摘要

随着大数据与人工智能技术的快速发展和广泛应用，以商业银行为代表的金融机构不断推出创新型信贷产品。然而，创新型信贷产品在投产初期面临着有标签样本不足、难以训练出有效模型的冷启动问题。该问题的传统解决方法是采用迁移学习方法，即利用外部机构相似信贷产品的丰富数据来辅助构建信贷违约预测模型。但是，随着技术的发展，网络犯罪分子不断利用新的手段来攻击数据，数据隐私保护和安全面临越来越大的挑战。银行业作为典型的数据密集型行业，在应用数据建模时会受到更多限制，导致不同机构间的数据无法交流互通，需要直接访问外部数据的迁移学习方法不再适用。

联邦学习为解决此问题提供了新的思路，其可以在保护数据隐私的前提下联合多方数据构建模型。然而，当数据间分布差异较大时，联邦学习的模型性能会因为数据异构性导致的“负迁移”现象受到抑制。联邦对抗域适应技术（FADA）可以通过对抗学习的方式对齐源域与目标域的特征，缓解“负迁移”现象，在保证模型精度的前提下提高模型泛化性能。但是，FADA 算法需要在各设备之间共享变换后的特征，该特征存在被解码的风险，会给数据安全保护尤为重要的银行业带来重大风险。因此，本文在 FADA 算法的基础上提出 Privacy-protected FADA（P2FADA）模型，一方面，该模型用共享梯度代替 FADA 中共享特征的做法，在数据隐私保护方面更加严格，在注重数据隐私和安全的金融信贷领域具有良好的适用性。另一方面，该模型通过对抗域适应技术缓解数据异构性带来的模型精度下降问题，模型性能优于经典的联邦学习算法 FedAvg，且在解决信贷风控冷启动问题时效率更高。

本文通过仿真实验和实证分析展示了 P2FADA 在数据异构性的场景下仍然具有优越的准确性和稳定性。首先，本文结合真实业务场景下信贷违约预测过程设计了数据生成规则。其次，选用源域与目标域条件分布的余弦相似

度和特征分布的 KL 散度两个指标从不同维度对源域和目标域之间的差异进行度量，并证明随着数据分布差异增大，模型准确率降低。接着，本文根据源域与目标域之间特征分布是否相同和条件分布是否相似梳理出四个典型信贷场景并进行仿真实验， P2FADA 在四个实验中均展现了突出的有效性和稳定性。最后，本文选取 2015-2018 年 Lending Club 平台的信贷数据进行实证分析。通过与逻辑回归、XGBoost、ANN 和 FedAvg 模型的对比，展现了 P2FADA 解决信贷风控冷启动问题和“负迁移”现象的优越性。

关键词：联邦学习；对抗域适应技术；信贷违约预测；隐私保护

Abstract

With the rapid development and wide application of big data and artificial intelligence technology, financial institutions represented by commercial banks continue to launch innovative credit products. However, in the early stage of production, innovative credit products are faced with the problem of cold start due to insufficient label samples and difficulty in training effective models. The traditional solution to this problem is to adopt the transfer learning method, that is, to use the abundant data of similar credit products of external institutions to assist in the construction of credit default prediction model. However, with the development of technology, cyber criminals constantly use new means to attack data, data privacy protection and security are facing more and more challenges. Banking industry, as a typical data-intensive industry, will be more restricted in the application of data modeling, resulting in the failure of data communication between different organizations, and the transfer learning method that requires direct access to external data is no longer applicable.

Federated learning provides a new way to solve this problem. It can combine multiple data to build a model under the premise of protecting data privacy. However, when the data distribution difference is large, the model performance of federated learning will be inhibited due to the "negative transfer" phenomenon caused by data heterogeneity. Federated adversarial domain adaptation (FADA) can align the features of source domain and target domain through adversarial learning, alleviate the "negative transfer" phenomenon, and improve the model generalization performance on the premise of ensuring the accuracy of the model. However, the FADA algorithm needs to share the transformed feature among the devices. The feature has the risk of being decoded, which will bring significant risks to the banking industry, where data security protection is particularly important. Therefore, this paper proposes a privacy-protected FADA (P2FADA) model on the basis of FADA algorithm. On the one hand, this model replaces the

shared feature in FADA with a shared gradient, which is more strict in data Privacy protection and has good applicability in the financial credit field that focuses on data privacy and security. On the other hand, the model alleviates the model accuracy decline caused by data heterogeneity through the counter domain adaptation technology. The model performance is better than the classical federal learning algorithm FedAvg, and it is more efficient in solving the cold start problem of credit risk control.

Through simulation experiments and empirical analysis, this paper shows that P2FADA still has superior accuracy and stability under the scenario of data heterogeneity. Firstly, this paper designs data generation rules based on the credit default prediction process in real business scenarios. Secondly, two indexes, cosine similarity of conditional distribution and KL divergence of feature distribution, are selected to measure the difference between source domain and target domain from different dimensions, and it is proved that the accuracy of the model decreases with the increase of data distribution difference. Then, according to whether the feature distribution is the same and the condition distribution is similar between the source domain and the target domain, this paper sorts out four typical credit scenarios and conducts simulation experiments. P2FADA has shown outstanding effectiveness and stability in the four experiments. Finally, this paper selects the credit data of Lending Club platform from 2015 to 2018 for empirical analysis. By comparing with logistic regression, XGBoost, ANN and FedAvg models, the advantages of P2FADA in solving the cold start problem of credit risk control and the phenomenon of "negative migration" are demonstrated.

Keywords: Federation learning; adversarial domain adaptation technique; credit default prediction; privacy preservation

目录

1.绪论

1.1 研究背景与意义

1.1.1 研究背景

近年来，随着大数据、云计算、人工智能、移动互联等技术的快速发展和广泛应用，互联网金融、数字经济等新兴领域的快速崛起为我国经济发展注入新动力，人们的消费观念也从保守型消费逐渐向信用消费转变，促使以商业银行为代表的金融机构进行自主创新，陆续推出多种纯信用类贷款产品，如融 e 借、快 e 贷、工薪贷、网捷贷等。信贷产品的业务流程主要有授信、用信和还款。在授信流程中，风险控制系统会对客户进行资质审核，评估其违约风险并决定是否授信，从而拒绝高风险客户，并优先授信给低风险高收益客户，帮助机构有效防范商业欺诈和保障盈利。由此可见，风险管理能力在商业银行拓展新信贷业务中占有核心地位，商业银行若想实现高质量发展，需要不断加强科技驱动风控的能力，而科技驱动风控的能力主要在于大数据风控技术，即利用大量外部数据（如社交数据、信用数据等）构建信贷违约预测模型，通过预测客户是否会发生违约行为，将客户分为“违约”和“守约”两类，从而拒绝违约客户，通过守约客户的授信申请[1,2]。由于信贷违约预测模型是一个有监督的二分类模型，若想训练出一个表现良好的信用评分模型，离不开大量有标签的训练样本的支持。然而，创新型信用贷款产品在投产初期缺乏甚至没有有标签的样本数据，难以训练出有效模型，该问题也称为信贷场景下风控系统的“冷启动”问题。

创新型信贷产品面临的冷启动问题，会导致银行等金融机构无法利用有限样本训练出有效的信贷违约预测模型，从而误导其做出错误的授信决策。一方面，若将违约客户误判为守约客户，则发放的贷款难以回收，使得银行

贷款规模下降，加剧其不良资产损失。另一方面，若将守约客户误判为违约客户，则会损失一部分收入并造成客户流失，阻碍银行进一步发展。此外，创新型信贷产品在冷启动阶段面临的违约预测困难问题，也打击了银行等金融机构在研发数字化、智能化创新产品的积极性，给金融行业的健康稳定发展带来影响。因此，创新型信贷产品面临的信贷风控冷启动问题在信贷领域普遍存在并影响巨大，成为学界与业界关注的重要议题。

目前，众多学者常用迁移学习的方法解决信贷风控冷启动问题，将基于外部相似信贷产品的丰富数据训练的分类模型迁移至创新型信贷产品进行违约预测。然而，这种迁移学习的方法需要直接访问外部数据。随着大数据的发展，对数据的使用是否会泄露隐私成为公众关注的焦点，隐私保护和数据安全领域立法兴起。中国相继发布了《人信息保护法》、《数据安全法》等法律法规，《个人信息保护法》对众多行业，尤其是金融行业的数据管理标准有重大影响[3,4]。《数据安全法》通过制定数据安全制度规范数据处理过程，并根据风险评估预警等机制实现全过程数据保护。金融行业的数据安全管理也与数据保护立法不断加快接轨，2020 年中国人民银行提出的金融行业标准《个人金融信息保护技术规范》，规定了个人金融信息在生命周期各环节的规范性要求，保障个人金融信息主体的信息安全与合法协议。次年，为了进一步保障金融数据安全流动，中国人民银行发布《金融数据安全 数据生命周期安全规范》和《金融大数据平台总体技术要求》作为金融行业标准，通过规范数据全生命周期各关键点的数据安全技术，确保金融数据安全管理策略的有效落实。

由此可见，数据安全治理已上升至国家安全的高度。银行业在提供金融服务过程中会接触到海量个人金融信息，不加限制地使用这些信息不仅会侵害客户的合法权益，对银行的声誉也会带来不利影响[5]。因此，作为数据密集的典型行业，银行业应用数据建模受到更多限制，这导致不同银行机构之间的数据无法交流互通，形成了“数据孤岛”现象，也使得风险控制能力难以进一步得到提升，此时，如何联合外部数据构建安全有效的模型是一大挑战。

联邦学习技术为解决此问题提供了新的思路，联邦学习是一种分布式机器学习框架，其可以在保护数据隐私的情况下联合多方数据构建模型，即仅

在本地使用数据集，通过交换模型参数进行训练。基于联邦学习框架训练的模型与传统训练方式下得到的模型效果基本一致，很好解决了“数据孤岛”问题。故可以使用联邦学习在保护用户数据隐私、银行信息安全的前提下，将拥有丰富数据的历史信用贷款产品的信贷违约预测模型迁移到创新产品上来解决冷启动问题。其中，拥有大量有标签数据的历史信贷产品为源域，而拥有少量数据的创新型信贷产品为目标域。

