编号 ××××××××

南京航空航天大学

本科毕业设计（论文）

题　目　　联邦学习本地模型贡献度评估机制

学生姓名	刘星麟
学　　号	162120317
学　　院	计算机科学与技术学院
专　　业	信息安全
班　　级	1621204
指导教师	周璐教授

二〇二五年五月

南京航空航天大学

本科毕业设计（论文）诚信承诺书

本人郑重声明：所呈交的毕业设计（论文）是本人在导师的指导下独立进行研究所取得的成果。尽我所知，除了文中特别加以标注和致谢的内容外，本设计（论文）不包含任何其他个人或集体已经发表或撰写的成果作品。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

作者签名：

日　　期：　20　　年　　月　　日

南京航空航天大学

毕业设计（论文）使用授权书

本人完全了解南京航空航天大学有关收集、保留和使用本人所送交的毕业设计（论文）的规定，即：本科生在校攻读学位期间毕业设计（论文）工作的知识产权单位属南京航空航天大学。学校有权保留并向国家有关部门或机构送交毕业设计（论文）的复印件和电子版，允许论文被查阅和借阅，可以公布论文的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编论文。保密的论文在解密后适用本声明。

论文涉密情况：

☐ 不保密

☐ 保密，保密期（起讫日期：　　）

作者签名：　　导师签名：

日　　期：　20　　年　　月　　日　　日　　期：　20　　年　　月　　日

摘　　要

联邦学习是一种允许参与多方在保留本地原始数据的情况下协同训练全局模型的分布式机器学习范式，但数据异构性问题给公平、准确地评估各参与客户端对全局模型的贡献带来了巨大挑战。现有方法（如 Shapley 值、模型更新分析）存在主观假设、高计算成本或噪声敏感性等局限性问题，因此如何在高异质性场景下实现高效、鲁棒的贡献度评估成为了推动联邦学习广泛应用的关键挑战。

本文提出了一种基于中心核对齐的联邦学习本地模型贡献度评估机制，针对传统方法依赖参数更新或浅层特征的局限性引入了中心核对齐算法，通过计算本地模型与全局模型在神经网络中间层特征空间的相似性，捕捉深层知识迁移的本质贡献。基于该方法，本文提出多层特征融合策略并结合线性核的中心核对齐的数学优化与无偏估计方法，实现了兼具鲁棒性和高效性的中心核对齐计算。进一步设计动态闭环评估框架，将中心核对齐算法嵌入联邦学习的流程中，从而精准地量化了异构环境下的客户端的实质贡献。

通过在 MNIST、CIFAR-10 和 CIFAR-100 数据集以及 SimpleCNN、ResNet18 模型上，从准确率、排名稳定性和抗噪性等不同角度进行分析，实验结果表明本文提出的基于中心核对齐算法的贡献度评估机制能够有效、稳定地区分在数据异构环境下不同数据质量客户端的贡献。此外，本文验证了 CKA 在贡献度评估中相较于 KL 散度、余弦相似度的优越性。本研究为联邦学习贡献度评估提供了一种高效公平的方法，在激励机制设计中具有广泛应用潜力。

关键词：联邦学习，贡献度评估，中心核对齐，数据异构性，模型表示相似性

ABSTRACT

Federated learning is a distributed machine learning paradigm that allows multiple parties to collaboratively train a global model while retaining their local raw data. However, the issue of data heterogeneity poses significant challenges to fairly and accurately evaluating the contributions of each participating client to the global model. Existing methods, such as Shapley values and model update analysis, have limitations such as subjective assumptions, high computational costs, or sensitivity to noise. Therefore, how to achieve efficient and robust contribution evaluation in high-heterogeneity scenarios has become a key challenge in promoting the widespread application of federated learning.

This paper proposes a federated learning local model contribution evaluation mechanism based on Center Kernel Alignment (CKA). To address the limitations of traditional methods that rely on parameter updates or shallow features, we introduce the CKA algorithm. This method calculates the similarity between the local model and the global model in the intermediate feature space of the neural network, capturing the intrinsic contributions of deep knowledge transfer. Based on this approach, we propose a multi-layer feature fusion strategy and combine the mathematical optimization and unbiased estimation method of CKA with a linear kernel, achieving both robustness and efficiency in the CKA computation. Furthermore, we design a dynamic closed-loop evaluation framework, embedding the CKA algorithm into the federated learning process to accurately quantify the substantial contributions of clients in heterogeneous environments.

Through experiments on the MNIST, CIFAR-10, and CIFAR-100 datasets, as well as on SimpleCNN and ResNet18 models, the results from different perspectives such as accuracy, ranking stability, and noise resistance show that the proposed contribution evaluation mechanism based on the CKA algorithm can effectively and stably distinguish the contributions of clients with different data qualities in heterogeneous environments. Additionally, this paper verifies the superiority of CKA in contribution evaluation compared to KL divergence and cosine similarity. This study provides an efficient and fair method for federated learning contribution evaluation, with broad application potential in incentive mechanism design.

Keywords: Federated Learning, Contribution Assessment, Centered Kernel Alignment (CKA), Data Heterogeneity, Model Representation Similarity

目 录

第一章 绪论

1.1 引言

联邦学习(Federated Learning, FL)是一种保护数据隐私的分布式机器学习范式，它的发展为解决数据孤岛问题提供了创新性思路。但是实际应用中广泛存在的数据异构性（Non-Independent and Identically Distributed, Non-IID）问题严重影响了客户端贡献度的公平评估，进而制约了联邦学习的激励机制设计和模型性能优化。本章从联邦学习的背景与挑战出发，系统梳理现有贡献度评估方法的局限性并明确本文的研究目标与创新方向。

1.2 研究背景与意义

联邦学习[1]是一种机器学习的分布式模式，在这种模式中，中央服务器（例如服务提供商）协调客户端（例如移动设备、边缘节点和组织）共同训练模型。联邦学习具有去中心化[2]的特点，在模型训练的过程中，服务器无需集中各客户端的本地敏感数据，即模型训练数据分散存储在客户端本地，客户端只需向服务器发送模型更新（如梯度更新等），这些更新相比原始数据包含的用户信息更少，从根本上降低了敏感数据暴露的可能性。联邦学习这种去中心化的特点决定了它能够解决人们对数据隐私保护的需要，使得它在医疗健康[3]、网络安全[4]以及物联网[5]等领域具有广泛的应用前景。

然而联邦学习在实际的大规模应用中通常面临着参与客户端之间数据分布、模型架构和网络环境等一系列的异构性问题[6]。数据异构性问题（即 Non-IID）是其中最为突出的问题：联邦学习中各客户端的数据分布不一致并且不服从相同采样的情况，它具体分为标签偏斜、特征偏斜、质量偏斜和数量偏斜。联邦学习训练在 Non-IID 数据的影响下会出现客户端的局部优化目标和全局模型的优化目标不一致的问题，这种情况会致使联邦学习的性能降低。异构性问题为联邦学习中的公平性带来了巨大的挑战[7]，诸多联邦学习的研究文献普遍认为公平性是关键问题之一，谷歌的 P. Kairouz 和 B. McMahan 在 2021 年的一篇评论[8]中将公平性确定为未来联邦学习研究的关键领域。

各客户端在联邦学习框架中使用私有的本地数据训练通用模型并通过共享本地模型更新来做出贡献，对这些参差不齐的贡献进行量化评估的过程称为贡献度评估[9]，贡献度评估在促进联邦学习的公平性具有举足轻重的作用。它在激励机制、模型优化和异常检测等方面发挥着关键作用：通过评估参与方的贡献度并建立有效激励机制能够显著提高客户端使用本地数据训练模型的积极性；通过筛选贡献度高的参与客户端可以引导模型训练资

源的合理分配并提高全局模型的性能；贡献度评估可以通过识别异常的贡献来减少损害联邦学习系统的公平性和有效性的行为。

1.3 联邦学习模型的现状

数字时代的飞速发展使得数据呈现爆炸式增长，由此引发的隐私风险也日益凸显。各国政府和组织纷纷出台日益严格的隐私保护法律法规，例如欧盟的《通用数据保护条例》(GDPR)和中国的《个人信息保护法》,这些法规的实施极大地推动了隐私保护需求的发展。在这种背景下联邦学习得到了广泛的应用[10]，它的核心优势是能够在保护用户数据隐私的前提下允许多个参与方协同训练共享的机器学习模型。联邦学习有效解决了数据孤岛的问题并推动了人工智能在金融[11]、医疗[3]、物联网[5]等多个领域的应用与发展，当前联邦学习模型的研究工作已经具有了显著的进展：基础理论、核心算法（如 FedAvg 及其变种[12]）、系统架构和安全隐私增强技术等方面不断取得新的发展[13]。近年来，联邦学习模型在理论研究和工程实践层面均取得了突破性的进展，其技术演进呈现出多维度的创新与跨领域融合的特征，主要的发展态势可归纳为以下五个核心方向。

1.3.1 模型架构创新与个性化演进

个性化联邦学习（Personalized Federated Learning, PFL）已然成为了破解联邦学习中数据异构性难题的关键研究方向。Dinh 等人[14]提出的 FedPer 框架通过参数解耦机制将模型划分为全局共享层与个性化私有层，该框架相比传统方法在医疗影像诊断场景中实现了15%-20%的准确率提升。Li 等学者[15]开发了动态模型缩放技术进一步提出一种简单通用的个性化联邦学习框架 Ditto，通过平衡全局模型与本地模型的协同优化，解决统计异构网络中公平性（跨设备性能一致性）与鲁棒性（抵御数据 / 模型中毒攻击）的矛盾。此外元学习与联邦学习的深度融合催生了 FedMeta 框架[16]，该框架在 CIFAR-100 数据集上通过元知识迁移机制仅需 5 轮通信即可达成 80%的收敛准确率，显著优于传统联邦平均算法。

1.3.2 通信效率优化与资源约束突破

Konečný 团队[17]面向边缘计算场景中的通信瓶颈问题提出了深度梯度量化方法，他们创新性地采用了 8 位自适应量化策略，该方法能够将通信带宽需求压缩至 FedAvg 基准的 1/6 并在移动端推荐系统中实现 40%的收敛加速。Qu 等人[18] 针对联邦学习通信效率优化的需求提出了 FedQClip 框架，通过结合量化（Quantized）与剪裁（Clipped）随机梯度下降（SGD）有效解决了传统联邦学习中通信开销大、收敛速度慢的问题。尤其是该方法在 Non-IID 数据分布下显著优于现有方法，为联邦学习的高效训练提供了理论和实践支持。

1.3.3 安全隐私增强与防御机制进化

作为联邦学习的重要优势之一的隐私保护也是研究的热点，差分隐私、同态加密、安全多方计算等安全技术被广泛应用于联邦学习中以进一步增强数据隐私保护[19]。Hu 等人首次系统地验证了中心化差分隐私（CDP）和本地差分隐私（LDP）在梯度泄露攻击下的实际效果[20]，并揭示 LDP 在非独立同分布（Non-IID）数据下隐私保护优势显著，但模型的性能在该方法中下降了 20%-35%。Zhang 等人[21]指出当前的防御机制依赖大量有标签干净数据及导致的性能下降并提出 BadCleaner 方案，该方案借助基于注意力的联邦多教师蒸馏技术，一方面利用少量无标签干净数据从多教师模型提取知识来调整后门联合模型，保证性能无损；另一方面通过注意力转移方法减少模型对触发区域的关注，从而消除隐藏后门模式。

1.3.4 跨模态融合与多源知识迁移

多模态联邦学习正成为医疗、自动驾驶等复杂场景的研究热点。最新研究中的自适应超图聚合框架[22]为多模态联邦学习提供了高效、鲁棒的解决方案，尤其在模态异质性和统计偏移场景中表现优异。该框架的目的是解决多模态联邦学习中模态不兼容和统计异质性挑战，实现模态无关的自适应聚合。此外最新提出的 MLA - BIN 方法[23]通过设计模型层面的注意力模块（MLA）以线性组合方式让全局模型泛化到未见域，以及批实例风格归一化（BIN）块解决域间图像风格差异对域泛化的影响，从而解决联邦学习在医学图像分割中的域泛化问题，实验证明该方法优于现有技术。

1.3.5 边缘智能融合与轻量化部署

物联网的发展以及相关设备的指数级增长推动了轻量化联邦学习的技术革新。Wu 等人[24]提出了一种端边云协同的联邦学习框架（FedAgg），该框架通过递归组织计算节点和桥接样本在线蒸馏协议（BSB ODP）解决了传统联邦学习中模型规模受限于终端设备的问题。该方法有效解决了联邦学习中模型规模受限和数据异质性问题，在准确率、收敛速度和通信效率上均显著优于现有方法，为大规模分布式 AI 模型训练提供了新范式。

1.4 联邦学习贡献度评估研究现状

联邦学习同人工智能一起展现出快速的发展速度，它作为一种能够保护数据隐私的分布式机器学习方法正在逐渐展现出了其独特的优势，但是联邦学习在机器学习中的广泛应用导致其面临的贡献度评估挑战也越发引人重视。联邦学习贡献度评估的目的是公平、准确地量化每个参与方（如客户端）在模型训练过程中的贡献大小。

现有的联邦学习贡献度评估研究主要为基于数据估值、合作博弈以及模型更新的三种方法。基于数据估值的方法的核心是量化每个参与方的数据价值，以此来评估其对联邦学习模型训练的贡献；基于合作博弈的方法将联邦学习中的贡献度评估视为一个合作博弈问题，通过博弈论中的公平分配方法（如 Shapley 值）来评估每个参与方的贡献；基于模型更新的方法则通过分析每个参与方上传的模型更新来评估其贡献。本小节主要介绍现有的主流联邦学习贡献度评估方法。

1.4.1 基于数据估值的方法

数据估值指标是衡量参与方的贡献度的基础，它的核心是对参与方的数据价值进行量化，以此来评估各参与方对联邦学习模型训练的贡献。数据估值指标需要对数据的质量、数量和对模型性能的影响进行合理评估，它又分为两类：测试集依赖指标（如准确率、R^2）和测试集无关指标（如统计指标、模型参数指标）[25]。Wenqian Li 等人[26]提出了一种名为 FedBary 的联邦学习数据评估和检测方法，利用 Wasserstein 距离，在无需共享原始数据和预先指定训练算法的情况下，实现客户端贡献评估。但这种方法存在特定假设和超参数依赖问题，且具有应用场景局限性。Hao Wu 等人[27]提出了一种名为 CoAst 的联邦学习贡献评估方法，通过参数修剪和跨轮估值两项关键技术完成贡献度评估，但模型修剪会使模型收敛时间变长。