然而，由于不同机构贷款产品的目标客户、贷款额度及还款方式不同，客户的“守约”与“违约”表现是“相对”的，客户在 A 机构“守约”，但有可能在 B 机构发生“违约”行为。如互联网贷款产品与银行贷款产品相比，虽然门槛低、覆盖量大，但其利率、资金成本高，更容易发生违约行为。从理论层面来看，这种客户在不同机构违约表现不同的现象是由机构之间数据分布不一致引起的，又称数据异构性。当源域与目标域数据分布不一致时，在源域学习到的知识会导致目标域模型做出错误判断，反而会抑制目标域的模型性能，从而降低模型准确率，这种现象称为“负迁移”现象。而在联邦学习中，由于数据异构性带来的负迁移现象，会使全局模型参数更新的方向朝着局部模型参数偏移，导致全局模型远离全局最优解，模型性能大大降低。因此，在应用联邦学习框架解决信贷产品冷启动问题时，如何解决“负迁移”现象也是具有重要意义的问题。

1.1.2 研究意义

近年来，随着公众对保护数据隐私和安全的重视，与数据安全和隐私保护相关的立法和监管工作也在不断完善加强中，这使得包括金融行业在内的众多行业发展受到阻碍。此外，由于在真实业务场景中，不同机构间数据存在差异，不符合机器学习的独立同分布假设，联合多方数据建立机器学习模型也面临极大挑战。

本文结合理论研究与实践开发研究，提出银行等金融机构应用联邦对抗域适应技术克服负迁移现象，并构建有效信贷违约预测模型的可能性和合理性，并通过实证进一步完善该理论的全面性。在保护数据隐私和安全的情况下，利用外部数据建立有效的信贷违约预测模型，可以使银行等金融机构利润最大化，并鼓励其积极进行数字化、智能化创新，从而推动整个金融市场

蓬勃发展。

同时，联邦学习作为一种通用的分布式机器学习框架，可以打通不同机构之间的数据壁垒，且在数据隐私保护方面具有突出的优越性。除了可以应用于银行等金融机构信贷违约模型的构建，还为政府数据治理与数字经济建设等领域实现数据的共建、共享、共通提供了样例。如打破车载数据和路况数据的壁垒，在保护数据隐私的前提下利用乘客的行驶数据、驾驶数据等辅助自动驾驶模型的训练；在保护病人隐私的基础上助力医疗机构进行跨域合作，有利于建立更准确的疑难病预测模型，赋能智慧医疗等。有利于增强我国数字化能力，推动数字化转型，构建智能化国家治理体系，降本增效，拓宽发展空间。

1.2 研究内容和创新点

在信贷场景下，传统的信贷违约预测模型主要是有监督的分类模型，无论是经典的二分类模型逻辑回归，还是集成学习，亦或是神经网络，这些模型均需有大量有标签的数据参与模型训练。随着人工智能等技术的发展，金融行业不断进行自主创新，推出创新型信用贷款产品，这类产品在投产初期往往缺乏足够的数据进行模型训练，只依靠少量有标签数据训练得到的模型效果并不理想，甚至会为金融机构带来巨大损失。基于此问题，我们可以利用其他相似信贷产品的丰富数据补充创新型信贷产品的训练集。然而，由于与数据隐私保护相关的立法与监管不断加强，导致不同银行金融机构之间的数据无法交流互通，数据应用场景受到限制。

本文利用联邦对抗域适应技术来解决信贷产品在冷启动阶段构建模型困难的问题，联邦学习可以在保护数据隐私的情况下联合多方数据构建模型，且其模型效果与传统训练方式下得到的模型效果基本一致，很好缓解了数据隐私保护带来的限制。同时，由于需要借助外部数据进行模型训练，模型可能会因为域偏移现象导致知识负迁移，从而降低模型准确率。而将对抗域适应方法嵌入到学习特征表示的过程中，可以缩小表示空间上源域与目标域的模型特征距离，使得在源域训练得到的模型分类器可直接应用于目标域，缓解“域偏移”问题，提高模型表现效果。

本文受到联邦对抗域适应算法（FADA）的启发，提出了 Privacy-protected FADA（下文简称 P2FADA）。作为一种创新型的联合建模方式，P2FADA 可以在保证数据安全和隐私的前提下利用多方数据训练出统一的全局模型，同时还能保证较好的模型效果，在金融信贷领域具有突出优势。本文的创新点主要有以下三个方面：

(1) 联邦学习发展至今，在金融信贷领域的研究较少，并且没有将联邦对抗域适应技术应用于信贷领域的相关研究。然而，联邦对抗域适应技术可以克服由数据异构性带来的负迁移现象，使得其非常适用于解决信贷产品的冷启动问题，本文将改进后的联邦对抗域适应算法应用于信贷领域，具有一定的研究价值和创新性。

(2) 本文在 FADA 的基础上提出 P2FADA，该模型性能优于经典的联邦学习算法 FedAvg，通过共享梯度代替 FADA 中的共享特征的做法，在数据隐私保护方面更加严格。同时，由于梯度与 batch size 大小无关，仅与输入和输出的数据维度有关，共享梯度也不用担心批数据量太大导致通信效率降低，故 P2FADA 支持大 batch size 训练。在数据隐私保护和数据安全制度日趋严格的大背景下具有良好的实用性和有效性。

(3) P2FADA 是一个通用的分布式机器学习框架，可以灵活应用于需要在保护数据隐私的前提下联合多方数据建模，并且数据之间存在差异性的场景中，如智慧医疗、智慧城市等，对我国的数字化转型很有帮助。

1.3 研究框架与路线

本文开篇阐述了信贷场景下风控系统的冷启动问题的研究背景和意义，并对本研究相关的文献进行整理，之后介绍了本文研究的相关理论和技术，方便读者理解本文研究。在实验分析部分，为了验证 P2FADA 解决信贷冷启动问题的有效性，首先，本文根据真实业务场景中信贷违约预测过程，设计了仿真实验的数据生成规则。其次，对源域和目标域之间的差异性进行度量，并说明数据差异性对单源域迁移模型预测效果 AUC 的影响。接着，本文从条件分布差异和特征分布差异两个方面刻画源域与目标域间的数据差异，梳理出四个典型场景并进行仿真实验，分析 P2FADA 在不同场景中的准确性和稳定性。最后，本文选取 Lending Club 平台的真实数据进行实证分析，验证

P2FADA 在表现效果。基于上述研究思路与研究方法，整理得出本文的研究路线图，如图 1-1 所示。

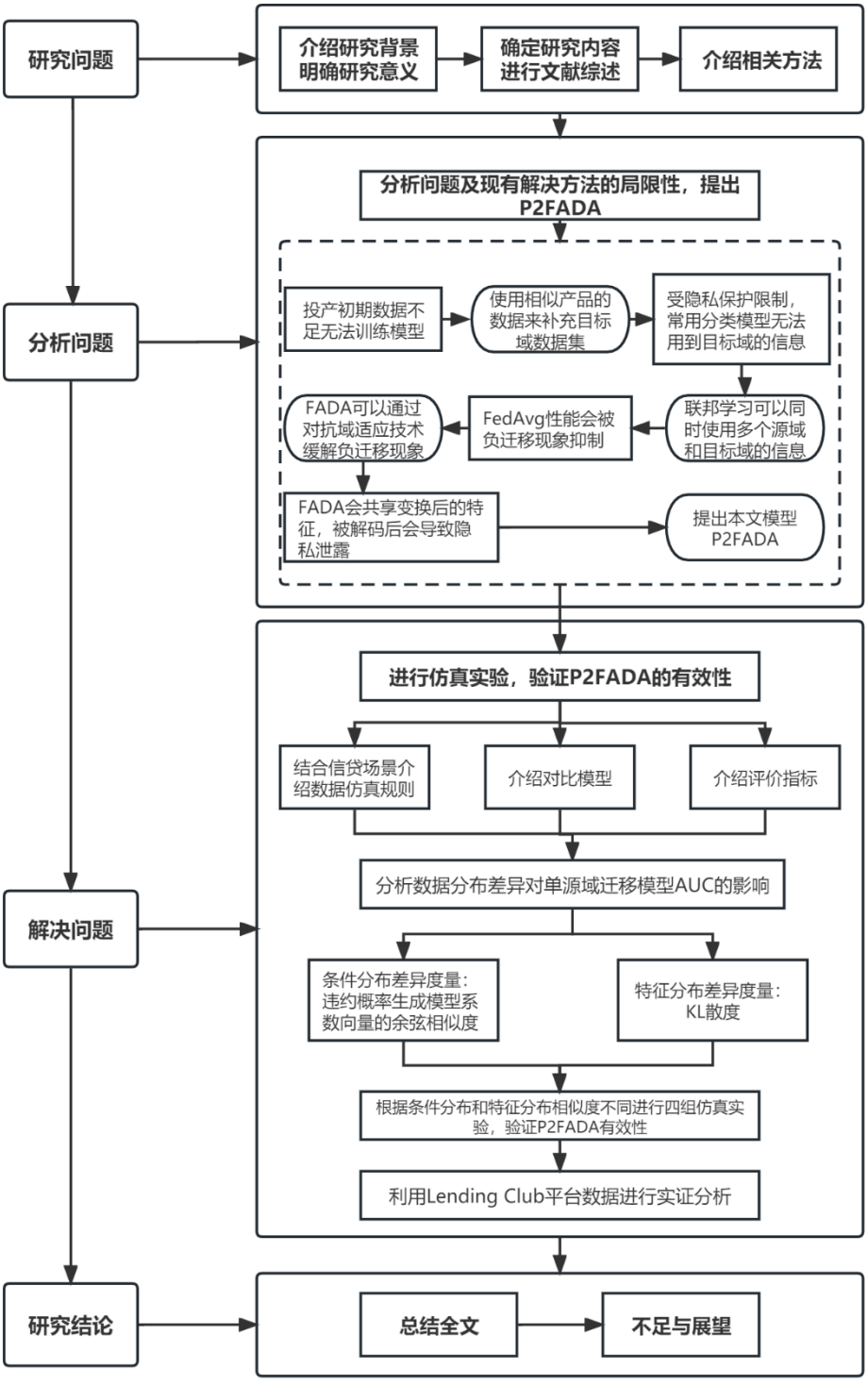

图 1-1 论文研究路线图

1.4 本文结构

本文共分为七个章节，各章节安排如下：

第一章为绪论，首先讲述了信贷产品冷启动问题的产生及其影响，并说明了迁移学习和联邦学习解决该问题的局限性，提出了使用了联邦对抗域适应技术解决信贷产品冷启动问题的重要意义，接着介绍了本文的创新点，并简要阐述了本文的研究路线和文章结构。