该方法的优势体现在直接从数据本身出发，且能比较直观地反应数据对模型训练的潜在价值。但是，数据估值存在主观因素影响大、依赖实际场景和恶意数据干扰等问题[25]。

1.4.2 基于合作博弈的方法

基于合作博弈的方法将联邦学习中的贡献度评估看作一个合作博弈问题，通过博弈论中的公平分配方法来评估每个参与方的贡献。该方法普遍使用合作博弈中的夏普利值(Shapley Value, SV)来评估参 与方在全局模型训练中做出的贡献，这种方法能够实现参与方一致认可的博弈平衡，完成公平的贡献度评估。

Peng Guo 等人[28]提出了一种基于 Shapley 值的贡献度评估方法，通过梯度复用避免额外模型训练，节省计算资源；引入新的聚合权重减轻数据分布异质性对贡献测量的影响，提高测量准确性。Nurbek Tastan 等人[29]提出的 ShapFed 算法利用合作博弈论中的 Shapley 值评估参与者对全局模型的贡献，该方法通过计算类特定 Shapley 值(CSSV)来细化评估。然而计算 Shapley 值的时间复杂度通常较高并且 Shapley 值的计算需要假设所有参与方的数据是独立同分布的，但是现实情况中 Non-IID 的存在是不可忽视的[25]。

1.4.3 基于模型更新的方法

基于模型更新的方法通过分析每个参与方上传的模型更新的大小、方向和质量来评估其贡献，因为模型的更新反映了参与方在本地训练模型时对全局模型的贡献。例如，CoAst方法[27]提出了一种跨轮次的评估机制，这种机制通过比较当前轮次的本地模型参数与后续几轮全局模型参数更新的相似性来评估贡献。该方法考虑了模型参数更新的长期影响，能够更准确地反映各参与方的贡献。

KL 散度[30]和余弦相似度[31]也是该类方法中常用的贡献度评估方式。例如 Zhijie Xie 等人[32]针对联邦强化学习(FRL)中的数据异质性问题提出了 FedKL 算法，该算法通过引入KL散度惩罚项约束本地策略与全局策略的差异。

这种方法不需要额外的验证数据集，因此在实际应用中更加灵活。然而由于模型训练过程中的随机性（如随机梯度下降的随机性），基于模型更新的评估方法可能会受到一定干扰导致评估结果不够准确[25]。

1.5 本文工作

在联邦学习实际应用中，数据拥有者往往不一定具有提供私有数据进行训练的动力，因此需要对其进行激励，最简单的激励办法就是经济激励。然而，由于不同数据通过训练后，对全局模型的贡献不尽相同，因此需要对不同贡献者的贡献度进行有效的评估，才能保证激励的公平性和合理性，从而更好地激励数据拥有者参与联邦学习训练。针对现有的联邦学习客户端贡献度评估机制的一些不足，本文设计了基于中心核对齐算法(Centered Kernel Alignment，CKA）的联邦学习本地模型贡献度评估机制。本文的主要工作内容如下：

- 联邦学习框架设计与实现。本文首先设计并实现了一套基于 PyTorch 的联邦学习框架，并采用经典的联邦平均(FedAvg)算法作为模型聚合方法。在架构上构建了完整的服务器-客户端结构，实现了参数分发、本地训练和全局聚合的核心流程并开发了灵活的配置系统。在数据集的选择上，分别使用 MNIST、CIFAR-10 和 CIFAR-100 数据集进行模型训练。
- 基于 CKA 的贡献度评估机制设计。通过线性核 CKA 算法和神经网络中间层特征提取策略构建了服务器-客户端间的特征表示相似度计算流程，并针对高维特征处理的计算效率进行优化，有效降低了计算资源的消耗，为联邦学习中的公平激励机制提供了技术基础。
- 数据异构性模拟实现。为验证所提出评估机制的有效性，本文设计并实现了三种

典型的数据划分策略：IID（独立同分布）、Non-IID（非独立同分布）和带噪声的数据，通过参数化的划分控制方法实现了不同程度的数据偏移配置，并实现了类别不平衡的 Non-IID 设置，更真实地模拟了实际应用场景中的数据分布情况。

- 对照实验设计与分析。本文设计一系列的对照实验，以此全面地评估了 CKA 算法的优越性。通过设置 CKA 不变性检验实验验证了 CKA 对正交变换的不变性；实现噪声敏感性测试用于评估 CKA 对高斯噪声的鲁棒性。

1.6 论文结构

本节主要阐述论文的结构安排，本文研究的问题、内容和目标如图 1.1 所示，本文的组织结构安排如下：

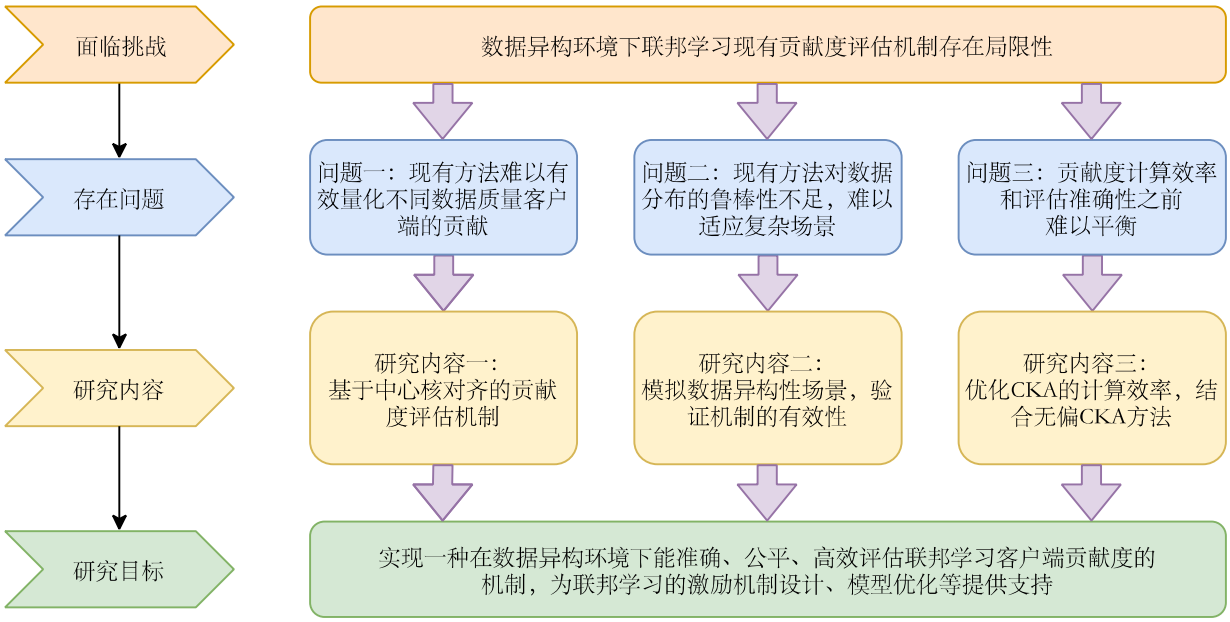

图 1.1 全文框架图

第一章：绪论。本章主要介绍了联邦学习的背景、研究意义以及阐述数据异构性给贡献度评估带来的挑战，同时本章梳理了现有的联邦学习贡献度评估方法的研究现状及其优缺点并明确本文的研究目标和主要工作内容。

第二章：相关理论与技术基础。该章节主要介绍联邦学习的基本概念、工作流程以及核心聚合算法——联邦平均算法。同时本章介绍模型/表示相似性度量的相关理论，其中重点阐述中心核对齐的方法的原理、数学推导、关键性质（如正交不变性、各向同性缩放不变性）以及无偏 CKA(Debiased CKA)的概念，为后续机制设计奠定理论基础。此外本章简要介绍其他相关的相似性度量方法，如余弦相似度和 KL 散度。

第三章：基于 CKA 的联邦学习贡献度评估机制设计。本章是论文的核心设计部分。首先在本章明确论文机制的设计目标与原则，然后给出系统的总体框架设计。接着，详细地阐述联邦学习基础模块（包括服务器端、客户端的设计以及数据异构性的模拟实现方法）和核心的 CKA 贡献度计算模块的设计细节，包括特征提取策略（利用 PyTorch 前向钩子）、CKA 计算的具体实现（包括标准 CKA 的优化计算和无偏 CKA）、贡献度的量化方法（多层特征融合）以及 CKA 计算在联邦学习流程中的嵌入位置。

第四章：实验设置与环境。这章详细地说明验证所提机制有效性的实验配置，它的内容包括所使用的数据集(MNIST、CIFAR-10、CIFAR-100)及其介绍、数据预处理与划分策略、数据异构场景（数据量不均衡、数据噪声、标签偏斜）的具体构建方法、选用的神经网络模型(SimpleCNN、ResNet18)、联邦学习和 CKA 计算的关键参数设置以及用于评估联邦学习性能和贡献度评估效果的评价指标体系。

第五章：实验结果与分析。 本章展示实验结果并深入分析，本章首先通过全局模型和本地模型的准确率/损失曲线验证所实现的联邦学习框架的有效性。然后本章通过重点分析 CKA 贡献度评估机制在不同数据异构程度下的表现验证其区分不同质量客户端贡献的有效性，并通过排名热力图等方式分析评估结果的稳定性。本章最后通过不变性检验和噪声敏感性测试实验，对比分析 CKA 方法相较于 KL 散度、余弦相似度等方法的特性和鲁棒性。

第六章：总结与展望。 本章对全文的研究工作进行全面总结、说明本文的主要创新点、客观分析当前研究存在的不足之处，并对未来可能的研究方向（如扩展应用领域、优化计算效率、去中心化评估、与激励机制结合等）进行展望。

1.7 本章小结

本章阐述了联邦学习的核心价值与数据异构性带来的公平性挑战以及总结了基于数据估值、合作博弈和模型更新的贡献度评估方法的特点和不足，并指出利用中心核对齐算法从模型表示相似性角度量化贡献的可行性。后续章节将围绕 CKA 的理论基础、机制设计及实验验证展开，为联邦学习的公平性研究提供新视角。

第二章 相关理论与技术基础

2.1 引言

联邦学习的核心优势是分布式协作和隐私保护，贡献度评估的关键则在于量化客户端对全局模型的实质性影响。本章详细地介绍联邦学习的核心算法（如 FedAvg）与流程、重点解析模型表示相似性度量方法，尤其是中心核对齐的数学原理、计算优化及特性，为后续机制设计奠定理论基石。

2.2 联邦学习

2.2.1 联邦学习简介

数据在数字化的浪潮中已然成为驱动人工智能发展的核心要素，然而数据孤岛和隐私安全问题在当今愈发的严重，这限制了人工智能的进一步发展。正是在这种背景下，联邦学习作为一种分布式的机器学习模式成为了一种新的选择：在保障数据和隐私安全的前提下，它允许在多个参与方或者计算结点之间协同训练机器学习模型。

联邦学习是一种分布式的机器学习框架，参与的多个客户端（从移动设备到企业）在中央服务器的组织下协作训练模型，但仍然保持训练数据的去中心化[33]。在传统的集中式机器学习中，需要将所有的数据整合得到一个数据集，以此训练机器学习模型。但在联邦学习中，各参与方不需要上传本地数据，只需使用本地数据训练模型，上传模型训练后的参数更新即可，由中央服务器聚合参数更新得到全局模型。联邦学习的训练流程如图 2.1 所示，具体如下：

- 模型初始化：中央服务器初始化全局模型并将此初始模型发送给各个参与训练的客户端。
- 客户端训练：客户端接受中央服务器发送的初始全局模型后利用本地数据对模型进行训练。
- 模型上传：客户端完成本地训练后将训练产生的模型参数或更新信息上传至中央服务器，这些上传的信息仅包含模型参数的变化。
- 模型聚合：中央服务器负责收集来自各个客户端的模型更新信息并运用特定的算法对这些信息进行聚合，如联邦平均算法。
- 模型更新：中央服务器将聚合后的新模型参数下发至各个客户端，客户端据此更新本地模型，之后开始下一轮的训练。如此循环往复，直至模型达到预设的收敛

条件。

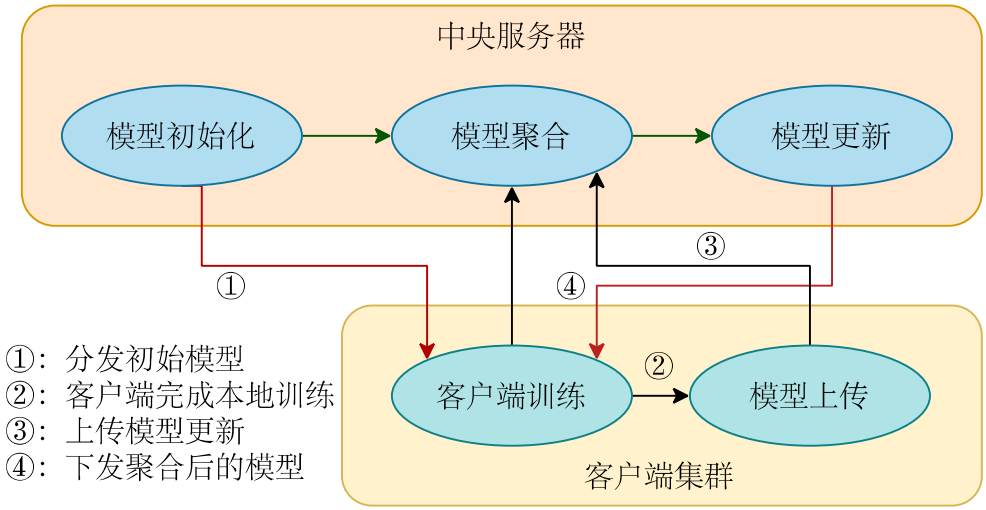

图 2.1 联邦学习流程

2.2.2 联邦平均算法

中央服务器聚合模型最常使用的算法为 H. Brendan McMahan 等人[34]提出的联邦平均算法，其是一种基于联邦学习场景设计的高效优化算法，适用于在去中心化数据源上训练深度模型：通过迭代模型平均的方式利用分布式设备上的数据训练共享模型。诸多实验和实际使用证明联邦平均算法相比传统的同步随机梯度下降(FedSGD)显著提高了通信效率，且对数据分布的不平衡和非独立同分布具有良好的鲁棒性，在实际的应用中具有较大的潜力。

具体来说，联邦平均主要步骤为：

- 首先，中央服务器进行全局模型初始化并生成全局模型参数 w_0。
- 服务器在每轮通信中随机选择参与方的 N 位（K 为参与客户端总数，C 为选择比例，N=C*K），并将当前的全局模型参数 wt（t 为当初的训练轮次数目）发送给被选中的客户端。
- 每个选中客户端基于本地数据进行模型训练，并在每个 epoch 中使用 SGD 更新参数,如式（2.1），训练完成后将更新后的参数 w_k 上传至服务器。
- 服务器对接收到的参数基于加权平均进行聚合，如式（2.2）。