第二章从信贷产品冷启动问题的解决方法入手，首先介绍了常用的解决方法及其局限性，并提出了使用联邦学习解决该问题的有效性和可能遇到的挑战，即数据异构性、访问受限等问题，并梳理相应解决方法。

第三章为本文的相关方法介绍部分。在理论方面，阐述了信用风险理论和常用的信用风险管理方法。在技术方面，对本文研究中涉及的逻辑回归、XGBoost 和 ANN 模型进行说明，并对本文涉及的横向联邦学习和迁移联邦学习内容及其代表算法进行详细介绍，对本文未涉及的纵向联邦学习进行简要说明。

第四章为研究设计部分，本章首先通过分析本文待解决问题及目前现有解决方法的局限性，提出本文所提出的联邦对抗域适应模型 P2FADA 对该问题的适用性及其具体框架。接着结合金融信贷场景介绍了仿真数据的模拟规则，并阐述了本文选取的对比模型及模型评价指标，之后提出了两个度量数据差异性的指标，并依据数据差异的不同设计出四个仿真实验。

第五章为实验分析部分，首先介绍了每个实验中目标域和源域的划分情况，接着展示实验结果，并分析 P2FADA 模型在不同场景下的准确性和稳定性。

第六章为实证分析。本章首先介绍了实证数据来源及其预处理过程，之后依据具体问题划分出实验所需的目标域与源域，最后对实验结果展开讨论与分析。

第七章为总结与展望。本章对本文的研究内容进行回顾，并提出了后续可改进和进一步研究的方向。

2.文献综述

近年来，随着人工智能算法的不断进步和大数据云计算的开发应用，促进数字化转型、布局数字经济已成为大势所趋。然而，在银行等金融机构不断推陈出新，向数字化转型的过程中依然面临着众多问题。其中一个重要的挑战就是信贷风控冷启动问题，该问题导致模型缺乏足够的有标签数据训练有效的信贷违约预测模型。此外，随着数据隐私和数据安全方面的法律法规不断兴起，导致各行业、各企业之间的数据并不能连通起来，内部数据就像一个个孤岛一样，无法与外部数据共享流通，我们将这种问题成为“数据孤岛”问题。

联邦学习（Federated learning，FL）的出现，提供了解决上述问题的新思路，联邦学习是一种在保护数据隐私的前提下使用机器学习建模的方法，该方法使得各客户端（如移动设备）在中央服务器（如可信的服务提供商）的协调下联合建模，实现高效安全的数据合作，其核心思路是“数据不动模型动，模型可用不可见”。使用联邦学习可以在保护数据隐私的前提下利用外部机构的历史信贷数据进行建模，丰富模型的训练集，是一种解决信贷产品冷启动问题的思路[6-8]。

然而，传统的联邦学习模型表现会因源域与目标域数据间的差异增大而变差，导致其在信贷领域的应用受到限制，故如何解决数据异构性对联邦学习的影响成为众多学者关注的议题。

本章首先梳理了解决信贷风控冷启动问题的常用方法，其次介绍了联邦学习联合多方数据解决信贷产品冷启动问题的适用性，接着阐述了数据异构性对联邦学习模型性能的影响及解决方法。

2.1 信贷风控冷启动相关研究

信贷违约预测模型是一个有监督的二分类模型，若想训练出一个表现良好的信用评分模型，离不开大量有标签的训练样本的支持。然而，创新型信用贷款产品在投产初期缺乏甚至没有有标签的样本数据，难以训练出有效模型，该问题也称为信贷场景下风控系统的冷启动问题（为叙述方便，下文称其为信贷产品冷启动问题）。

根据是否有样本数据，可将信贷产品冷启动问题的解决方法分为两种：第一种是无样本信贷产品冷启动解决方法，即创新型信贷产品在投产初期没有任何无样本数据；第二种是有样本信贷产品冷启动解决方法，即创新型信贷产品在投产初期有少量有标签样本或者大量无标签样本。

(1)无样本信贷产品冷启动解决方法

由于没有可以使用的样本数据，对于无样本信贷产品冷启动问题，宋捷[9]等在 2020 年提出依靠行业“替代”数据或经验丰富的专家解决该问题，可以将其分为基于外部评分数据的冷启动方法和基于德尔菲法的冷启动方法。

基于外部评分数据的冷启动方法指的是利用可获得的且较为权威的外部评分数据进行客户违约判断，如将中国人民银行征信中心发布的征信报告或第三方征信机构（如百行征信）出具的个人信用报告作为是否授信的主要决策依据。

基于德尔菲法的冷启动方法[10]指的是在该领域的专家凭借其对业务的深入理解和丰富经验，通过对客户个人特征、财务特征等指标进行分析评分，构建风险评级模型，基于该模型对客户进行违约判断。这些方法不够有针对性，且主观因素影响较大，需要寻找更加有效的方法解决冷启动问题。

(2)有样本信贷产品冷启动解决方法

当拥有少量有标签样本和大量无标签样本时，可以考虑将迁移学习应用于解决冷启动问题[11]。迁移学习通过在具有丰富数据的源域D_s上学习，将在源任务T_s上学习的“知识”应用于目标域D_t，以此来提升与原任务相似但不相同的目标任务T_t的模型表现。Pan 错误!未找到引用源。等人于 2010 年发表的迁移学习综述中，将迁移学习按照学习方法进行分类并指出了负迁移现象对模型性

能的影响。其中，基于模型的迁移学习可以用于解决信贷产品冷启动问题，即将具有丰富数据且与创新型贷款产品相似的历史贷款产品数据作为源域，将源域上学习所得的模型直接应用于创新型贷款产品。然而，该方法的前提是源域任务与目标域任务相似，否则模型将会把在源域学习到的“错误知识”应用于目标域，从而降低模型性能[13-16]。此外，迁移学习在模型训练过程中需要直接访问数据，不满足数据隐私保护的要求。

2.2 联邦学习与信贷风控冷启动

联邦学习可以在保护数据安全的基础上，利用外部金融机构的历史信贷产品的丰富数据扩充模型训练集，从而解决信贷产品的冷启动问题[17-19]。

2016 年，谷歌的 McMaha[20]等人为解决安卓系统更新时用户隐私保护问题提出的联邦平均算法（Federated Averaging，FedAvg）是联邦学习的开山之作。其主要思想为：由各个移动设备通过其本地数据集训练模型，之后向中央服务器发送更新的模型参数，中央服务器通过聚合收到的模型参数得到更新后的全局模型，并将其下发至各个设备。这种方法用上传更新后的模型参数替代直接上传数据训练模型，以此在不泄露数据的情况下不断迭代更新模型，保证每个设备的用户数据不出本地，各个设备又可以共享同一个全局模型，完成模型的构建。FedAvg 算法为目前最常用的联邦学习算法，其经常作为联邦学习研究中的基准算法。与 FedAvg 算法相似，Yang[21]等人在联邦学习框架下进行信用卡欺诈检测，提出 FFD（Federated learning for Fraud Detection）方法，该方法可以使各银行利用其本地数据集训练欺诈检测模型，之后将各个银行本地欺诈检测模型更新的参数聚合起来得到全局的欺诈检测系统，使得各银行可以在不共享数据的前提下共享表现更好的全局模型。Long[22]等人说明了将联邦学习应用于以“数据共享”为核心思想的开放银行场景中的适用性，并探讨了开放银行项目时可能遇到的挑战，如数据异构性、访问限制等问题，同时也讨论了相应的解决方案。

将联邦学习应用于信贷领域，可以在保护数据隐私的前提下，联合外部数据建立有效模型，解决了由于有标签样本不足带来的信贷风控冷启动问题。然而，由于各方数据存在分布不一致的问题，在联合建模时会因负迁移现象

导致全局模型性能下降。接下来对数据异构性给联邦学习带来的局限性进行说明，并梳理现有解决方法。

2.3 异构联邦学习的局限性与解决方案

2.3.1 异构联邦学习及其局限

在实际场景中，由于不同银行的数据是独立产生、互不相通的，客户可能来自于与目标银行具有不同特征空间的银行，也可能在不同银行均有同一个客户的特征信息，故各设备上的数据往往是非独立同分布的，即各设备本地数据的分布不一定能从模型的总体数据分布中得出，该特点称为联邦学习的数据异构性[23,24]。

Li[25]等人分析了 FedAvg 算法在非独立同分布数据上的收敛性，发现数据的异构性会导致 FedAvg 算法难以实现线性加速，收敛速度变慢。此外，当联邦学习基于非独立同分布数据进行模型训练时，可能会出现客户端漂移现象，即中央服务器模型参数更新的方向会向设备模型参数偏移，导致全局模型远离全局最优解，模型性能大大降低[26]。可以看出，数据的异构性不仅给算法设计带来挑战，也使得理论分析变得困难。

2.3.2 异构联邦学习解决方案

为了解决数据异质性带来的模型难以收敛、性能变差的问题，现有解决方法主要有平衡数据、对齐本地模型与全局模型以及个性化学习[27]。

(1)联邦学习平衡数据方法

目前联邦学习中数据的不平衡性主要体现在数据类别分布不平衡和数据量分布不平衡两个方面。Zhao[28]等人通过构造高度倾斜的异构性数据集进行实验后发现，当每个客户端只使用 CIFAR-10 数据集中单一类别的数据进行训练时，FedAvg 算法训练的神经网络准确性显著降低了 55%。作为解决方案，Zhao 等人提出通过构造一个均匀分布的公共数据集来改善异构性数据的影响。受此方法启发，Huang[29]等人提出了 LoadAdaBoost FedAvg 算法，将

共享均匀数据集的思想应用于医疗场景，有效提高了应用异构性医疗数据建模的模型准确率。然而，在联邦学习的架构下，不一定总是能维护起一个合适大小的共享数据集。Duan[30]等人提出了一种联邦自平衡框架 Astraea，该算法通过平衡数据类别来缓解数据异构性带来的联邦学习模型性能下降的问题。首先，在模型训练之前，Astraea 根据 Z-score 划分数据集，并对少数类数据进行数据增强（对数据进行旋转、移动等操作），对多数类数据进行下采样来缓解全局不平衡。其次，在训练过程中，该算法通过引入一个协调器，根据各设备数据分布的 KL 散度（Kullback-Leibler divergence，KLD)来选择数据分布均匀的设备进行聚合，从而达到新的平衡。实验表明，该算法在类别不平衡的数据集上表现良好，准确率有所提高。除了采用数据增强的方法外，Li[31]等人提出了一种有效缓解数据异质性的算法 FedBN，该算法与 FedAvg 算法类似，但在训练过程中增加了局部批处理归一化层，以此来缓解局部特征偏移的问题。虽然结果显示 FedBN 算法的收敛速度优于 FedAvg 算法，但是 Hsieh[32]等人的研究表明批处理归一化层也会带来精度损失问题。Lin[33]等人将集成蒸馏学习拓展至联邦学习，提出了联邦蒸馏融合算法 FedDF，该方法利用未标记或人工生成（如利用 GAN 生成）的数据来对中央服务器上聚合的模型进行融合训练，从而避免批处理归一化层带来的精度损失并打破了异质性设备之间的知识壁垒，提升了模型精度。进一步地，Sattler[34]等对联邦蒸馏法进行改进，提出了一种基于软标签量化、增量编码和双重蒸馏的高效联邦蒸馏方法：压缩联邦蒸馏（Compressed Federated Distillation，CFD），该方法从知识蒸馏中的协同蒸馏法受到启发，通过提取数据管理、上游量化、增量编码、双重蒸馏和下游量化物种技术扩展了传统的联邦蒸馏框架，通过无损压缩增量编码降低了通信的信息量，加快模型收敛速度。

由于各设备数据量相差太大也会导致数据异质性，Smith[35]等人将多任务学习与联邦学习相结合，提出了联邦多任务学习算法，该算法的主要思想为：中央服务器采用多任务学习的思路，根据各设备上传的模型参数来学习不同模型之间的关系，之后本地模型就可以通过交替优化算法协同训练，根据中央服务器学习到的模型关系来更新自己的模型，从而缓解数据量不平衡带来的异质性问题，获得高质量模型。此外使用生成模型生成假数据也是一

种缓解数据量不平衡的方法，Jeong[36]等人提出了一种联邦增强算法 FAug，该算法联合各设备共同训练一个生成模型，并通过此生成模型生成独立同分布的数据集供模型训练，值得注意的是，该算法要求各设备上传少量样本至中央服务器以供其训练生成模型，故会增加隐私泄露风险和通信开销。

(2)联邦学习对齐本地模型与全局模型方法

为了解决数据异构性带来的客户端漂移问题，主要有在各设备本地的损失函数中加入惩罚项和对抗对齐特征两种优化算法的做法。对于加入惩罚项优化算法，Li[37]等人提出了 FedProx 算法，该算法在各设备本地的优化函数中加入一个惩罚项，限制本地模型更新不会过分偏离上一轮的全局模型，最终使得算法更加稳定，收敛速度更快。与 FedProx 算法中对所有参数使用相同的惩罚力度不同，Shoham[38]等人采用终身学习的思路，提出使用 Fisher 信息矩阵来选择性地惩罚那些偏离全局模型太远的参数，来保护每个任务中的重要参数。除了加入惩罚项来限制本地模型更新方向，也可以使用全局知识来协调模型的更新方向。Karimireddy[39]等人提出的 SCAFFOLD 算法就是通过全局知识来协调本地模型的更新方向，防止本地模型更新出现“局部漂移”的现象。然而，这种在训练中加入惩罚项和利用全局知识的方法增加了通信开销，降低了通信效率。

对于对抗对齐特征，Hoffman[40]等人提出用对抗域适应技术来缓解“负迁移”现象，将各设备上的本地训练数据视为源域，将中央服务器上的训练数据视为目标域，则域适应算法可以在保持源域分类模型精度的前提下，缩小表示空间上源域与目标域的模型特征距离，使得利用源域数据训练的模型可以直接应用于目标域。然而，对抗域适应算法需要直接访问源域与目标域的数据，为了保护数据隐私，Peng[41]等人提出了联邦对抗域适应算法 FADA，其将联邦学习框架与对抗域适应技术结合在一起，可以在一定程度上保护数据隐私，解决域偏移问题。

(3)个性化联邦学习

当各设备上的数据分布相差很大时，单一的全局共享模型很难在每个设备上都表现得很好，需要对每个设备做针对性训练，而这种挑战正是个性化学习尝试解决的问题。Sim[42]等人提出了一种“联邦训练+本地适应”的联邦个性化学习框架，首先学习一个全局模型，之后在本地通过梯度下降等操作

对模型进行个性化处理，提高数据异质性下的模型性能。然而，采用这种“联邦训练+本地适应”两步走的方法会将训练与个性化过程割裂开来，难以得到最优的个性化模型。Jiang[43]等人将元学习算法（Model Agnostic Meta Learning, MAML）应用于 FedAvg 算法，提出了 FedAvg+算法。元学习可分为两个阶段。第一个阶段是 meta-training，该阶段可以找到一组最优的初始化参数，快速适应一系列新任务。第二个阶段是 meta-testing 阶段，该阶段可以通过对模型微调得到较优的个性化模型。与 FedAvg+算法类似，Canh[44]等人提出了 pFedMe 算法，该算法通过 Moreau envelope 函数将个性化模型与全局模型分离，使得优化全局模型和个性化模型两个过程可以并行，从而达到较快的收敛速度。个性化学习作为一种联邦学习解决数据异构性的新兴技术，还需要进一步探索和突破。

综上所述，目前解决联邦学习数据异构性的方法都或多或少存在增加通信开销和泄露隐私的风险，如何保证在更少的通信开销、更强的隐私保护下学习到有效的全局最优模型是数据异构性联邦学习下一步研究的方向。

2.4 文献评述

从上文对信贷风控冷启动和联邦学习相关的文献梳理可以看出，联邦学习在信贷领域的研究应用较少。在已有的研究中，绝大多数研究重点在于如何在保护数据隐私的前提下联合多个银行等金融机构建立信贷违约预测模型。而对真实业务场景中各机构之间的数据异构性带来的模型性能局限关注度较低。现有解决异构联邦学习的方法，或多或少存在增加通信开销和泄露隐私的风险。对抗域适应技术一方面可以在保持源域分类模型精度的前提下，缩小表示空间上源域与目标域的模型特征距离，使得基于源域训练的模型可以直接应用于目标域[45]。另一方面，对抗学习与非对抗学习相比，优势在于可以学习足够好的特征表示，帮助缩小源域与目标域的特征距离，将对抗域适应技术与联邦学习融合具有一定可行性。对抗域适应技术发展至今，在信贷领域的研究较少，且目前还没有将对抗域适应技术与联邦学习融合并应用于信贷领域的研究，故将联邦对抗域适应技术应用于信贷领域具有一定的创新性和研究价值。已有的联邦对抗域适应技术还存在一些不足，研究[41]在实际

应用中，没有考虑各银行机构之间共享变换后的特征存在被解码的风险，会给数据安全保护尤为重要的银行业带来重大风险。本文在联邦对抗域适应技术（FADA）的基础上提出 P2FADA 算法，该模型表现与 FADA 相当，通过共享梯度代替 FADA 中的共享特征的做法，在数据隐私保护方面更加严格，且不受模型 batch size（批数据量）限制，支持大 batch size 训练。在数据隐私保护和数据安全制度日趋严格的大背景下具有良好的实用性和有效性。

3.相关方法介绍

本文对信贷风控的相关理论进行介绍，列举了传统的信用风险管理方法及其代表算法，之后将联邦学习分为三个类别并进行介绍，并对本文涉及的横向联邦学习和迁移联邦学习展开详细阐述，并介绍其经典算法。

3.1 信贷风控相关理论

3.1.1 信贷产品的风险介绍

个人信用贷款是指银行等金融机构依据贷款申请人的信誉发放的贷款，目前信贷业务逐渐占据重要地位，成为商业银行的核心业务。在为银行带来利润的同时，信贷业务也伴随着诸多风险，其中信用风险占主要地位[46,47]。信用风险是指在金融交易过程中，贷款者由于信用变差或失去偿还本息能力等原因无法按照合同还款导致交易对手产生金融损失的风险，故又称违约风险，该风险造成的损失比重很大[48]。