重复上述步骤直至模型收敛或达到预设通信轮数，具体流程如算法 2.1 所示。

$$w^{(k)} \leftarrow w^{(k)} - \eta \nabla F_k\left(w^{(k)}\right) \tag{2.1}$$

其中 $F_k(w)$是客户端 k 的本地损失函数, $w^{(k)}$ 表示客户端 k 的模型参数，η 为学习率。

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w^{(k)} \tag{2.2}$$

其中 n_k 是客户端 k 的数据量，n 为所有客户端的数据总量。

算法 2.1 联邦平均算法

```
服务器执行：   //K为客户端数量，k为客户端索引
初始化w0
for 每轮 t = 1,2,… do
    m ← max(C*K,1) //C为比例系数
    St ← (随机选择 m 客户端)
    for 每个客户端k ∈ St do
        w_{t+1}^{(k)} ← 客户端更新(k, w_t)
    w_{t+1} ← Σ_{k=1}^{K} (n_k/n) w^{(k)}
客户端更新(k, wt)：  //运行在客户端k
    β ← （将Pk分成大小为B的批次） //B是本地小批次大小，E是本地epoch的数量
    for 每个本地的epoch i from 1 to E  do
        for batch b∈β do
            w^{(k)} ← w^{(k)} − η∇F_k(w^{(k)})   //η是学习率
    返回 w^{(k)} 给服务器
```

2.3 模型/表示相似性度量

2.3.1 度量方法分类

模型/表示相似性度量方法的大多数都可以被归类为属于四个类别[35]:

- 基于表征相似性的测量（全局刺激关系结构比较）
- 基于相似性的测量（全局几何、神经元级对齐）
- 基于最近邻的测量（局部邻域相似性度量）
- 基于典型相关分析的测量（线性空间最大相关子空间对齐）

2.3.2 中心核对齐 (CKA)

（1） 核方法与希尔伯特空间

核方法是机器学习中一类基于正定核函数的算法，如式（2.3）中所示它通过核函数计算数据点间的相似度[36]。核函数将数据从原始空间隐式映射到高维特征空间，使得原来线性不可分的问题在高维空间中变得线性可分并且无需显式计算映射，避免了高维计算的复

杂性。，核方法根据核函数的不同类型适用于不同数据与任务，它在支持向量机、核主成分分析、核岭回归等任务中发挥着重要作用，能够有效的解决降维、分类、回归等问题。核函数的类别主要有线性核、多项式核、径向基核（RBF 核）等。

$$k\left(x,x^{'}\right)=\left\langle \Phi(x),\Phi\left(x^{'}\right)\right\rangle \tag{2.3}$$

其中 Φ 是特征映射，$\langle\cdot,\cdot\rangle$ 是高维空间内积。核矩阵（Gram Matrix）$K_{ij}=k\left(x_i,x_j\right)$ 需满足正定性。

希尔伯特空间(Hilbert Space) $\mathcal{H}$ 是核方法的数学基石，它是满足以下条件的向量空间：

- 内积定义：对任意 $f,g\in\mathcal{H}$，内积 $\langle f,g\rangle$ 满足对称性、线性和正定性。
- 范数完备性：所有柯西序列在 $\mathcal{H}$ 中收敛，即 $\mathcal{H}$ 是完备的。

希尔伯特空间中的内积度量数据点相似性，正交变换（如旋转）和各向同性缩放不改变内积结构，为核方法的不变性提供理论支撑。

（2） CKA 原理与计算

中心核对齐(Centered Kernel Alignment,CKA)[37]是一种衡量两个核矩阵之间相似性的方法，广泛应用于比较神经网络不同层的表示或者不同模型之间的相似性。CKA 的核心思想是计算两个表示（假设为 X 和 Y）对应的中心化核矩阵之间的归一化 Hilbert-Schmidt 独立性准则(Hilbert-Schmidt Independence Criterion, HSIC)[38]。

$X\in R^{nxp_1}$ 和 $Y\in R^{nxp_2}$为两组样本表示，通常为神经网络的激活值矩阵。其中，n 代表样本数量，p_1 和 p_2 分别是两个表示空间的维度。CKA 的具体计算步骤如下：

核矩阵构建

首先，利用核函数将样本映射到再生核希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS)，并计算对应的 Gram 矩阵，即核矩阵。通常选择线性核，其计算方式为：

$$K=XX^{\top}\in R^{n\times n} \tag{2.4}$$

$$L=YY^{\top}\in R^{n\times n} \tag{2.5}$$

K_{ij} 表示第 i 样本和第 j 样本在 X 表示空间中的内积（相似度），L_{ij} 同理。

中心化核矩阵

核矩阵中心化的操作通过减去行均值和列均值然后加上总均值来实现，能够去除核矩阵的均值影响，使得数据在特征空间中对齐原点,同时确保了原始表示加上任意常数向量后 CKA 的计算结果也不变，其计算方式如下:

$$K_c = HKH, \quad L_c = HLH \tag{2.6}$$

其中 $H = I - \frac{1}{n}11^\top$ ，I 为 n 阶单位矩阵，1 为全 1 列向量。

HSIC 计算

HSIC 是一个衡量两个随机变量在 RKHS 中统计独立性的指标[38]。对于中心化后的核矩阵 K_c 和 L_c，HSIC 定义为它们在弗罗贝尼乌斯范数(Frobenius norm)下的协方差的平方，其计算公式为:

$$\mathrm{HSIC}(K_c, L_c) = \frac{1}{(n-1)^2}\mathrm{tr}(K_c H L_c H) \tag{2.7}$$

其中 $\mathrm{tr}(\cdot)$ 表示矩阵的秩。HSIC 值的大小反映了 K_c 和 L_c（及其对应的原始表示 X 和 Y）之间的独立程度，若两者相互独立，HSIC 趋近于零。

归一化与 CKA 定义

HSIC 对各向同性缩放没有不变性，但是可以通过归一化使其不变。这种归一化指数被称为中心核对齐(CKA)[39] ，如式(2.8)所示:

$$\mathrm{CKA}(K_c, L_c) = \frac{\mathrm{HSIC}(K_c, L_c)}{\sqrt{\mathrm{HSIC}(K_c, K_c) \cdot \mathrm{HSIC}(L_c, L_c)}} \tag{2.8}$$

当 K_c, L_c 成比例（表示完全对齐）时，CKA = 1；当两者统计独立时，CKA $\approx$ 0。

（3） CKA 的性质

CKA 是一种衡量高维表示空间相似性的度量方法，它具有一系列重要的数学性质。这些性质使得其在分析和比较神经网络表示或者不同模型时具有强大的有效性和鲁棒性，具体性质如下[37, 39]：

- 正交不变性：CKA 对于输入表示空间的正交变换（如选择、反射）具有不变性。
- 各向同性缩放不变性：CKA 对于输入表示空间的均匀缩放具有不变性。
- 平移不变性：对核矩阵进行中心化处理（使用中心化矩阵 $H = I - \frac{1}{n}11^\top$ ）后，CKA 对原始表示数据的平移仍然具有不变性。

（4） 无偏 CKA

CKA[37]已然成为量化和比较不同系统（特别是在生物与人工神经网络中）内部表示相似性的重要工具，然而在特征维度远大于样本数目时等数据受限的情况下，标准 CKA 的估计值可能存在偏差。Alex Murphy 等人的研究[40]明确指出了这些偏差的来源，并提出了修正后的无偏 CKA 估计方法。

无偏 CKA 通过引入无偏 HSIC 估计器，修正了中心化步骤。对于矩阵 $A \in R^{n \times n}$，其

中 n 为样本数，其无偏中心化公式为：

$$\tilde{A}_{i,j} = \begin{cases} a_{i,j} - \frac{1}{n-2}\sum_{\ell=1}^{n} a_{i,\ell} - \frac{1}{n-2}\sum_{k=1}^{n} a_{k,j} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{k,l} & 若 i \neq j, \\ 0 & 若 i = j. \end{cases} \tag{2.9}$$

其中 $a_{i,j}$ 为核矩阵中第 i 行第 j 列的元素，表示样本 i 和样本 j 的核函数输出。$\tilde{A}_{i,j}$ 为无偏中心化后的核矩阵元素，满足以下性质：主对角线元素为 0（i = j 时），消除了原始核矩阵中与样本均值相关的偏差，使度量结果不再依赖于特征维度和样本数量的比例。

在修正后的 HSIC 基础上，无偏 CKA 的公式为：

$$\text{Debiased CKA}(K, L) = \frac{\text{HSIC}_{\text{unbiased}}(K, L)}{\sqrt{\text{HSIC}_{\text{unbiased}}(K, K) \cdot \text{HSIC}_{\text{unbiased}}(L, L)}} \tag{2.10}$$

相较于标准的 CKA 计算，无偏 CKA 具有更高的准确性、更强的鲁棒性。

2.3.3 其他相似性度量方法

（1） 余弦相似度(Cosine Similarity)

余弦相似度[41]通过计算两个非零向量夹角的余弦值来衡量它们在方向上的相似性并且忽略了其幅度的差异，结果在 [-1, 1]之间。对于局部模型的更新g_i和服务器模型的更新g_0，它们之间的余弦相似度c_i计算公式如式（2.11）所示：

$$c_i = \frac{\langle g_i, g_0 \rangle}{|g_i| \cdot |g_0|} \tag{2.11}$$

其中，$\langle \cdot, \cdot \rangle$表示两个向量的内积，$||\cdot||$表示向量的 l_2 范数。

余弦相似度经常被用于比较客户端模型更新之间的方向一致性，高余弦相似度表明客户端之间可能具有相似的数据分布或者学习任务，而低相似度或负值则可能意味着数据具有异质性或者更新具有潜在的对抗性。在联邦学习中的应用中，服务器使用余弦相似度来量化客户端本地模型更新与服务器模型更新之间的方向相似性[42]。

（2） KL 散度(Kullback-Leibler Divergence)

KL 散度，也称为相对熵(Relative Entropy)，它是由 Solomon Kullback 等人[43]于 1951 年提出的一种衡量两个概率分布之间差异的非对称性度量。对于两个离散概率分布 p 和 q，KL 散度的计算公式如下：

$$D_{KL}(p|q) = \sum_{x} p(x) \log \quad \frac{p(x)}{q(x)} \tag{2.12}$$

它主要用于衡量两个概率分布之间的差异，在联邦学习中可以比较不同模型对于同一组输入产生的预测概率分布的相似性[32]。

2.4 本章小结

本章系统地介绍了联邦学习的运行机制与关键技术，并深入分析了 CKA 的数学基础及其正交不变性、抗噪声等核心特性。在本章的最后阐述了余弦相似度、KL 散度等传统方法的特点和局限性。这些理论为设计基于 CKA 的贡献度评估机制提供了重要支撑，下一章将具体阐述该机制的设计框架与实现细节。

第三章 基于 CKA 的联邦学习贡献度评估机制设计

3.1 引言

本章针对数据异构场景下的贡献度评估难题提出了一种基于CKA的贡献度量化机制，该机制通过融合多层特征表示相似性、优化计算效率等方法在异构数据分布下实现了公平、高效且的贡献度评估。本章详细阐述机制的设计目标、联邦学习框架构建、CKA 计算模块实现以及数据异构模拟方法。

3.2 机制设计目标与原则

在联邦学习的实际应用中，参与方之间数据分布的异构性和贡献度的不平衡导致如何公平有效地评估各参与方对全局模型训练的贡献度成为了一个急需解决的问题。因此本项目设计了基于 CKA 的贡献度评估机制，项目机制设计的核心目标与原则为：

- 公平性：在各种不同的数据分布情况（包括 IID 数据分布、Non-IID 数据分布）下仍然能够公正地评估各客户端的实际贡献。
- 有效性：准确地衡量各参与方对全局模型性能提升的实质性贡献。
- 鲁棒性：在存在数据噪声、对抗样本或者不同的训练轮次下，保持相对稳定的贡献度评估效果。
- 计算效率：贡献度评估的过程不应该增加联邦学习的通信负担并且 CKA 计算花费的时间、资源消耗在可接受范围内。

3.3 总体框架设计

本文研究中提出的联邦学习本地模型贡献度评估框架采用模块化设计为四个核心组件：联邦学习框架中的服务器模块、客户端模块和贡献度评估框架中的特征提取模块、CKA 相似度计算模块，以及对项目参数进行设置的配置文件，如所图 3. 1 示，项目执行的核心流程如下：

（1） 加载配置

程序启动首先从配置文件文件中加载所有的配置参数，包括了联邦学习设置、模型类型、数据集的类型以及 CKA 贡献度评估等一系列相关参数。

（2） 初始化训练器

根据配置初始化服务器、所有的客户端、贡献度计算模块以及加载指定的模型和数据集，并根据配置（IID/Non-IID、数据量、噪声）为每个客户端分配相应的数据索引。

（3） 联邦学习迭代

启动联邦学习模型训练的主循环：

- 客户端选择: 服务器根据配置的参与率随机选择一部分客户端参与本轮训练。
- 模型分发: 服务器将当前的全局模型参数分发给被选中的客户端。
- 本地训练: 每个被选中的客户端在其本地数据集上使用接收到的全局模型执行训练操作，得到更新后的本地模型并对模型进行评估。
- 模型聚合: 参与训练的客户端将更新后的本地模型参数送回服务器。服务器使用联邦平均算法聚合这些本地模型参数然后生成新的全局模型并对模型进行评估。

（4） CKA 贡献度计算

- 在每轮联邦学习全局模型训练完成后，贡献度计算模块加载每个参与训练的客户端的本地模型和本轮聚合后生成的新全局模型。
- 特征提取模块使用一个共享的数据加载器提取模型在指定层的特征表示。
- CKA 相似度计算模块计算每个客户端本地模型和全局模型之间的 CKA 相似度。
- 根据 CKA 相似度评估各参与方的贡献度。