除了信用风险，信贷业务风险的另一种表现形式是欺诈风险，即贷款人通过伪造个人信息等方式恶意骗取贷款，给借款银行带来损失。本文主要对信用风险管理进行研究。

3.1.2 信用风险管理方法

目前主流的信用风险管理方法可以分为三类，即专家评定法、贷款评级模型和信用卡评分模型。

(1)专家评定法

专家评定法是指由具有丰富经验的信贷专家依据其专业知识、对借贷人

信息整合以及对关键因素的权衡做出是否贷款的决策。专家评定法是一种较为传统的信用风险评估方法，该方法通过对影响企业信用的因素建立评价指标并赋予权重来做出最终决策。根据构成评价指标的影响因素不同，专家评定法也被分为不同形式，其中比较常见的方法为“5C”要素分析法[49]，其主要由债务人品格(Character)、企业未来的资本质量及价值(Capital)、债务人偿还债务能力(Capacity)、企业经营环境(Capital)和企业抵押品质量(Collateral)五个要素构成。

虽然专家评定法具有操作便捷、适应性强和应用广泛等优点，但由于其受个人主观性影响很大，且对高素质专业人员极度依赖，成本较高，面对日益复杂的银行经营形式已渐露疲态。

(2)贷款评级模型

贷款评级模型是根据借款人的还款能力、贷款项目的盈利能力、贷款的担保以及信贷资产质量等因素对贷款进行分类。2007 年，银监会下发了《贷款风险分类指引》，在银行业推广贷款质量的五级分类，分别是：正常贷款、次级贷款、关注贷款、可疑贷款和损失贷款，关注、可疑和损失贷款损失率在 30%-100%，风险较大，为不良贷款。目前贷款五级分类的评价标准是我国银行业较为主流的信用风险评估方法，不少银行在此基础上对分类方法进行进一步细化，形成个性化贷款分类管理体系[50]。

然而，由于贷款五级分类法主观性较强且缺乏数据支撑，对信息科技的利用不足且自主性不足，依据此方法做出的决策不够科学合理。

(3)信贷违约预测模型

信贷违约预测模型依据借款人的个人特征、财务特征和贷款特征等数据预测借款人是否会违约，是一个有监督场景下的二分类模型，目前构建模型的主流算法主要有分类算法、集成学习算法以及神经网络模型[51-53]。

a)分类算法

选择一个合适的分类器是构建信贷违约预测模型的关键部分，传统的分类器有很多，如：逻辑回归（Logistic Regression)、SVM、K 近邻（KNN）、决策树等，逻辑回归是一种经典的二分类模型，可解释性强，本文对其进行介绍。

逻辑回归是一种二分类算法。二分类任务的本质是找到一个单调可微的

函数将模型线性回归的预测值映射到样本真实标签上[54]，在逻辑回归中选取的是对数几率函数，则可得到：

$$p=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-\left(w^{T}x+b\right)}} \tag{3-1}$$

于是有：

$$ln\frac{p}{1-p}=w^{T}x+b \tag{3-2}$$

对于二分类问题来说，令p表示x的标签Y为 1 的概率，即 $p=P\left(Y=1\middle|x\right)$，则$1-p$表示 x 的标签Y为 0 的概率，即$1-p=P\left(Y=0\middle|x\right)$，两者的比值称为几率（odds）则有：

$$w^{T}x+b=ln\frac{P\left(Y=1\middle|x\right)}{1-P\left(Y=1\middle|x\right)} \tag{3-3}$$

$$P\left(Y=1\middle|x\right)=\frac{1}{1+e^{-\left(w^{T}x+b\right)}}=\frac{e^{w^{T}x+b}}{1+e^{w^{T}x+b}} \tag{3-4}$$

$$P\left(Y=0\middle|x\right)=1-P\left(Y=1\middle|x\right)=\frac{1}{1+e^{w^{T}x+b}} \tag{3-5}$$

由式 3-3 可知，$Y=1$的对数几率可由 x 的线性组合得到。由式 3-4 可知，通过建立 x 的线性组合与分类概率的联系，就可以得到客户的违约和守约概率，且 $w^{T}x+b$ 的值越接近于正无穷，$P\left(Y=1\middle|x\right)$的值越接近于 1，即违约概率越高。对于类别预测，一般会从实际问题出发设置一个阈值，将分类概率与阈值相比较，若分类概率大于阈值，则为正例，否则为反例。将式 3-5 写成似然函数的形式，再利用牛顿法、梯度下降法等即可求解参数w。

b)集成学习

通过单一分类器进行分类预测，极易发生过拟合现象，模型泛化性能较差，集成学习（ensemble learning）通过构建和组合多个基学习器（Base learner）来解决单一模型预测泛化效果差的问题，其核心思想是将多个分类器的独立预测结果通过线性加权融合、预测融合等方法进行组合，使得模型的泛化和分类性能进一步提升，示意图如下。

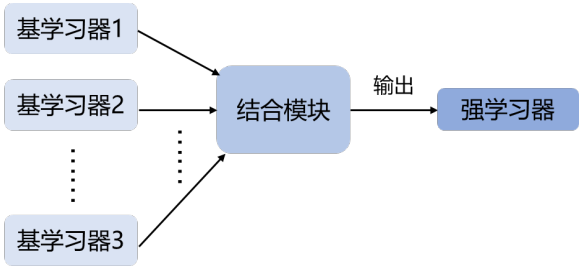

图 3-1 集成学习示意图

集成学习算法主要由构建基学习器和组合基学习器两部分构成。其中，构建基学习器是指用平行方法或顺序化方法生成一系列基学习器，使用平行方法构建的基学习器可以并行生成，并独立进行学习和预测任务，它们之间是互不依赖的，代表算法有随机森林（Random Forest）。而使用顺序化方法生成的基学习器是串行生成的，新的基学习器需要依赖上一个基学习器的预测结果来生成，代表算法有 Adaboost、XGBoost 和 GBDT 等。组合基学习器的方法有投票法、平均法、学习法等。下面对本文用到的 XGBoost 算法进行介绍。

XGBoost 是一种大规模并行决策树的 Boosting 框架，其在分类问题上表现很好，被广泛应用于工业界。Boosting 框架的基本思想为：基于初始训练集从初始化模型，得到一个基学习器，再根据基学习器的在初始训练集上的分类结果对样本权重进行调整，即给予分类错误的样本更大的权重，给分类正确的样本较小的权重，不断重复此步骤，直至达到预设准确率后停止训练，即可得到性能较好的强学习器[55]。XGBoost 的目标函数为：

$$L(\Phi)=\sum_{i=1}^{n} l(\hat{y}_i - y_i)+\sum_{k}^{K}\Omega(f_k) \tag{3-6}$$

其中 $l(\hat{y}_i - y_i)$ 是第 i 个样本的预测误差，$\sum_k^K \Omega(f_k)$ 表示 K 棵树的复杂度求和，为正则化项，用来控制模型的复杂程度，其展开式为：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2 \tag{3-7}$$

其中 T 为叶子节点的个数，w 为叶子节点权重，γ 和 λ 为控制叶子数权重的参数，可以看到，叶子节点数越少，叶子节点权重越低，模型复杂度越低，泛化能力越强。由于 XGBoost 是并行生成基学习器的，其基于前一个决策树的预测结果来生成新决策树，故目标函数可以写成：

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) + C \tag{3-8}$$

其中 n 为决策树的数量，C 为常数项。

泰勒公式可以将非线性问题转化为线性问题，在近似计算上很有优势，下面将 XGBoost 进行二阶泰勒展开来求解，其中二阶泰勒展开的公式为:

$$f\left(x+\Delta x\right) \approx f\left(x\right) + f^{'}(x)\Delta x + \frac{1}{2} f^{''}(x)\Delta x^2 \tag{3-9}$$

为了使表达形式更简洁，令：

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{3-10}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \widehat{y_i}^{(t-1)}\right) \tag{3-11}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{3-12}$$

则可以得到目标函数的表达式：

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y_i}^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^{\,2}\left(x_i\right)\right] + \Omega\left(f_t\right) + C \tag{3-13}$$

故在训练时，目标函数可以写作：

$$Obj^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{3-14}$$

其中I_j表示每棵树第 j 个分裂节点上的样本集合，进一步地，令$G_j = \sum_{i\in I_j} g_i$，$H_j = \sum_{i\in I_j} h_i$，则可得到：

$$Obj^{(t)} = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] + \gamma T \tag{3-15}$$

由于G_j和H_j是前$t-1$步的计算结果，可视为常数，故未知参数只有w_j，对目标函数求一阶导并令其为 0，即可得到w_j的值，即：

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{3-16}$$

将w_j^*代入目标函数，即可得到目标函数的最优解：

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{3-17}$$

c)神经网络

神经网络（NN）是由大量并行运算的简单计算单元（神经元）联结组成的非线性网络，因其一定程度上模仿了生物神经网络的结构和功能而得名，其核心思想是将分布式存储的信息通过神经元并行协同运算后进行求解。神经网络具有很强的自适应学习能力，且不依赖研究对象的数学模型，能够很好地胜任分类预测任务。目前，神经网络在金融、医疗、交通、工业经济等领域得到了广泛应用，是机器学习领域的主流算法之一，根据网络结构的不同，神经网络有很多变种，如：人工神经网络（ANN）、卷积神经网络（CNN） 和循环神经网络（RNN）。下面以结构较简单的 ANN 为例介绍神经网络。

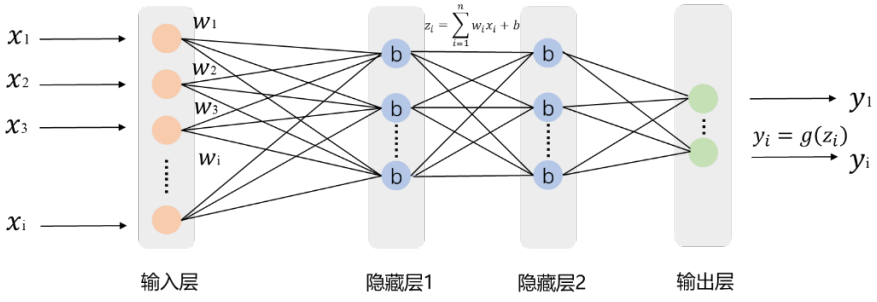

图 3-2 ANN 结构

ANN，也被称为多层感知机，其网络结构如图 3-2 所示，可以看到，ANN 由输入层、隐藏层和输出层组成。输入层将输入的数据x_i分发给隐藏层，隐藏层将x_i乘上对应权重w_i再加上偏置项b得到计算结果z_i，则有：$z_i = \sum_{i=1}^{n} w_i x_i + b$，之后对$z_i$进行非线性映射得到输出$y_i$，$y_i = g(z_i)$，该过程称为前向传播。

在构建神经网络时可以根据实际需求自行设置隐藏层和每个隐藏层神经元的数量，通常神经元数量可以参照以下公式设计：

$$l = \sqrt{n+m} + a \tag{3-18}$$

其中，l 为隐藏层神经元数，n为输入层神经元数量，m 为输出层神经元数量，$a \in [1,10]$，是一个调节常数。

对z_i进行非线性变换的函数g称为激活函数，激活函数是神经网络可以学习复杂任务的核心要素，如果没有激活函数，神经网络只能依据权重w和

偏差b做线性变换，即使叠加多层神经网络，也依旧相当于一个线性回归模型，限制了神经网络的学习能力，而激活函数可以对输入进行非线性变换，使得神经网络可以适应更复杂的学习场景。常用的激活函数有：Sigmoid 函数、Tanh 函数、Relu 函数等，不同激活函数适用的场景不同，在构建神经网络时需要根据实际问题去选择激活函数[56]。

在构建神经网络时，可以根据不同学习任务选择不同损失函数，对于分类任务，常用的损失函数为交叉熵函数，而均方误差（MSE）常用作回归任务的损失函数，每个训练样本损失函数的总和为代价函数。有了目标函数，就可以进行参数更新了，神经网络采用梯度下降的方式来不断更新权重w和偏差b的参数，使权重沿着损失变低的方向移动，这一过程称为反向传播。

ANN 模型的训练流程如下图所示：

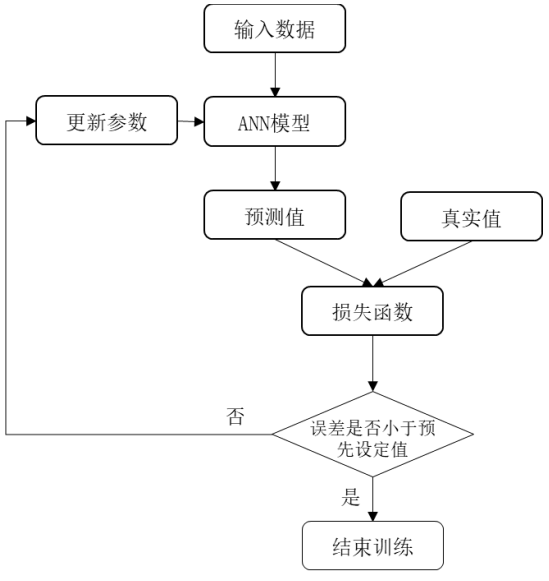

图 3-3 ANN 模型训练流程

3.2 联邦学习

3.2.1 联邦学习概念

联邦学习（Federated Learning）是一种分布式机器学习框架，在这种框架下，多个参与方可以在保护数据隐私的前提下协同训练，从而得到一个更优的全局模型。这种方法只需要交换模型参数或中间计算结果，不需要共享

数据，极大程度上保证了数据安全，可以在合法合规的前提下联合多方数据建模，进一步提升模型的效果[1,6]。根据联邦学习框架中各参与方上数据重叠情况的不同，联邦学习主要可以分为横向联邦学习（Horizontal Federated Learning）、迁移联邦学习（Federated Transfer Learning）和纵向联邦学习（Vertical Federated Learning）三个类别，本文主要涉及到前两类联邦学习，故下文对横向联邦学习和迁移联邦学习展开详细阐述，并介绍其代表算法，对本文未涉及的纵向联邦学习进行简要概述。

3.2.2 横向联邦学习

对于各个参与联邦学习的参与方上的本地数据，若满足数据特征重叠多而样本重叠少的条件，如在不同地区的两家银行，其业务可能是相似的（数据特征空间相似），而用户不同（样本空间不同），则可以将各参与方的数据集按照横向（样本维度）进行切分，并联合数据特征相同而样本不同的部分进行训练，这种将数据集进行横向切分进行训练的方法叫做横向联邦学习，示意图如下图所示。

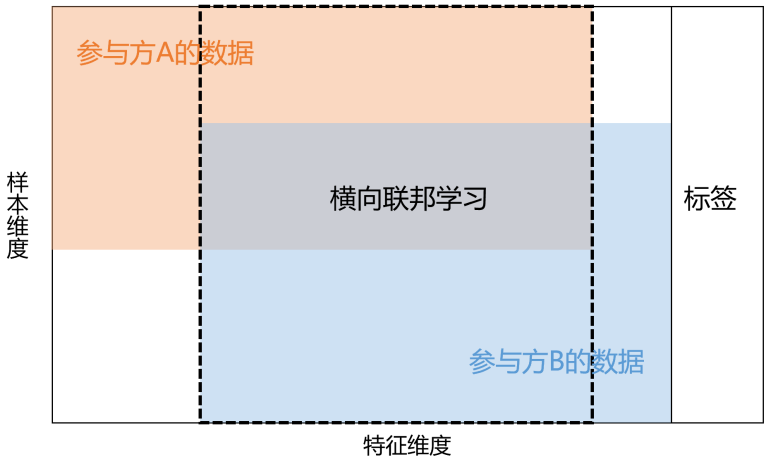

图 3-4 横向联邦学习

横向联邦学习常用架构主要有“中央服务器-客户端”架构和“对等网络”架构，下面对这两个架构进行介绍。

(1)中央服务器-客户端

在“中央服务器-客户端”架构中，参与方被分为中央服务器（server）

和客户端（client），其中客户端为参与训练的设备，负责独立训练机器学习模型，中央服务器是一个安全可信任的服务器，主要负责协调模型训练并进行模型聚合等工作。该架构的训练过程如下图所示，其主要包括以下四个步骤：

Step1：各客户端根据本地数据集训练模型，并将计算得到的模型梯度通过同态加密、差分隐私等加密技术加密后发送给中央服务器。

Step2：中央服务器对收到的模型梯度通过加权平均等方式进行梯度聚合。

Step3：中央服务器将聚合后的结果发送给各参与方。

Step4：各参与方将收到的聚合结果解密后用于更新各自的模型参数。

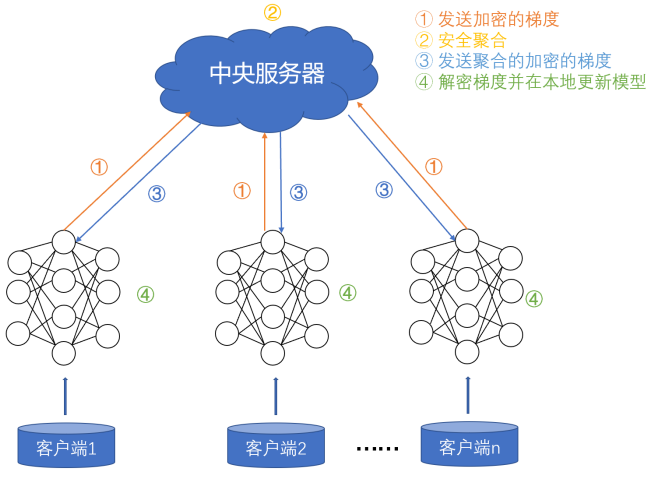

图 3-5 中央服务器-客户端架构

(2) 对等网络

在“对等网络”架构中，联邦学习参与方不再被分为中央服务器和客户端，而是每个参与方都作为训练方（trainer）参与训练，每个训练方只负责利用本地数据进行机器学习模型训练，之后使用公共密钥等加密措施对训练得到的模型参数或梯度进行加密，之后每个训练方之间通过安全链路（channels）交换计算结果，结构如下图所示。

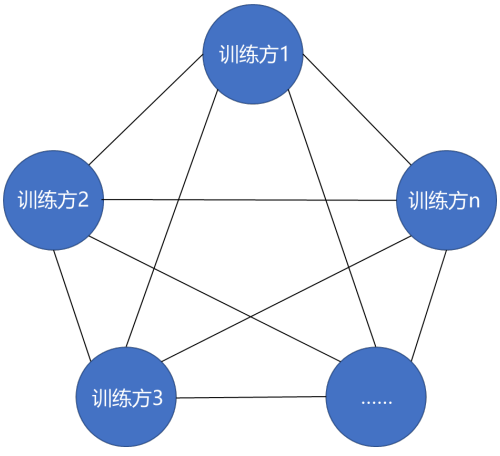

图 3-6 对等网络架构

(3) 联邦平均算法介绍

联邦平均算法（Federated Averaging，FedAvg）是一种横向联邦学习中“中央服务器-客户端”架构的经典算法，其于 2016 年被谷歌的 McMahan [错误!未找到引用源.] 等人提出，最初这个算法是用来防止手机用户在移动设备上更新模型时隐私被泄露的，该算法的伪代码如表 3-1 所示。

表 3-1: FedAvg 算法伪代码

算法 1 FedAvg 算法 K个客户端编号为k，B为本地批大小，E是本地迭代轮次，η是学习率。

中央服务器端：
初始化$w0$
for 每一通信迭代轮次$t = 1,2,\ldots$**do**
$m \leftarrow \max(C \cdot K, 1)$
$C_t \leftarrow$ 随机选取m个客户端组成集合C_t
for 每个客户端$k \in C_t$并行 **do**
$w_{t+1}^k \leftarrow$客户端基于(k, w_t)进行模型更新
$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$
客户端更新(k, w)：//在客户端 k 上运行
$\mathcal{B} \leftarrow$（将P_k分成大小为B的批）//P_k为客户端k的样本索引，其大小为n_k
for 每个迭代轮次i（1~E）**do**
for batch $b \in \mathcal{B}$ **do**
$w \leftarrow w - \eta \nabla l(w; b)$
将w上传给中央服务器