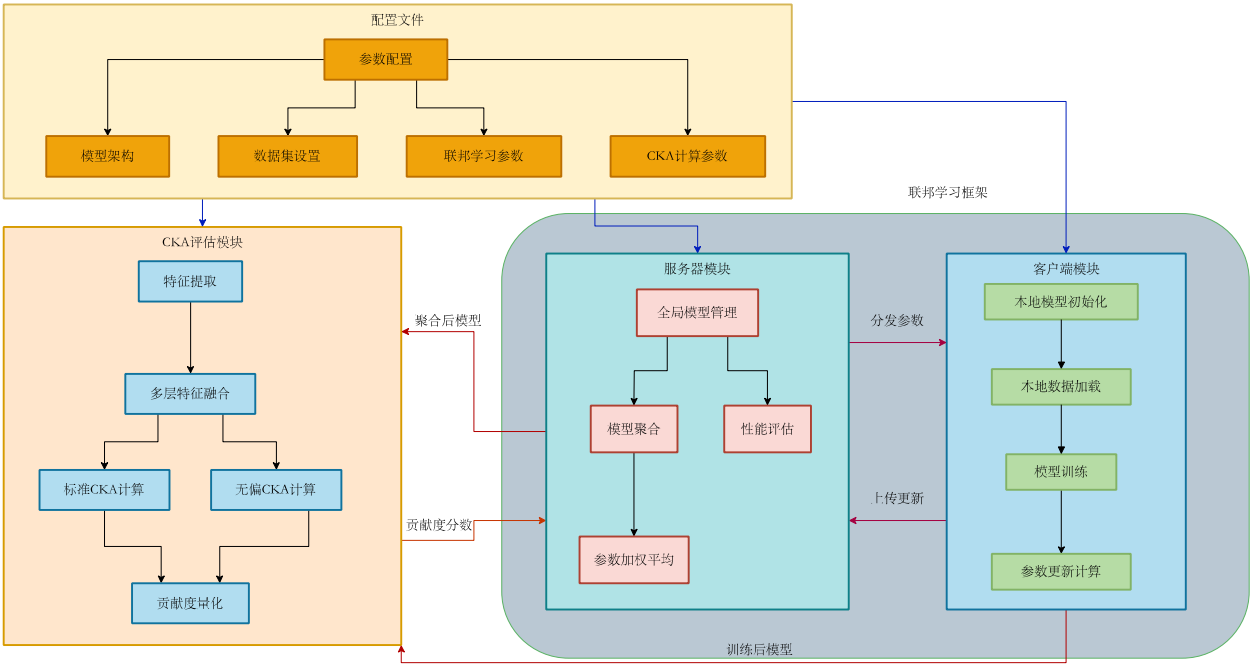

图 3.1 项目总体框架

3.4 联邦学习基础模块设计

3.4.1 服务器端设计

服务器端在联邦学习框架中负责维护全局模型状态、聚合来自客户端的更新以及评估

全局模型的性能。本小节将详细阐述服务器端的具体实现。

初始化与全局模型管理。在联邦学习流程运行的开始阶段，服务器端读取配置文件中的参数、加载配置文件中相应的模型和数据集并创建初始的全局模型。该初始化创建的全局模型是服务器管理的核心资产，代表着联邦学习训练过程中的共享知识。在最后服务器程序会基于提供的评估数据集和配置中批处理大小创建一个基于 PyTorch 的数据加载器，该加载器会用于模型聚合后的模型评估阶段。

模型聚合机制。服务器端的参数聚合是联邦学习流程中的核心环节，其目的是聚合来自所有参与方客户端的本地模型参数信息，然后根据聚合后的信息更新全局共享模型，它通过基于联邦平均算法[34]的聚合策略完成客户端参数集聚合，如算法 2.1 所示。在聚合过程中，服务器首先遍历自身维护的全局模型状态字典，其包含了全局模型的可学习参数（例如权重、偏置）、对应层名称以及参数张量。然后，对于全局模型中的每一个参数张量，聚合器根据参与方客户端上传的本地模型参数信息计算得到该参数在本轮的平均更新量。最后，服务器使用计算得到的平均值更新全局模型的对应参数：利用 PyTorch 中的 add 函数将平均更新量加到当前的全局模型参数上，实现高效的原地加法更新。

全局模型性能评估。调用 eval()函数将全局模型设置为评估模式，并使用 torch.no_grad()禁用梯度计算。使用初始化过程中创建的数据加载器加载评估数据集，迭代计算得到聚合后全局模型的准确率和平均损失。

3.4.2 客户端设计

客户端是联邦学习框架中的基本执行单元，负责在本地数据上执行模型训练任务并上传参数更新。本小节将详细阐述客户端的设计与实现，如图 3. 2 所示。

客户端初始化。客户端接受配置文件以及服务器分发的参数，加载与全局模型相同的架构将本地模型进行初始化。该步骤的关键在于，它使用服务器分发的全局模型状态字典初始化本地模型的参数，确保了本地训练从当前的全局状态开始训练操作。同时，准备本地的数据集和数据加载器。

本地模型训练。客户端的核心任务是在其本地数据上执行模型训练任务。首先，在本地模型完成初始化后，初始化一个 SGD 优化器，用于更新本地模型的参数。然后，将本地模型设置为训练模式，进行本地模型训练批次迭代，对于每个数据批次：

- 将数据和标签移动到模型所在的设备
- 清零优化器梯度
- 执行前向传播，得到模型输出

- 计算损失
- 执行反向传播，计算梯度
- 更新模型参数
- 记录并累加批次损失

返回模型参数更新。本地训练结束后，通过用本地训练结束后的模型状态减去传入的初始全局模型状态得到本地模型训练前后的参数差值，同训练结束时的完整本地模型状态字典一起返回服务器端。返回差值是 FedAvg 算法的标准做法，但后续计算 CKA 需要用到完整的最终状态。

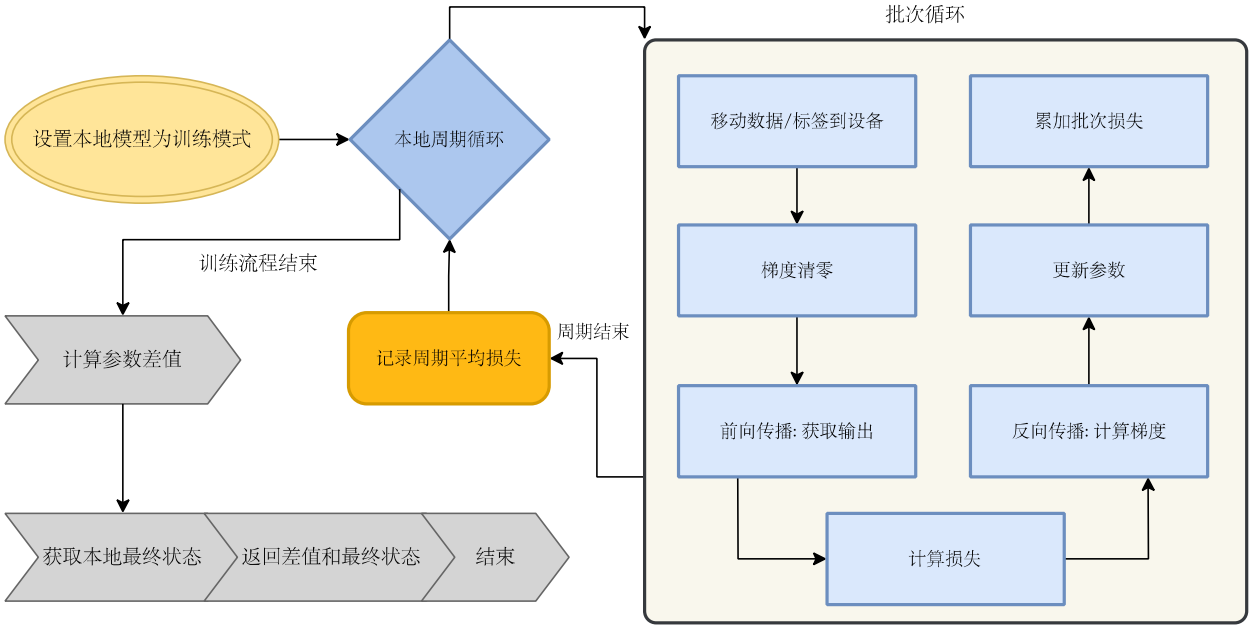

图 3.2 本地模型训练流程

客户端设计是构建联邦学习框架的基础，它封装了本地数据处理、模型训练和结果上报的核心逻辑，并能对本地训练后的模型进行性能评估。

3.4.3 数据异构性模拟

数据异构性[44]是联邦学习在实际应用中所面临的关键性挑战之一，在联邦学习中，它指的是不同客户端拥有的本地数据在数量、分布和质量上存在差异。

在毕业设计项目中，通过对客户端进行分类，进行不同的数据划分操作实现了数据异构性的模拟：

- IID + 数据量不均衡 (前 50%)：这些客户端的数据是独立同分布的，它们拥有所有类别的数据，但是它们拥有的数据量不同。这一类客户端模拟了参与方的设备

存储容量不同或者数据产生速率的差异。

- IID＋ 数据噪声 (20%)：这些客户端的数据同样是独立同分布的，数据量为平均水平，但它们的数据中会被添加不同程度的高斯噪声。这一类客户端模拟了不同客户端数据采集过程中可能引入的噪声或者数据质量的差异。
- Non-IID (后 30%)：这些客户端的数据是非独立同分布的，每个客户端只会被分配数据集中部分类别的数据。该类别模拟了联邦学习中不同客户端拥有的数据类别不同的情况。

在客户端初始化的过程中，跟据客户端的编号和配置获取该客户端的类别，创建该客户端对应的本地数据加载器。通过这样的数据划分操作，在本地训练迭代数据批次时，每个客户端实际上是在精心划分以及可能添加了噪声的数据子集上训练本地模型，因此能够有效地模拟联邦学习训练过程中存在的数据异构性。

3.5 CKA 贡献度计算模块设计

3.5.1 特征提取策略

特征提取策略用于从训练使用的神经网络模型中提取指定层的中间激活（特征表示），为后续的 CKA 贡献度计算提供输入。

该策略的核心机制是 PyTorch 中的前向钩子 (Forward Hooks) 机制。前向钩子允许在模型执行前向传播的过程中注册一个函数，该函数会在特定模块（层）完成它的前向传播操作之后、执行将其输出传递给下一层的操作之前自动调用。该策略会在每个指定的特征提取层上创建一个特定的钩子函数，遍历模型并在所有的目标层上注册对应的钩子后，执行一次前向传播。这次前向传播会触发所有注册好的钩子函数，获取对应层的特征张量，在对特征张量进行扁平化处理后存储在特征字典中，以便 CKA 计算模块调用。最后，在前向传播完成后，立即遍历所有注册的钩子句柄移除钩子函数，防止这些钩子在后续的模型使用中继续存在，导致性能下降和内存泄漏的危害。

该特征提取策略的核心是利用 PyTorch 的前向钩子动态地、非侵入式地捕获模型在单次前向传播过程中指定层的输出,如算法 3.1 所示：

算法 3.1 神经网络模型特征提取

```
输入：
- model：神经网络模型
- input_data：输入数据批次
```

算法 3.1 神经网络模型特征提取

```
- target_layers: 待提取的特征层名称集合

输出：
- features: 字典（键为层名称，值为特征向量）

流程：
1. 初始化空字典 features = {}
2. 定义钩子函数 hook_function(name):
   当目标层的前向传播完成时：
   a. 捕获该层输出 output
   b. 将 output 展平为一维向量
   c. 存入字典：features[name] = output
3. 遍历模型，为 target_layers 中的每一层注册钩子函数
4. 执行一次前向传播：model(input_data)
5. 移除所有已注册的钩子
6. 返回 features
```

3.5.2 CKA 计算实现

中心核对齐(CKA)是一种有效的神经网络表示相似性度量方法，在本节中，将详细阐述标准 CKA 和无偏 CKA 计算的实现。CKA 使用的核函数主要为 RBF 核和线性核，根据 Kornblith 等人的研究[37]，在绝大多数的情况中，RBF 核和线性核的计算结果是等价的，因此本文选择计算复杂度更低的线性核完成 CKA 计算的实现，并通过数学优化的方法大幅降低了 CKA 的计算复杂度。

（1） 标准 CKA 计算

标准的 CKA 计算过程中涉及到构建大型核矩阵，时间复杂度为 $O(n^2d)$，空间复杂度为 $O(n^2)$，此外，中心化操作的时间复杂度也为 $O(n^3)$，导致了其在处理大规模数据集时效率低下。本文实现了一种数学等价的优化版本[45]，通过直接中心化特征矩阵而非构建核矩阵，显著降低了计算和存储的负担，优化后的 CKA 计算流程如下：

特征矩阵中心化

在特征提取模块从模型的指定特征提取层中完成特征提取后将参数传递给 CKA 计算模块，CKA 计算模块通过直接中心化特征矩阵避免构建核矩阵。该步骤消除了特征矩阵的均值偏移，能够确保相似性度量对平移操作的鲁棒性。

对于任意的客户端本地模型的特征矩阵 $X \in R^{n\times d1}$ 和全局模型的特征矩阵 $Y \in R^{n\times d2}$，其中 n 为样本数，d1,d2 为特征维度，中心化方式如式（3.1）所示。

$$\mathrm{X_c} = \mathrm{X} - \frac{1}{\mathrm{n}} 1_\mathrm{n} 1_\mathrm{n}^\top \mathrm{X}, \quad \mathrm{Y_c} = \mathrm{Y} - \frac{1}{\mathrm{n}} 1_\mathrm{n} 1_\mathrm{n}^\top \mathrm{Y} \tag{3.1}$$

其中1_n为全 1 列向量，$1_\mathrm{n}1_\mathrm{n}^\top$为 n×n 的全 1 矩阵。

高效计算 HSIC 项

然后，直接使用中心化的特征矩阵和 Frobenius 范数计算 CKA 的分子和分母，HSIC 分子项如式（3.2）所示，HSIC 分母项如式（3.3）所示。

$$\mathrm{HSIC}_{xy} = |X_c^\top Y_c|_F^2 = \mathrm{tr}(X_c^\top Y_c Y_c^\top X_c) \tag{3.2}$$

$$\mathrm{HSIC}_{xx} = |X_c^\top X_c|_F^2, \quad \mathrm{HSIC}_{yy} = |Y_c^\top Y_c|_F^2 \tag{3.3}$$

tr(·) 表示矩阵的迹（对角线元素之和），通过迹的运算将高维矩阵运算转化为标量计算。

CKA 计算

根据 HSIC 项的结果计算最终 CKA 值，将 HSIC 项归一化至 [0,1] 区间，得到 CKA 值，如式（3.4）所示，其中 $\epsilon = 10^{-10}$。

$$\mathrm{CKA}(X, Y) = \frac{\mathrm{HSIC}_{xy}}{\sqrt{\mathrm{HSIC}_{xx} \cdot \mathrm{HSIC}_{yy}} + \epsilon} \tag{3.4}$$

这种优化将 CKA 计算中的时间复杂度从 $O(n^2d)$降低到 $O(nd^2)$（当 $n \gg d$ 时，提速显著），空间复杂度从 $O(n^2)$降低到 $O(d^2)$。该优化对于联邦学习环境中的模型贡献度评估至关重要，使其在实际应用中的大规模分布式系统中具有可行性。

（2） 无偏 CKA 计算

本文基于 Murphy 等人的研究[40]实现了使用无偏 HSIC 估计方法的 CKA 方法，通过修正 HSIC 估计中的系统性偏差提供更加可靠的相似度度量，有效解决了小样本情况下的偏差问题。与标准 CKA 计算相比，当样本数不大于 2 时无偏估计会因为分母为零的问题无法计算，此时需要回退到标准 CKA 的计算。下面为无偏 CKA 的计算方法：