在 FedAvg 算法中，假设一共有K个客户端，其索引为0~k，该算法步骤为：

Step1：由中央服务器进行模型参数初始化，之后在每一轮通信迭代$t =$

1,2, …中都随机选取m个客户端参与训练，其中m（$1 \leq m < K$）为每轮选取的客户端个数，C_t为随机选取的m个客户端的集合。

Step2：由每轮中被选中的客户端使用本地数据对中央服务器下发的该轮模型w_t进行训练，得到更新后的模型参数w_{t+1}^k，并将其回传给中央服务器。

Step3：中央服务器对收到的模型参数加权平均进行聚合得到w_{t+1}，权重为$\frac{n_k}{n}$，其中n_k为客户端k的样本数量，n为本轮所有被选中客户端的总样本数量。

由于 Step2 中客户端下载中央服务器更新后的全局模型并向中央服务器上传其更新后的参数的过程需要迭代很多次才能完成最终模型的构建，在此过程中通信成本十分高昂，故 FedAvg 采用增加本地计算量（增加本地设备训练模型时的计算轮数）的方法来优化算法，提高通信效率。本地的计算轮数由C（每一轮参加训练的客户端比例）、E（每一轮每个客户端将全部样本训练一遍的次数）和B（每个客户端本地训练的批大小，$B = \infty$代表使用所有样本进行训练）。McMaha [错误!未找到引用源。] 等人证明了 FedAvg 算法不仅在独立同分布的数据上表现很好，在非独立同分布的数据上也能收敛，但收敛速度和模型精度较独立同分布数据有所下降。

3.2.3 迁移联邦学习

如图 3-7 所示，对于各个参与联邦学习的参与方上的本地数据，若其特征和样本都重叠地比较少，如在不同地区的银行和保险公司，则可以考虑用迁移学习的方法克服数据和标签不足的情况，这种在联邦学习框架下利用数据或任务之间的相似性进行模型训练的方法称为迁移联邦学习。迁移联邦学习不要求参与方之间的特征或样本相似，可以在任何数据分布、任何实体之间进行，通用性较强。

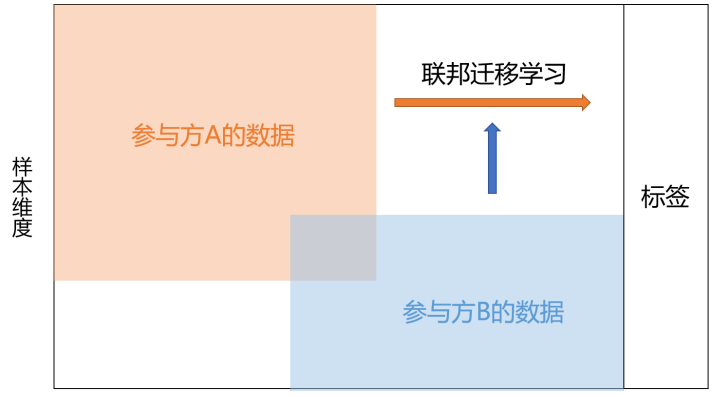

图 3-7 迁移联邦学习

迁移联邦学习中将参与方分为源域和目标域，其核心思想是：找到源域和目标域之间的数据或任务相似性，即隐层特征不变量，将源域学习的模型应用于目标域[57][58]。迁移联邦学习的架构如图 3-8 所示：

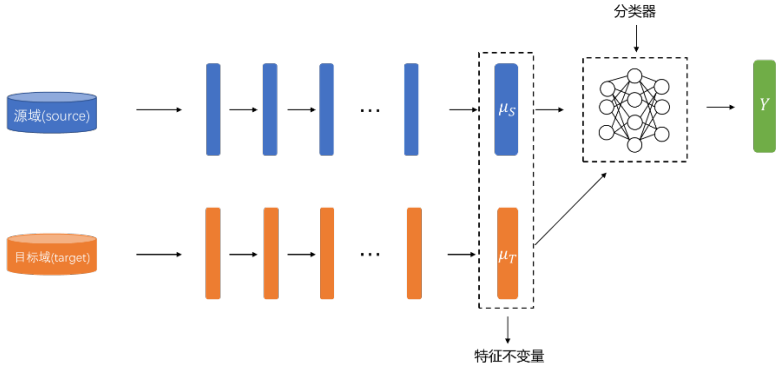

图 3-8 迁移联邦学习架构

假设源域和目标域之间存在共同样本通过一定策略对源域和目标域上的数据做变换，可以得到它们之间的隐层特征不变量μ_S、μ_T，设对目标域的分类器为$\varphi(\mu_i^T)$，Θ^S、Θ^T分别表示源域和目标域上的模型参数，λ为正则化参数，则迁移联邦学习的目标函数为：

$$min_{\Theta^S,\Theta^T} L = L_1 + \lambda L_2 + \frac{\lambda}{2}\left(\left\|\Theta^S\right\|^2 + \left\|\Theta^T\right\|^2\right) \tag{3-19}$$

其中$L_1 = \sum_i l_1\left(y_i^S, \varphi(\mu_i^T)\right)$表示分类器在源域上的标签预测损失，用于训练分类器的分类性能。$L_2 = \sum_i l_2(\mu_i^S, \mu_i^T)$表示$\mu_i^S, \mu_i^T$之间的差异损失，目的是学习源域和目标域之间的特征不变量。整个过程就是通过学习源域和目标域之间的特征不变量μ_i^S, μ_i^T，使得在目标域可以使用由y_i^S和μ_i^S训练的分类器。“迁移”体现在目标域迁移了源域的分类能力，“联邦”体现在整个过程不

需要源域和目标域的本地数据参与，仅需交换中间的计算结果。

接下来对迁移联邦学习的经典算法 FADA 进行介绍，无监督域适应(Unsupervised Domain Adaptation, UDA)指的是将从有标签的源域上学到的知识迁移到无标签的目标域上[59]，在此过程中，若源域的标签分布和目标域的标签分布相差很大，即发生“域偏移”现象，很可能导致源域数据的学习对目标域学习分类器产生干扰或抑制的作用，这种问题被称为“负迁移”。为缓解负迁移问题，Peng[41]等人提出了一种基于对抗域适应技术的迁移联邦学习框架 FADA，该算法利用对抗技术解决域偏移问题，并通过特征解耦进一步提高模型的性能，接下来以训练 k 类别分类器为例对该算法进行介绍。

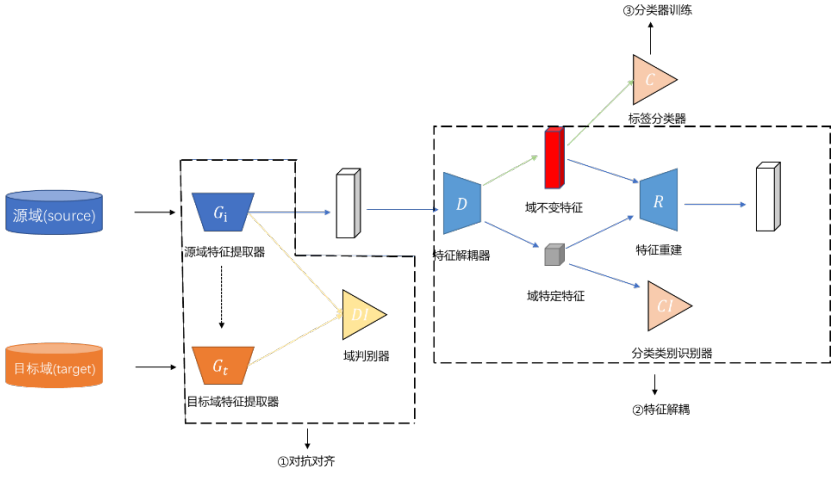

图 3-9 FADA 模型框架

如图 3-9 所示，该算法过程可被分为对抗对齐、特征解耦和分类器训练三个部分。在对抗对齐部分，采用对抗生成的思想来训练源域和目标域的特征提取器，使得提取后的特征非常相似，可以缓解域偏移问题。该方法分为两个步骤：(1)在源域和目标域上分别训练对应的特征提取器G_S、G_t。(2)训练一个域判别器DI以对抗性的方式来对齐分布：首先训练DI来识别经过特征提取器变换的特征$G_S(X_s)$、$G_t(X_t)$来自于哪个域，之后训练G_S、G_t，使得它们非常相似，从而混淆域判别器DI。即：

Step1：固定特征提取器G_S、G_t，训练域判别器DI。

$$\min_{D} L_{adv}(X_s, Y_s, G_s, G_t) = -E_{x_s \sim X_s}\left[logDI\left(G_S(X_s)\right)\right] - E_{x_t \sim X_t}\left[log\left(1 - DI\left(G_t(X_t)\right)\right)\right] \tag{3-20}$$

Step2：固定域判别器DI，训练特征提取器G_S、G_t。

$$\min_G L_{adv}(X_s, Y_s, D) = -E_{x_s \sim X_s}\left[logD\left(G_S(X_s)\right)\right] - E_{x_t \sim X_t}\left[logD\left(G_t(X_t)\right)\right] \quad (3\text{-}21)$$

交替训练以上两步直至模型收敛。

同样地，在特征解耦部分也是采用对抗性生成的方法来提取域不变特征，从而去除无关的以及域特定的特征，只使用对影响分类任务的特征进行建模，提高模型的准确性。该方法也可以分为两个步骤：