计算线性核矩阵

和上述中的标准 CKA 优化版本不同，无偏 CKA 显式地构建核矩阵，因为它需要访问核矩阵的对角线元素。该步骤捕捉非线性关系为偏差修正提供基础。

使用线性核构建核矩阵，计算方式如式（3.5）所示。

$$K_X = X_c X_c^\top, \quad K_Y = Y_c Y_c^\top \tag{3.5}$$

计算去偏差的 HSIC

使用去偏差校正因子计算无偏 HSIC 的值，随着样本数 n 的增大，该因子会和标准 HSIC 的归一化因子一致。

同时，无偏 HSIC 需要在标准 HSIC 的值中减去核矩阵对角线元素的乘积。这是无偏估计的核心，能够消去自相关带来的偏差、提升小样本场景下的评估准确性。无偏 HSIC 的计算方式如式（3.6）和（3.7）所示。

$$\mathrm{HSIC}_{\mathrm{unbiased}}(K_X, K_Y) = \frac{1}{n(n-3)}\left[\mathrm{tr}(K_X K_Y) - \frac{\mathrm{tr}(K_X)\mathrm{tr}(K_Y)}{n-2}\right] \tag{3.6}$$

$$\mathrm{HSIC}_{\mathrm{unbiased}}(K_X, K_X) = \frac{1}{n(n-3)}\left[\mathrm{tr}(K_X^2) - \frac{\mathrm{tr}(K_X)^2}{n-2}\right] \tag{3.7}$$

同理计算 $\mathrm{HSIC}_{\mathrm{unbiased}}(K_Y, K_Y)$ 。

计算去偏差 CKA

使用式（3.4）归一化后得到无偏 CKA 值。

3.5.3 贡献度量化方法

本节阐述联邦学习本地模型贡献度评估机制中的量化方法：多层特征融合的贡献度评估框架。通过计算本地模型与全局模型在不同层次表示空间的相似性实现了对客户端贡献的精确量化。

不同层次捕获的特征表示在神经网络中具有不同的抽象级别，本文设计了多层特征融合的贡献度量化方法，这有利于全面、公平地量化参与客户端的贡献。该方法从模型的关键层提取特征表示并针对每一层计算客户端模型与全局模型特征表示之间的 CKA 相似度，然后基于不同层级的权重（层次越深，权重越高）将各层的 CKA 相似度进行加权平均，得到综合贡献度分数。

在本文的设计中，CKA 相似度得分被用来衡量客户端模型与全局模型在特征表示上的相似性，这种相似性被进一步解读为客户端对全局模型的贡献。 CKA 相似度是逐层计算客户端模型和全局模型在共享数据集上提取的特征之间的 CKA 相似度并聚合平均值得到的。CKA 得分越高（越接近 1），表示两个模型在对应层上学习到的特征表示越相似。在联邦学习的背景下，如果一个客户端模型经过本地训练后它的特征表示与全局模型的特征表示高度相似，那么就可以认为该客户端有效地学习了与全局目标一致的知识。所以较高的 CKA 相似度被视为该客户端对全局模型形成了更有价值或更一致的贡献，它的更新方向与全局模型的优化方向更为吻合，从而判断该客户端的贡献度更高。

3.5.4 CKA 计算在联邦学习流程中的嵌入点

联邦学习的运行流程按照"模型分发-本地训练-参数聚合"的形式迭代进行。在本文中，CKA 贡献度评估模块在联邦学习迭代过程中的嵌入点位于模型聚合后、全局评估前的阶

段，从而形成了完整的"模型分发-本地训练-参数聚合-贡献度评估"闭环，如图 3.3 所示。

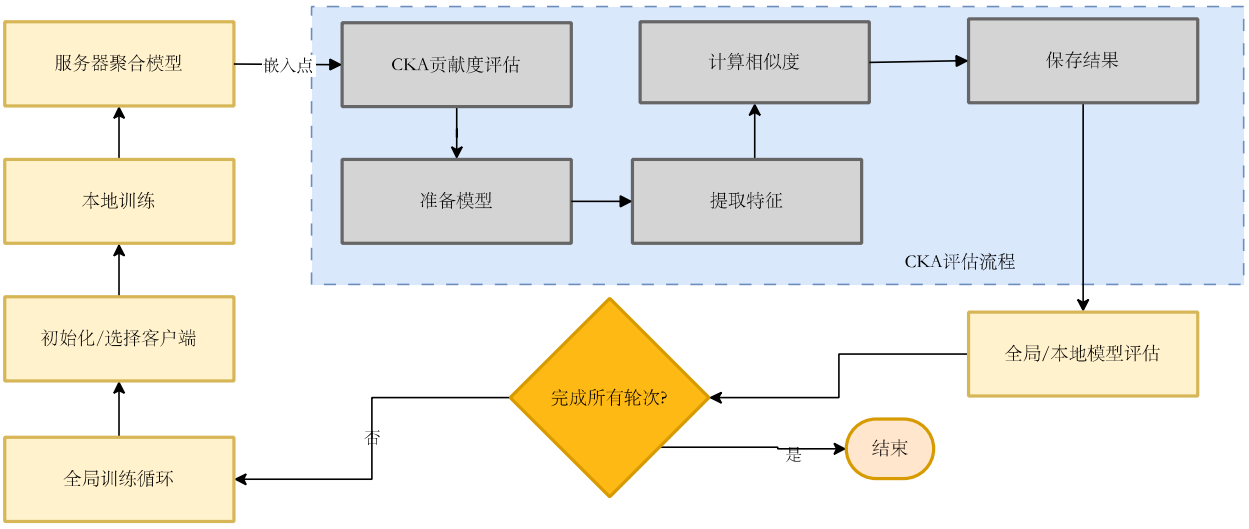

图 3.3 嵌入 CKA 贡献度评估的联邦学习流程

3.6 本章小结

本章说明了设计中构建的联邦学习框架、特征提取模块和 CKA 计算模块并提出了多层特征融合的策略以及 CKA 的数学优化方法，同时通过模拟 IID/Non-IID 混合数据、噪声干扰等场景为后续实验验证提供了多样化测试环境。下一章将基于此设计实验参数与评价指标。

第四章 实验设置与环境

4.1 引言

本章通过设计包含 MNIST、CIFAR 等数据集的实验方案并构建数据量不均衡、标签偏斜及噪声干扰的异构场景，验证基于 CKA 的贡献度评估机制的有效性。本章还配置了联邦学习与 CKA 计算模块的参数，为后续分析提供数据基础。

4.2 数据集与模型

4.2.1 数据集介绍

本文实现的贡献度评估机制支持多种标准图像分类数据集，包括 MNIST、CIFAR-10 和 CIFAR-100，并且通过模块化的设计实现了数据集的灵活配置和加载。同时，由于采用函数映射表的方式，数据集加载的方式具有可扩展性，后续能够集成新的数据集类型。

MNIST 作为手写数字分类领域的经典数据集，主要用于 0 到 9 这十个数字的识别任务。该数据集共计包含 70,000 张灰度图像，每张图像的尺寸为 28×28 像素，其中 60,000 张图像被划分为训练集，剩下的 10,000 张则作为测试集，在机器学习入门场景中应用十分广泛。

CIFAR - 10 是常用于自然图像分类的数据集，聚焦于 10 类通用物体的识别。它由 60,000 张 32×32 像素的彩色图像构成，整体划分为 50,000 张训练图像和 10,000 张测试图像。具体来说，其涵盖的类别包括飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车这十种常见物体。

CIFAR - 100 属于细粒度图像分类数据集，和 CIFAR - 10 有着相同的来源，但在类别设置上更为精细，包含 100 类物体。同样是 60,000 张 32×32 像素的彩色图像，不过其数据划分有所不同，每个类别仅有 500 张训练图像和 100 张测试图像，更适合对分类精度要求较高的场景。

4.2.2 数据预处理与划分

为了确保 CKA 贡献度评估的准确性，对于不同的数据集本文使用了特定的标准化参数，确保模型接收到的输入分布在统一的范围内：

```
# MNIST标准化参数
transforms.Normalize((0.1307,), (0.3081,))

# CIFAR-10标准化参数
cifar10_mean = (0.4914, 0.4822, 0.4465)
```

```
cifar10_std = (0.2023, 0.1994, 0.2010)

# CIFAR-100标准化参数
cifar100_mean = (0.5071, 0.4867, 0.4408)
cifar100_std = (0.2675, 0.2565, 0.2761)
```

在数据划分环节，为保障实验结果稳定性，系统运用多层次随机性控制机制。考虑到客户端数据分布的随机划分可能对模型训练产生潜在干扰，本方案通过以下方式实现随机过程的可控性：

随机种子绑定机制

针对非独立同分布客户端的类别分配建立客户端级种子绑定机制。为每个客户端分配唯一随机种子，该种子与客户端索引形成严格映射关系。通过固定种子值，在多次实验运行中可确保：

- 客户端类别洗牌序列完全一致
- 起始索引计算过程输出确定
- 类别子集选择逻辑具备可复现性

循环覆盖的动态分配算法

借助数学模运算构建确定性分配规则，如式（4.1）所示。

$$起始索引 = (k \cdot c)\ mod\ N \tag{4.1}$$

其中，k 表示客户端索引，c 为单客户端类别配额，N 是总类别数。该公式通过参数间的确定性运算关系，将随机洗牌后的类别序列转化为可预测的循环分配模式，在保证类别组合差异化的同时，消除随机索引跳跃带来的不可控波动。

噪声生成参数约束体系

对独立同分布客户端的噪声注入过程实施双重控制：基础噪声基线，动态波动阈值。前者设定标准差参考基准值作为噪声强度的标定锚，后者通过预设比例系数限定噪声参数的最大偏移范围。该机制在保留必要数据扰动的同时，通过参数边界约束有效抑制噪声随机性对数据分布的过度干扰。

此外，本文为不同数据集设计了特定的增强策略，包括：

- MNIST 增强策略：通过应用轻微的随机旋转(RandomRotation(10))完成增强，适合处理手写数字数据集的特点；
- CIFAR 增强策略：通过采用随机裁剪和水平翻转的组合完成，更适合自然图像数据集的特点；

- 条件性应用：通过特定的参数控制增强策略的使用，为训练和评估阶段提供不同配置。

4.2.3 数据异构场景

本节阐述中心化核对齐(CKA)贡献度评估机制中的联邦学习异构数据场景构建。本文通过系统化、多维度的方法模拟数据异构场景（其中包括数据量分布不均衡、高斯噪声干扰和类别分布偏斜等），为 CKA 贡献度评估机制提供了系统化的测试环境。

（1） 数据量不均衡场景

数据量的分布不均衡是联邦学习中普遍存在的异构形式，由现实中不同参与方的数据采集能力不同导致。在设计实现中，采用线性增长的缩放因子实现客户端的数据量在保持 IID 分布的同时具有梯度变化的特性。

模拟从资源受限的小型参与者到数据丰富的大型机构的场景，为观察数据量对于各参与方的贡献度的影响提供了理想的实验条件。

（2） 数据质量异构场景

数据质量的异构性模拟通过对数据添加高斯噪声的方式实现：通过在指定客户端类别的数据输入中添加不同程度的高斯噪声，能够为探索数据质量对模型训练和贡献度评估的影响提供途径。

（3） 标签分布偏斜场景

Non-IID 标签分布是联邦学习中最具挑战性的异构形式，本文通过为不同客户端分配数据集中特定的类别子集实现。此外本文采用了自适应的类别分配策略，程序会根据数据集的总类别数量动态调整预分配给客户端的类别数。具体实现如下：

类别数量动态分配

根据数据集特性动态设定每个客户端的类别配额，以平衡异构性与可学习性：对于 MNIST、CIFAR-10 等类别数较少（10 类）的数据集，每个 Non-IID 客户端分配 3 个类别；对于 CIFAR-100 等细粒度分类数据集（100 类），每个客户端分配 10 个类别。

确定性类别分配机制

使用全局类别随机化，基于客户端唯一标识符生成确定性随机种子对全局类别标签进行客户端特定的排列，并在随机化后的类别列表中实现循环索引覆盖。

类别分布特性分析

不同客户端因随机种子差异获得独立的类别排列顺序，但均从同一全局类别池中选取子集。例如，CIFAR-10 中客户端 A 可能分配类别{3,0,6}，客户端 B 分配{9,2,0}，二者在

类别 0 上重叠，但整体分布差异显著。此类部分重叠现象更贴近实际场景中用户数据的有限相关性，实现了部分重叠性。

每个 Non-IID 客户端仅包含配额内的类别样本，其余类别数据完全缺失。以 CIFAR-10 为例，客户端本地数据仅覆盖 30%的全局类别，形成严重的标签分布偏斜，实现了强制稀疏性。

本分配策略的设计参考了联邦学习研究中广泛采用的 Non-IID 模拟方法，其合理性体现在以下方面：通过限制客户端可见类别数量实现真实性适配，精准地刻画了现实场景中的标签分布偏斜问题，如边缘设备仅能采集局部环境数据；调整参数可灵活调节数据异构程度从而实现异构可控性，满足从温和偏斜（如 5 类别/CIFAR-10）到极端稀疏（如 1 类别/MNIST）的实验需求。

4.2.4 实验模型选择

本节介绍联邦学习本地模型贡献度评估机制中使用的两种核心神经网络架构：SimpleCNN 与 ResNet18。研究表明，这两种架构代表这不同架构复杂度和能力水平的模型选择，其对于准确测试贡献度评估机制的功能具有重要意义。

SimpleCNN 是本文设计中专为联邦学习场景设计的中等复杂度模型，具有良好的表征能力与计算效率平衡。它采用现代 CNN 设计理念，拥有明确的功能分离、适度的深度与宽度以及抗过拟合设计。

ResNet18 是现代深度学习模型架构中的代表模型架构，它引入了残差学习机制，具有高级语义表征能力。该模型架构具有残差连接机制、层次化残差块、模型训练知识基础以及全局的池化设计，使其适合用于联邦学习贡献度评估机制。

4.3 实验参数设置

本文设计通过在配置文件中设置参数调控联邦学习和 CKA 计算的配置。

4.3.1 联邦学习参数

联邦学习的基础参数会影响模型训练的过程和数据表征学习的质量，进而对 CKA 贡献度评估的结果产生显著影响。联邦学习参数配置包括：

- 客户端数量(no_models)：设置为 10，是参与联邦学习训练的总客户端数量。
- 每轮参与客户端数量(k)：同样配置为 10，它表示每个全局轮次中参与训练的客户端数量。
- 全局训练轮次(global_epochs)：设置为 20，设置联邦学习的总迭代次数。

- 本地训练轮次(local_epochs)：配置为 3，设置客户端在一次全局迭代中的本地模型训练次数。
- 学习率(lr)：设置为 0.005，保证了收敛速度与稳定性之间的平衡。
- 动量系数(momentum)：配置为 0.9，采用了联邦学习中 SGD 优化器的标准动量配置参数。
- 权重衰减(weight_decay)：设置为 1e-4，提供 L2 正则化以防止过拟合。
- 训练批次的大小(batch_size)：设置为 32，在内存占用与随机梯度下降有效性之间取得平衡。
- 工作线程数量(num_workers)：配置为 4，优化数据加载过程的并行性。

4.3.2 CKA 计算参数

本文为 CKA 贡献度评估机制设计了一套特有的参数配置，确保了表征相似性计算的准确性和计算效率：

- 特征提取层集合(cka_feature_layers)：根据使用的模型架构配置。对于 SimpleCNN 模型，配置为 ["features.3", "features.6", "classifier.3"]，对于 ResNet18，配置为 ["layer3", "layer4", "fc"]。
- CKA 批次大小(cka_batch_size)：设置为 64，用设置于 CKA 相似度计算中的批次大小。
- 去偏差 CKA(use_debiased_cka)：设置为 true 或 false，表明是否启用去偏差 CKA 计算。

4.4 评价指标

本文的研究通过构建一套完整的评价指标体系，全面的评估所提出的基于 CKA 的联邦学习本地模型贡献度评估机制的有效性、公平性和实用性。

4.4.1 联邦学习性能指标

联邦学习性能指标用于评估联邦学习框架的有效性和训练质量，主要包括全局模型和本地模型的准确率、损失。通过在测量聚合后的全局模型和训练完的本地模型在测试集上的分类准确率，得到模型的性能评估。其中，准确率通过计算累加预测正确的样本数和总样本数计算，如式（4.1）所示；损失使用交叉熵损失函数计算，累加批次损失并除以总样本的数量得到平均损失，如式（4.2）。

$$\text{准确率} = \frac{\text{正确预测样本数}}{\text{总样本数}} \times 100\% \tag{4.1}$$

$$\text{平均损失} = \frac{1}{N}\sum_{i=1}^{N} \text{交叉熵损失}\left(\textit{输出}_i, \textit{目标}_i\right) \tag{4.2}$$

4.4.2 贡献度评估指标

贡献度评估指标是本文研究的核心内容，用于量化和衡量参与客户端对全局模型的贡献。该指标是 CKA 相似度值，通过计算客户端模型与全局模型在指定特征提取层的表示空间相似度得到客户端的贡献度，如式（3.4）所示，它的取值范围为[0,1]，值越高表示对应客户端的贡献越大。

4.4.3 对比实验指标

对比实验指标用于验证 CKA 方法的优越性和可靠性，通过与其他相似度度量方法进行对比得到。

不变性测试得分：该对比实验通过对特征表示进行正交变换的操作，测量变换前后相似度度量值的变化程度，从而验证 CKA 的旋转不变性特性。

噪声敏感性曲线：噪声敏感性实验通过向模型表征添加不同强度的高斯噪声，从而测量不同的相似度度量方法对噪声的响应曲线，评估 CKA 对噪声的稳健性。

4.5 本章小结

本章详细说明了数据集选择、预处理策略、异构场景构建及模型配置并明确了联邦学习参数（如本地训练轮次、学习率）与 CKA 计算参数（如特征提取层、批次大小）。实验设计覆盖性能验证、贡献度区分度测试及方法特性分析，为第五章的结果解读提供依据。

第五章 实验结果与分析

5.1 引言

本章通过全局模型收敛性验证、贡献度排名稳定性分析及 CKA 特性实验系统地评估所提机制的有效性。实验涵盖不同数据分布（IID/Non-IID/噪声）下的贡献度量化结果，并通过对比 KL 散度、余弦相似度验证 CKA 的优越性。

5.2 联邦学习有效性验证

本节旨在评估所实现的联邦学习框架的有效性，在设定的数据集（MNIST、CIFAR-10 和 CIFAR-100）上运行联邦学习后，得到全局模型的准确率和损失曲线。下面以 CIFAR-10 数据集的结果为代表进行分析。

如图 5.1 所示，随着联邦学习训练轮次的不断迭代，全局模型的准确率呈现明显的上升趋势。全局模型在初始阶段（前 5 轮）准确率快速提升（从 42.76 %升至 78.34 %），表明模型迅速学习到数据中的基础模式；第 5-15 轮进入平稳上升期（78.4 %→83.61 %），学习速率放缓但仍保持正向增益；第 15 轮后准确率呈现小幅波动（82.9 %→83.91 %），可能由于数据异构性导致局部最优震荡。最终准确率稳定在 84%左右，验证了联邦学习框架的有效性。

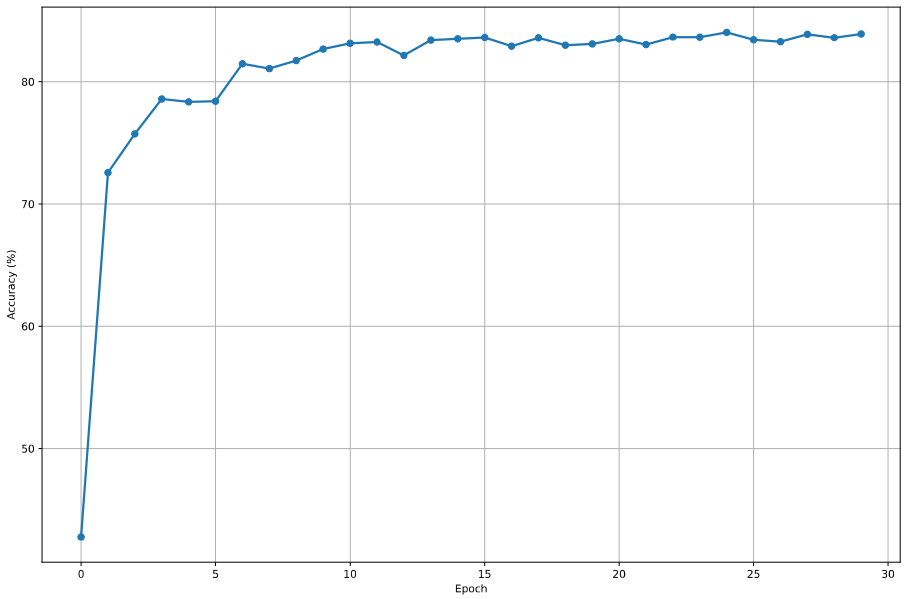

图 5.1 CIFAR-10 下全局模型准确率曲线

该实验结果表明本文所实现的联邦学习框架能够有效地聚合客户端的模型更新，从而

使得全局模型逐渐学习到数据中的有效模式并提升模型的性能。此外结果显示，在训练初期时模型准确率较低但准确率随着训练的进行快速上升，但在训练的中后期时准确率呈现波动上升的趋势。

同时全局模型的损失值随着训练轮次的增加而持续下降，如图 5.2 的折线所示。初始阶段损失值急剧下降（交叉熵损失从 1.47 降至 0.62），对应模型快速拟合数据分布；中后期损失下降趋缓（0.62→0.49），符合深度学习中损失函数收敛规律；第 18 轮后损失值波动范围（±0.03）显著小于初期，表明全局模型趋于稳定。该结果表明了联邦学习维护的全局模型在不断学习并减少预测误差，损失曲线的下降趋势也反映了全局模型对训练数据的拟合程度在逐步的提高。这些都反映了本文设计的联邦学习框架所使用的模型训练过程是有效的。

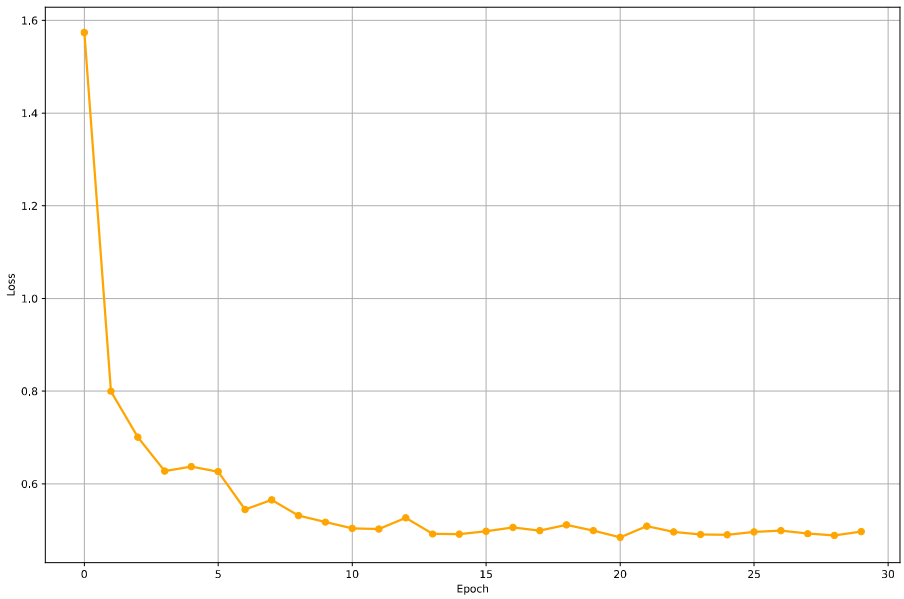

图 5.2 CIFAR-10 下全局模型损失曲线

此外，本文设计并模拟了实际应用中的数据异构性场景，导致各客户端的模型性能可能不同，同样以 CIFAR-10 数据集上运行的结果为例，如图 5.3 和图 5.4 所示。

在图 5.3 中展示了全局模型与本地模型在准确率上的对比：IID 客户端本地模型准确率与全局模型差距较小，表明其数据分布与全局模型优化方向一致，贡献度较高；Non-IID 客户端准确率偏低，反映其局部数据偏斜导致模型泛化能力下降；噪声客户端的表现最差，且随着训练轮次波动剧烈，验证了数据质量对模型性能的负面影响。

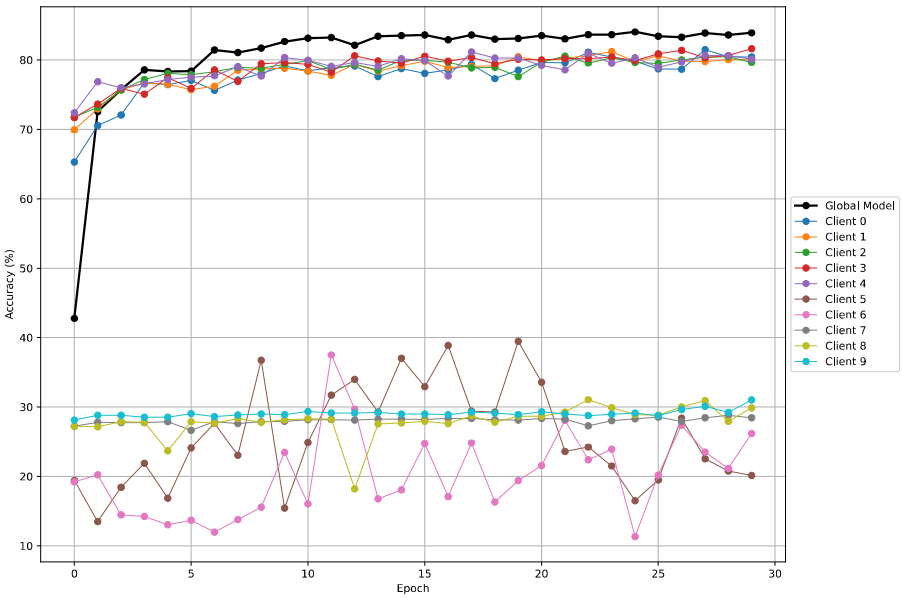

图 5. 3 模型准确率：全局 vs 本地

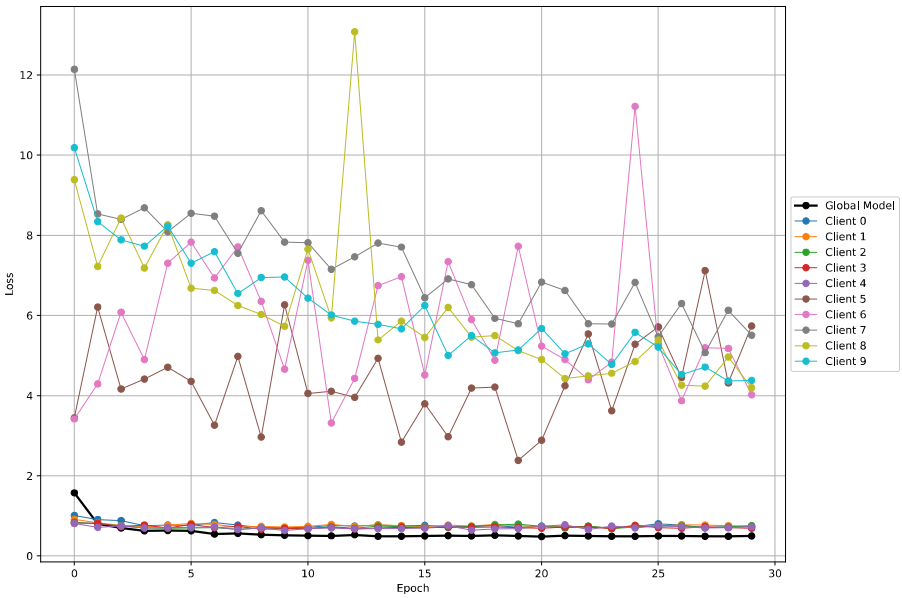

图 5. 4 模型损失：全局 vs 本地

在图 5. 4 中全局模型和 IID 客户端模型的损失值远低于非 IID 和噪声客户端的本地模型损失值，表明了全局模型在整合各客户端信息后具有更好的泛化能力和稳定性。非 IID 和噪声客户端的模型由于数据分布的偏差或质量的降低，其损失值相对较高，对全局模型的贡献较小。

在联邦学习的模型训练中，拥有高质量数据分布的 IID 客户端模型性能随着模型训练轮次的迭代稳步上升并接近全局模型的性能，表明它对全局模型做出了较大的贡献。而噪声客户端和 Non-IID 客户端的模型性能表现低迷，符合实现的客户端数据划分情况，说明

了联邦学习框架实现的有效性。

实验的结果表明本文实现的联邦学习框架能够有效地训练全局模型并达到一定程度的收敛。此外，该框架能够有效地聚合来自具有不同数据特征（IID, Non-IID, 噪声）的客户端数据特征知识，从而生成一个性能稳健、具有良好泛化能力的全局模型，验证了本文的联邦学习框架在分布式环境下协调客户端进行模型训练的可行性与有效性。

5.3 CKA 贡献度评估实验

5.3.1 不同异构程度下的贡献度分析

在联邦学习的实际应用中，参与方客户端的数据往往不是呈现独立同分布（Non-IID），因此某些参与客户端的模型更新可能会对全局模型的性能产生负面的影响，或者导致评估各参与方客户端的贡献变得极为困难。所有，在数据分布不均匀的场景下，研究如何公平、有效的评估参与方客户端的更新是一个富有意义的举动，这对于理解模型训练动态、设计激励机制和提升联邦学习系统的鲁棒性非常重要。本文的设计采用 CKA 作为贡献度评估的核心指标，探索了该指标在数据异构的场景下，它能否分辨具有不同数据质量、分布的客户端，并量化各客户端的贡献度。