Step1:基于域不变特征$f_{\mathrm{d}} = D\left(G(\boldsymbol{X_s})\right)$和域特定特征 $f_{ds} = D\left(G\left(X_s\right)\right)$训练 k 类别分类器C_i 和类识别器CI对标签进行预测，通过计算交叉熵损失来实现目标。

$$Lcross - entropy = -\mathrm{E}_{(x^s, y^s) \sim D_s} \sum_{k=1}^{K} 1[k = y^s]\, log\left(C(f_{\mathrm{d}})\right) - \mathrm{E}_{(x^s, y^s) \sim D_s} \sum_{k=1}^{K} 1[k = y^s]\, log\left(CI(f_{\mathrm{ds}})\right) \quad (3\text{-}22)$$

Step2:固定类识别器CI不动，通过训练特征解耦器D生成域特定特征f_{ds}来混淆类识别器CI，并通过最小化类分布的负熵损失来实现目标。

$$L_{ent} = -\frac{1}{N_s}\sum_{j=1}^{N_s} log\left(CI\left(f_{ds}^{j}\right)\right) = -\frac{1}{N_s}\sum_{j=1}^{N_s} log\left(CI\left(D\left(G(X_s)\right)\right)\right) \quad (3\text{-}23)$$

特征解耦通过保留f_{d}并消除f_{ds}来促进知识迁移。在分类器训练部分，就是通过生成的域不变特征f_{d}来训练分类器，最后将训练好的分类器应用于目标域。

3.2.4 纵向联邦学习

对于各个参与联邦学习的参与方上的本地数据，若满足数据特征重叠少而样本重叠多的条件，如在保险公司与银行，其用户可能是相似的（样本空间相似），而业务不同（特征空间不同），则可以将各参与方的数据集按照纵向（特征维度）进行切分，并联合样本相同而数据特征不同的部分进行训练，这种将数据集进行纵向切分进行训练的方法叫做纵向联邦学习，示意图如下。

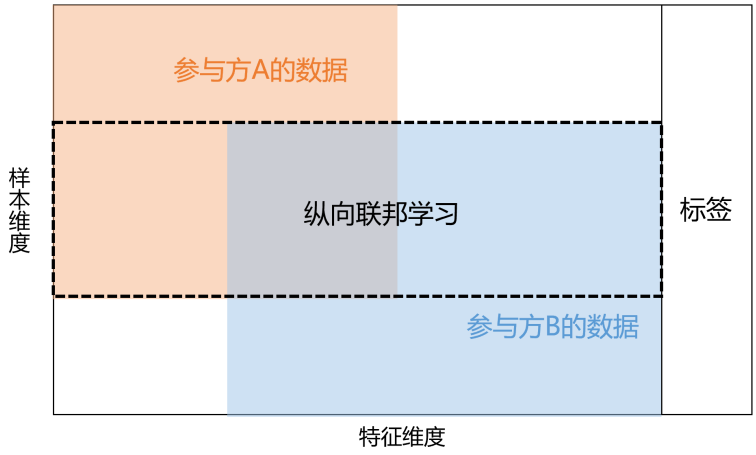

图 3-10 纵向联邦学习

纵向联邦学习由若干个参与方和一个协调者 C 构成，假设协调者 C 是成实且不与参与方共谋的，而参与方是诚实和好奇的，则纵向联邦学习的架构可以由下图表示：

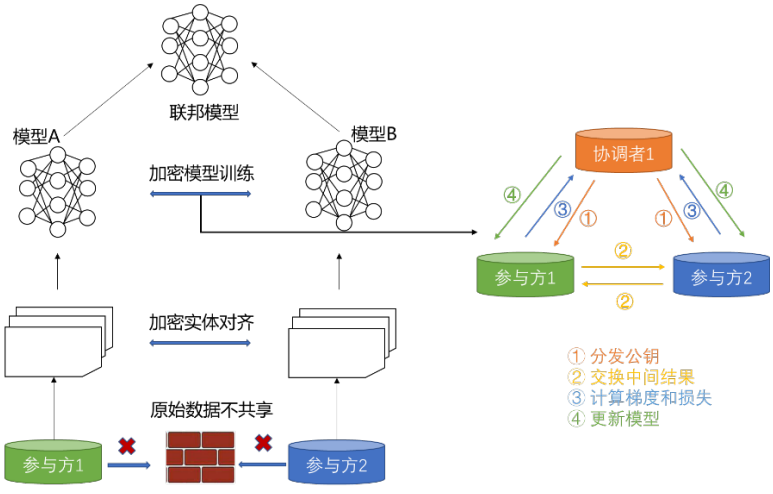

图 3-11 纵向联邦学习架构

该架构的训练过程主要分为加密实体对齐和加密模型训练两个部分，在加密实体对齐部分，各参与方的共同用户通过基于加密的用户 ID 对齐技术进行对齐，该技术可以保证在用户 ID 对齐的过程中，各参与方的本地数据不被泄露。在加密模型训练部分，各参与方主要通过以下几个步骤进行训练：

Step1：由协调者将公共密钥发送给参与方。

Step2：参与方将计算结果进行加密后交换。

Step3：参与方基于的都的结果计算加密梯度并加上随机掩码，有标签的

参与方还将计算加密损失，最后各参与方将梯度与损失加密后发送给协调者 C。

Step4：协调者 C 将收到的计算结果解密并回传给参与方，各参与方根据破解随机掩码后的梯度来对本地模型进行更新。

3.3 本章小结

本章首先对信用风险相关理论及技术进行介绍，介绍了逻辑回归、集成学习、神经网络三种常用的信用风险管理方法，并对其代表算法进行详细介绍。之后，对联邦学习技术分类进行介绍，对本文涉及的横向联邦学习和迁移联邦学习进行详细阐述，并对其经典算法进行较少。同时，对本文未涉及的纵向联邦学习进行简要概述，为本文的研究内容夯实了理论基础和技术基础。

4.研究设计

本文主要研究在联邦学习框架下，如何更好地利用大量已知标签的源域数据集S_i帮助拥有少量已知标签的目标域数据集t进行有监督域适应学习，提升目标域分类器的分类效果。本文的研究从金融信贷领域的实际需求出发，受到 FADA 迁移联邦学习框架的启发，提出了一种新的联邦对抗域适应算法 P2FADA，创新性地对 FADA 算法进行改进并应用于解决信贷产品冷启动问题。为了刻画 P2FADA 模型在不同信贷违约预测任务上的准确性和稳定性，由于仿真数据可控性强，可以根据不同信贷场景生成相应数据，本文采用数据模拟的方式生成数据，分别使用人工智能神经网络 ANN、典型的二分类模型逻辑回归、常用的机器学习模型 XGBoost 和经典的联邦学习算法 FedAvg 作为基模型进行不同实验，来评价本文提出的 P2FADA 算法在信用贷款产品投产初期风险评估的能力。

4.1 研究思路

信贷场景下风控系统的冷启动问题是指创新型信用贷款产品在投产初期缺乏有标签样本数据，难以训练出有效模型来进行风险管理。基于此问题，本文提出联邦对抗域适应算法 P2FADA 来联合多方数据建模，提升模型表现。下面叙述该算法与传统解决方法相比，在信贷场景中冷启动问题上的适用性和有效性。

由于创新信用贷款产品投产初期数据不足，我们可以想到使用迁移学习方法，利用相似信用贷款产品的丰富数据来辅助信贷违约预测模型训练。然而，由于数据安全和隐私保护相关的立法限制，不同银行等金融机构之间的数据无法交流互通，形成了“数据孤岛”现象。故常用的分类模型 ANN、逻辑回归和 XGBoost 在此场景中应用数据受到限制，只能使用单一源域的数据训

练分类器，不仅无法联合多方源域建模，还浪费了已有的目标域数据资源。

联邦学习为解决此问题提出了新的思路，它是一种分布式机器学习框架，可以在保护数据隐私的情况下联合多方数据构建模型，且该模型与传统训练方式下得到的模型效果基本一致，能很好解决“数据孤岛”问题。FedAvg 就是一种经典的联邦学习算法，它可以在保护数据隐私的前提下利用源域和目标域数据建模，充分利用已有数据资源。但在金融信贷场景中，客户的“守约”与“违约”表现是“相对”的，即客户在 A 机构“守约”，但有可能在 B 机构发生“违约”行为。此时使用 FedAvg 算法将基于各域训练的分类模型进行简单平均聚合，得到的全局模型效果非但不会提升，反而会变差，即发生迁移学习的负迁移现象。这是因为 FedAvg 算法没有考虑到“域偏移”问题，即当源域的数据分布和目标域的数据分布相差很大时，利用源域数据训练模型可能会对目标域学习分类模型产生干扰或抑制的作用，从而发生负迁移现象。Long [错误!未找到引用源。] 等人通过此现象研究得出机器学习模型的性能会随着域之间的差异增大而下降的结论。为了解决此问题，Hoffman[39]等人提出用对抗域适应技术来缓解负迁移现象，一方面，域适应算法可以在保持源域分类模型精度的前提下，缩小表示空间上源域与目标域的模型特征距离，使得基于源域训练的模型可以直接应用于目标域。另一方面，对抗学习与非对抗学习相比，优势在于可以学习足够好的特征表示，帮助缩小源域与目标域的特征距离，故对抗学习与域适应技术的结合，可以最大化缓解源域数据分布与目标域数据分布不一致带来的负迁移问题，从而提高了模型泛化性能。

然而，对抗域适应技术需要同时访问源域和目标域数据，无法解决现实情况中各银行之间数据无法交流的问题。Peng[41]等人提出了联邦对抗域适应算法 FADA，该算法将联邦学习框架与对抗域适应技术结合在一起，可以在一定程度上保护数据隐私，解决域偏移问题。

由上文对 FADA 算法的介绍可知，在使用对抗学习进行源域与目标域特征表示对齐时，各源域为了训练判别器，会将经过特征提取器变换的特征$G_S(X_s)$、$G_t(X_t)$上传至中央服务器，虽然此时数据已经经过变换，但仍然有被解码的风险，从而造成数据隐私泄露，危害数据安全。同时，经过特征提取器变换的特征$G_S(X_s)$、$G_t(X_t)$维度与 batch size（批数据量）大小高度相关，