在第四章实验环境设置中提到，本文设计模拟了实际应用中的数据异构性场景，将客户端分为拥有不同数据质量的分类，并为其分配对应的数据。在实验设置的数据集（MNIST、CIFAR-10 和 CIFAR-100）上运行联邦学习贡献度评估机制，得到三种不同分类的客户端的贡献度分布情况。

以 CIFAR-10 的结果为例，图 5. 5 反映了不同数据质量分类的客户端在联邦学习训练过程中的平均准确率变化。在图中 IID 客户端的平均准确率（79.7%）显著高于 Non-IID（28.5%）和噪声客户端（26.3%），三类客户端间的差异具有统计显著性。结果表明在联邦学习的模型训练中，拥有高质量数据分布的 IID 客户端训练完成的模型性能在模型训练轮次迭代中普遍高于 Non-IID 客户端和噪声客户端的本地模型性能。基于 CKA 的贡献度评估机制能够识别出这种差异，为公平地评估各客户端的贡献提供了依据。

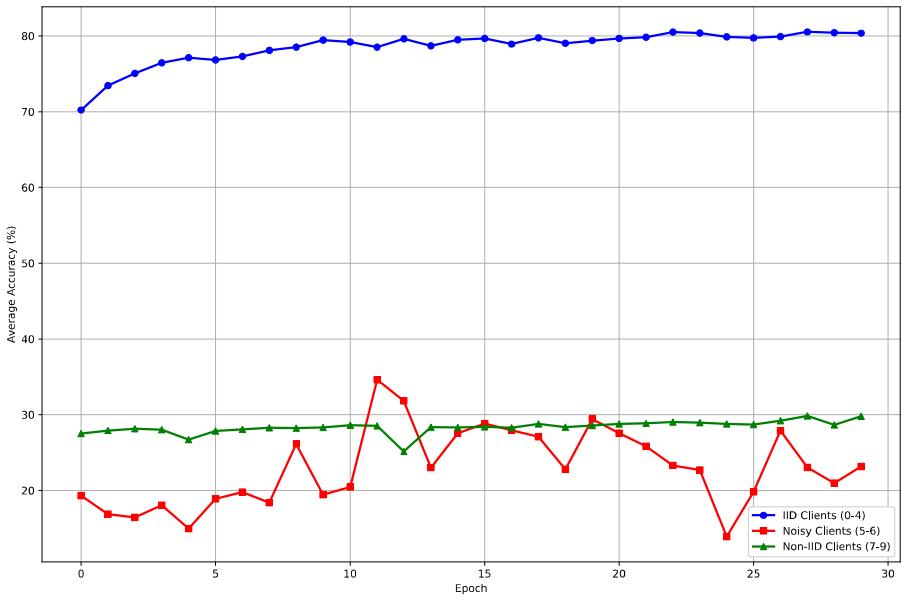

图 5.5 不同分类客户端的平均准确率

因此，拥有不同数据质量的客户端对全局模型训练的贡献度也同样应该具有显著的区别，高质量的客户端模型更新会对全局模型产生更积极的作用。实验结果显示，如图 5.6 和图 5.7 所示，本文实现的基于 CKA 的联邦学习本地模型贡献度评估机制能够有效、公平地识别在数据异构场景下各参与方客户端的贡献度，分辨拥有不同数据质量的客户端种类。

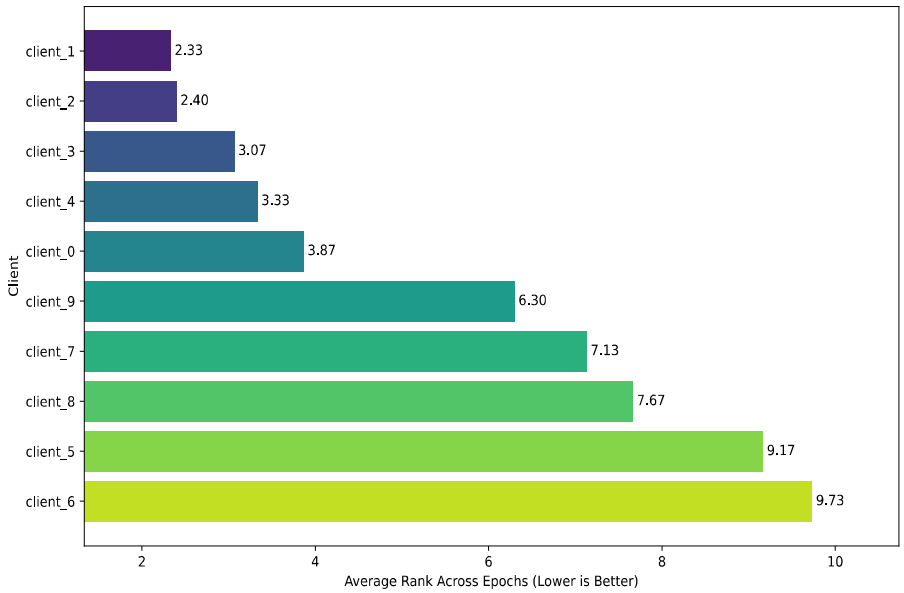

图 5.6 客户端的平均贡献度排名

图 5.6 中的排名结果呈现了清晰的层次结构：前 5 位均为 IID 客户端，中间 3 位为

Non-IID，末 2 位为噪声客户端。拥有高质量数据的 IID 客户端具有较高的平均贡献度排名，而非 IID 和噪声客户端的排名相对于前者则较低。这种稳定的排名差异表明本文所提出的贡献度评估机制能够有效地量化各客户端对全局模型的贡献，为激励机制的设计提供可靠依据。

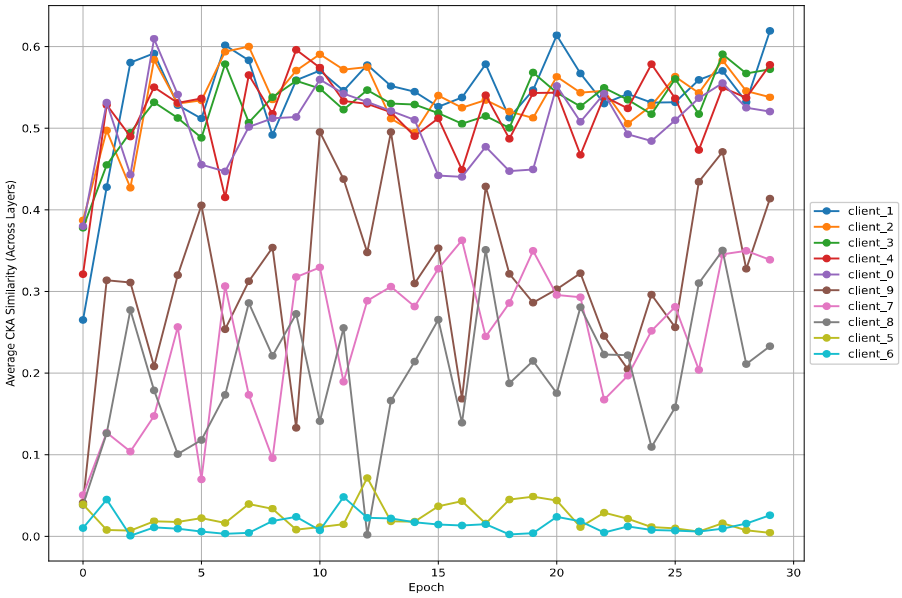

图 5.7 客户端的 CKA 贡献度

图 5.7 展示了各客户端的 CKA 贡献度值，直观地反映了不同客户端对全局模型的贡献大小。IID 客户端 CKA 值分布主要集中在[0.45,0.6]区间，Non-IID 组则集中在[0.15,0.35]区间且分散性较高，噪声组最低且波动大。贡献度值较高的客户端通常对全局模型的性能提升有较大贡献，相反的是贡献度值较低的客户端可能由于数据质量问题或数据分布的偏差对全局模型的贡献较小，基于 CKA 的贡献度评估机制能够准确地捕捉到这种差异。

实验结果表明，本文的贡献度评估机制具有良好的有效性和强大的鲁棒性。它能在数据分布不均匀的场景下准确识别各参与方客户端模型更新的质量，从而能够公平、有效地评估各参与方客户端的贡献度。

5.3.2 贡献度评估稳定性分析

在联邦学习贡献度评估机制中，贡献度评估的稳定性同样重要，其是确保系统公平性、可持续性和效率的核心要素。良好的贡献度评估稳定性能保障真正提供高价值数据的拥有者（如数据多样性高、质量优的参与客户端）的正确识别，避免因偶然波动导致误判参与方的贡献。此外，良好的稳定性也能抵御恶意攻击和噪声干扰。

在本文的设计中，通过统计所有全局轮次中各参与方客户端的评估下排名，分析其贡

献度排名是否稳定，如图 5.8 所示。热力图直观地展示了各客户端在多个全局轮次中的贡献度排名位次及其变化情况。拥有高数据质量的 IID 客户端因为其数据对于模型训练的积极作用，所以均处于贡献度排名的前列。与之相反的是，另外两类低质量数据的客户端因为其数据对模型训练的较小作用或消极作用，所以处于贡献度排名的低位。

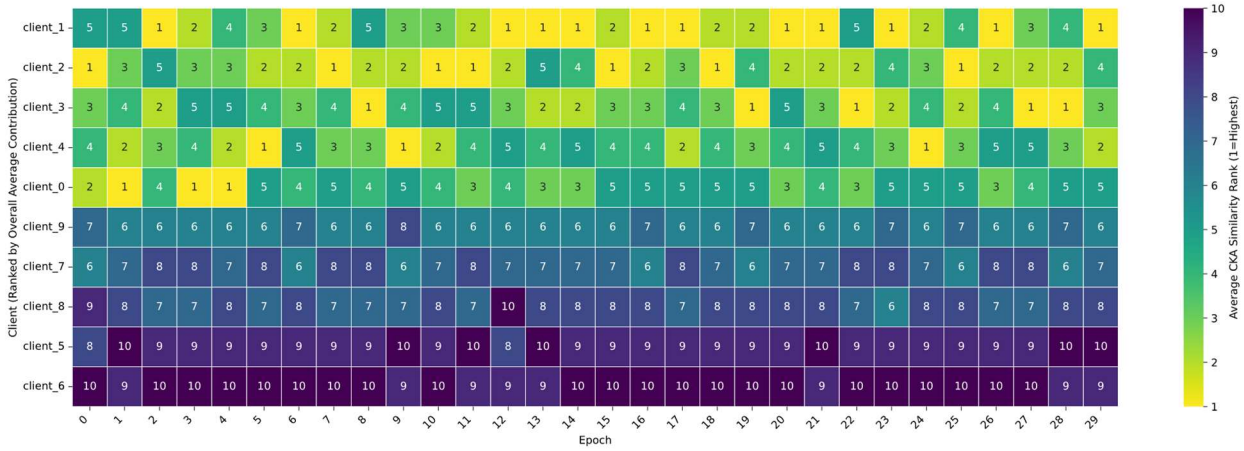

图 5.8 CKA 贡献度排名热力图

实验结果表明，拥有高质量数据的客户端种类始终保持着较高的贡献度排名，且位次波动范围较小。而 Non-IID 种类的客户端和噪声客户端的贡献度排名始终在较低的位次波动，其中数据噪声比例最高的客户端贡献一直位于最低位。这表明本文的贡献度评估机制具有良好的稳定性，能够保持对各类数据质量不同客户端进行稳定的评估。

5.4 计算效率比较

本文实现了一种数学等价的 CKA 优化版本，通过直接中心化特征矩阵而非构建核矩阵，显著降低了计算和存储的负担，图 5.9 展示了优化前后 CKA 花费的计算效率对比。

实验结果显示在一次完整的联邦学习的流程中优化后 CKA（灰色菱形虚线）的计算时间在全部训练周期（Epoch 0-30）内始终低于优化前 CKA（黑色方块虚线）的计算时间，平均改进达 49.30%。这表明了实验优化的有效性，这种优化将 CKA 计算中的时间复杂度从 $O(n^2d)$降低到 $O(nd^2)$（当 $n \gg d$ 时，提速显著），空间复杂度从 $O(n^2)$降低到 $O(d^2)$，从而使得 CKA 计算花费的时间大幅度减少，增加了计算效率。并且优化后的时间曲线波动较小，表明方法在不同训练阶段均保持高效稳定，未引入额外计算瓶颈。

该方法通过数学等价变换，将 CKA 的计算复杂度从平方级降至线性级，尤其适用于大规模数据集或深层模型，为贡献度分析提供了高效的实现方案。

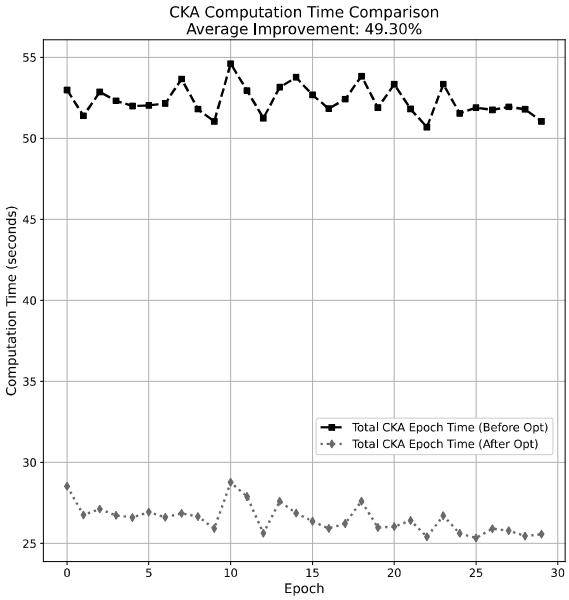

图 5.9 计算效率对比

5.5 CKA 方法特性分析

5.5.1 不变性检验结果与分析

本节验证 CKA 的正交不变性，具体的做法如下。

首先，随机生成一个模型特征，并计算它与自身的 CKA 相似度、KL 散度和余弦相似度作为基线对照。

然后，对模型特征施加一个正交变换（如乘以正交矩阵），在和原来的模型特征计算上面三种相似度。

收集实验运行的结果，比较施加正交变换前后三者的差距，如图 5.10 所示。图中的实验结果验证了 CKA 的正交不变性。对模型特征施加正交变换后 CKA 的相似度变化幅度很小，仍然及其接近完全相同（从 1→0.9917），KL 散度因对正交变换敏感而变化较大（从 0→0.9815）。这表明 CKA 在衡量模型表示相似性时具有良好的正交不变性，能够更好地捕捉模型表示的本质特征，而不受特征空间旋转等变换的影响。

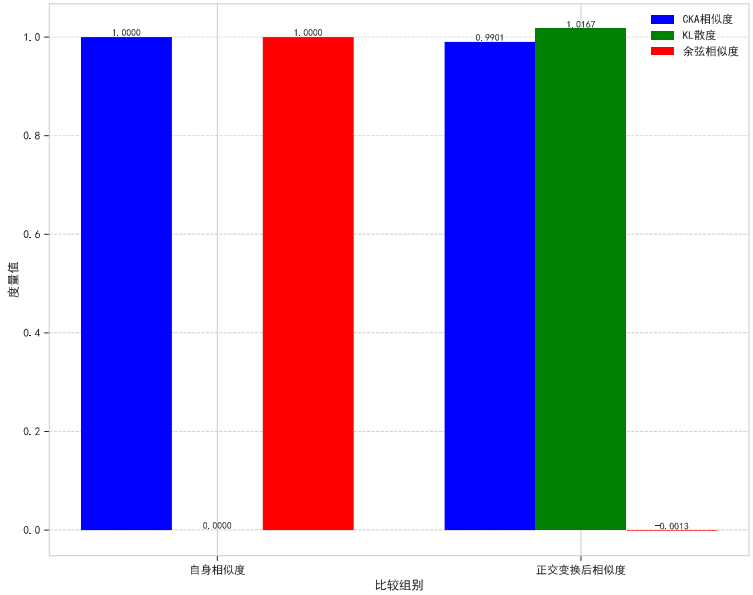

图 5.10 正交不变性检验

5.5.2 噪声敏感性测试结果与分析

在该实验中，通过对特征添加不同强度的高斯噪声，并逐步递增迭代，然后分别计算CKA、KL 散度和余弦相似度三种度量的均值并进行归一化，如图 5.11 所示。

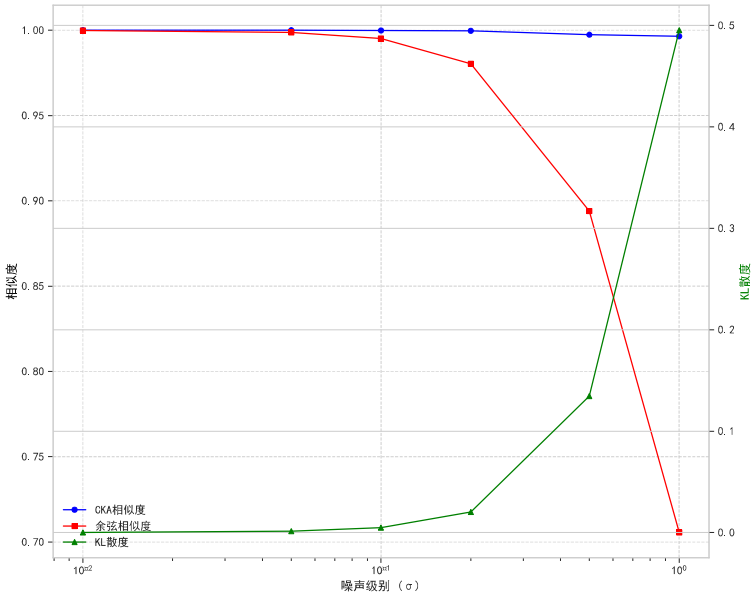

图 5.11 噪声敏感性实验

图中的实验结果表明，CKA 在不同程度的高斯噪声扰动下均表现出较高的鲁棒性，随着噪声比例增加相似度下降较慢，标准差和变化率都比较小。相比之下，KL 散度和余弦相似度则对噪声的干扰更敏感，相似度变化更快、变化幅度更大。这个实验结果表明了

CKA 更适合衡量特征空间的结构相似性，尤其在存在噪声的情况下能够更稳定地评估模型表示的相似性。

5.6 本章小结

实验结果表明，基于 CKA 的机制能稳定区分各种贡献度不同的客户端，且对正交变换与噪声干扰具有强鲁棒性。CKA 在贡献度排名一致性、抗噪声等方面显著优于传统方法，验证了其在联邦学习公平性评估中的实用价值。

第六章 总结与展望

6.1 引言

本文针对联邦学习中的数据异构性问题，提出了一种基于 CKA 的贡献度评估机制。本章总结全文工作，提炼创新点与不足，并探讨未来研究方向，为联邦学习的公平性优化与扩展应用提供参考。

6.2 全文工作总结

联邦学习作为一种新兴的分布式隐私保护机器学习范式，在解决数据孤岛和隐私安全问题方面展现出了巨大的潜力。然而，在实际应用中，普遍存在的数据异构性给参与方的贡献度评估带来了严峻挑战，影响了联邦学习系统的公平性、效率和鲁棒性。针对这一核心问题，本文深入研究并提出了一种基于中心核对齐的联邦学习本地模型贡献度评估机制，并通过系统性的研究和实验验证，主要的工作如下：

联邦学习框架搭建：设计并实现了基于 PyTorch 的联邦学习框架，支持数据异构场景，其中采用了联邦平均(FedAvg)算法作为客户端参数更新的聚合机制。

CKA 贡献度评估机制设计：引入 CKA 相似度作为联邦学习贡献度的评估指标，并提出了基于多层特征融合的贡献度量化方法。同时对 CKA 的计算方式进行了等效的数学优化，显著降低了计算复杂度。

数据异构性模拟：通过对客户端的数据进行 IID/Non-IID 混合划分、标签偏斜以及高斯噪声注入，实现了多维度的异构数据场景模拟，其覆盖数据量、质量与分布差异的典型数据异构性挑战。

实验验证与分析：在 MNIST、CIFAR-10/100 数据集上运行，验证机制的有效性：

- 有效性：基于 CKA 的贡献度评估机制能够准确地区分 IID、Non-IID 与噪声客户端，高质量客户端贡献度比噪声客户端和 Non-IID 客户端高。
- 稳定性：贡献度的排名波动范围处于可接受的范围，并且噪声客户端始终稳定处于末位。
- 鲁棒性：CKA 对正交变换保持严格不变性，且对噪声的干扰保持了强大的鲁棒性。

6.3 主要创新点

本文提出了基于 CKA 的联邦学习本地贡献度评估机制，其在联邦学习面对的数据异构性挑战中，展现了优秀的有效性和鲁棒性。相较于其他的依赖数据估值、博弈论（如

Shapley 值）或者仅仅关注模型更新数值的贡献度评估方法，CKA 从模型内部表征相似性的角度提供了一种新颖、更深层次的贡献度量化方法，特别是在应对数据异构性问题展现出了独特的潜力。此外，CKA 相似度的抗噪特性、正交不变性以及模型深层相似性提取能力，使得其是一种适合联邦学习场景的相似度度量方法，可以用于评估联邦学习中各参与方的贡献度。

此外，本文对 CKA 进行了两个创新性的应用。首先，对 CKA 计算方式进行优化，通过避免直接构建显式的特征矩阵，减少了 CKA 的计算复杂性。其次，在面对高维度小数据样本时，标准 CKA 会出现评估失真的问题，采用了无偏 CKA 计算相似度。这两个创新使得本文的联邦学习本地模型贡献度评估机制能够根据面对的计算需求，灵活采用 CKA 的计算方法，并保持对评估效果的有效性，同时减少了 CKA 需要的计算资源。

6.4 存在的不足

当前实验的数据应用场景主要聚焦于图像分类任务，尚未涵盖文本、时序数据等多模态数据场景，因此难以验证该机制在跨模态数据场景中的普适性。此外，尽管通过优化使 CKA（Centered Kernel Alignment）计算方法的效率获得显著提升，但在面临超大规模客户端（如客户端数量超过 1000 个）或深层神经网络模型（例如 ResNet-152）时，特征提取与相似度计算过程仍对计算资源和内存空间构成较大压力。

本次实验采用 10 个客户端参与联邦学习训练的规模，虽能有效验证所提机制的基础功能与核心特性，但在模拟真实大规模联邦学习场景时存在以下局限性：

首先，模型泛化能力验证的充分性不足。小规模客户端构成的异构场景复杂度有限，难以全面模拟真实环境中非独立同分布数据的高度碎片化特征（如超大规模 Non-IID 分布或动态变化的客户端参与模式）。例如，当客户端数量激增时，全局模型可能面临更显著的收敛震荡风险或陷入局部最优解，而本文实验尚未验证 CKA 贡献度评估机制在这类极端场景下的适应性。

其次，贡献度评估的稳定性面临挑战。在客户端数量较少的情况下，单次实验中贡献度排名的波动可能被局部数据特性放大，而大规模场景下客户端间的贡献差异可能呈现更复杂的分布模式。例如，当存在大量低质量客户端时，CKA 算法对有效客户端的区分能力是否仍能保持高鲁棒性，需要进一步通过大规模实验验证。

此外，计算效率的验证存在场景局限。尽管本文通过数学优化降低了 CKA 的计算复杂度，但相关实验仅在小规模客户端环境下验证了效率提升效果。当客户端数量扩展至千

级规模或模型深度显著增加时（如采用 ResNet-152 等深层架构），特征矩阵的维度膨胀可能导致计算量与内存占用呈非线性增长，需结合分布式计算框架或分层聚合策略进行进一步优化。

未来研究需将实验场景扩展至更大规模的联邦学习环境，系统探索该机制在超多客户端、动态参与模式和复杂异构数据分布下的性能边界。同时，需设计轻量化计算策略（如基于子空间投影的近似计算方法或自适应采样技术），以提升机制在资源受限场景中的实际应用价值。

6.5 未来工作展望

本文希望在未来将基于 CKA 的贡献度评估机制用于更多的领域，而不是仅限于图像分类领域，并探索更多的 CKA 优化方法，缓解 CKA 面临的计算资源消耗大的问题。

此外，本文的实验环境为模拟的联邦学习场景，因此不存在通信资源、数据保密等需要解决的问题，未来需要在真实联邦学习系统中验证机制的可行性和性能。

6.6 本章小结

本文通过理论分析、机制设计与实验验证，证明了 CKA 在贡献度评估中的有效性，其优化计算与多层特征融合策略提升了实用性。未来工作需扩展至多模态场景、优化计算效率，并探索与激励机制的结合，以推动联邦学习在更复杂环境中的落地。

参考文献

[1] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and Open Problems in Federated Learning[J]. Found. Trends Mach. Learn., 2019,14: 1-210.

[2] COLLINS E, WANG M. Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence[J]. arXiv, 2025.

[3] BASHIR A K, VICTOR N, BHATTACHARYA S, et al. A Survey on Federated Learning for the Healthcare Metaverse: Concepts, Applications, Challenges, and Future Directions[J]. ArXiv, 2023,abs/2304.00524.

[4] 赵英, 王丽宝, 陈骏君, 等. 基于联邦学习的网络异常检测[J]. 北京化工大学学报（自然科学版）, 2021,48(2): 92-99.

[5] SHAHEEN M, FAROOQ M S, UMER T, et al. Applications of Federated Learning; Taxonomy, Challenges, and Research Trends[J]. Electronics, 2022,11(4): 670.

[6] YE M, FANG X, DU B, et al. Heterogeneous Federated Learning: State-of-the-art and Research Challenges[J]. ACM Computing Surveys, 2023,56(3): 1-44.

[7] DILLEY O, PARRA-ULLAURI J M, HUSSAIN R, et al. Federated Fairness Analytics: Quantifying Fairness in Federated Learning[J]. ArXiv, 2024,abs/2408.08214.

[8] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and Open Problems in Federated Learning[J]. 2019.

[9] SIOMOS V, PASSERAT-PALMBACH J. Contribution Evaluation in Federated Learning: Examining Current Approaches[J]. ArXiv, 2023,abs/2311.09856.

[10] S. R H, I. A, M. B, et al. Decision Trees in Federated Learning: Current State and Future Opportunities[J]. IEEE Access, 2024,12: 127943-127965.

[11] KANG R, LI Q, LU H. Federated machine learning in finance: A systematic review on technical architecture and financial applications[J]. Applied and Computational Engineering, 2024,102(1): 61-72.

[12] LIU B, LV N, GUO Y, et al. Recent Advances on Federated Learning: A Systematic Survey[J]. Neurocomputing, 2023,597: 128019.

[13] ZHANG C, XIE Y, BAI H, et al. A survey on federated learning[J]. Knowledge-Based Systems, 2021,216: 106775.

[14] T DINH C, TRAN N, NGUYEN J. Personalized federated learning with moreau envelopes[J]. Advances in neural information processing systems, 2020,33: 21394-21405.

[15] LI T, HU S, BEIRAMI A, et al. Ditto: Fair and robust federated learning through personalization[C]//. International conference on machine learning: PMLR, 2021: 6357-6368.

[16] CHEN F, LUO M, DONG Z, et al. Federated meta-learning with fast convergence and efficient communication[J]. arXiv preprint arXiv:1802.07876, 2018.

[17] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint arXiv:1610.05492, 2016.

[18] QU Z, JIA N, YE B, et al. FedQClip: Accelerating Federated Learning via Quantized Clipped SGD[J]. IEEE Transactions on Computers, 2024.

[19] 吴文泰, 吴应良, 林伟伟, 等. 横向联邦学习:研究现状、系统应用与挑战[J]. 计算机学报, 2025,48(1): 35-67.

[20] HU J, DU J, WANG Z, et al. Does differential privacy really protect federated learning from gradient leakage attacks?[J]. IEEE Transactions on Mobile Computing, 2024.

[21] ZHANG J, ZHU C, GE C, et al. Badcleaner: defending backdoor attacks in federated learning via attention-based multi-teacher distillation[J]. IEEE Transactions on Dependable and Secure Computing, 2024,21(5):

4559-4573.

[22] QI F, LI S. Adaptive hyper-graph aggregation for modality-agnostic federated learning[C]//. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 12312-12321.

[23] ZHU F, TIAN Y, HAN C, et al. Model-level attention and batch-instance style normalization for federated learning on medical image segmentation[J]. Information Fusion, 2024,107: 102348.

[24] WU Z, SUN S, WANG Y, et al. Agglomerative federated learning: Empowering larger model training via end-edge-cloud collaboration[C]//. IEEE INFOCOM 2024-IEEE Conference on Computer Communications: IEEE, 2024: 131-140.

[25] CHEN Y, LI K, LI G, et al. Contributions Estimation in Federated Learning: A Comprehensive Experimental Evaluation[J]. Proceedings of the VLDB Endowment, 2024,17(8): 2077-2090.

[26] LI W, FU S, ZHANG F, et al. Data Valuation and Detections in Federated Learning[J]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 12027-12036.

[27] WU H, ZHANG L, LI S, et al. CoAst: Validation-Free Contribution Assessment for Federated Learning based on Cross-Round Valuation[C]//. Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC, Australia: Association for Computing Machinery, 2024: 1839-1847.

[28] GUO P, YANG Y, GUO W, et al. A Fair Contribution Measurement Method for Federated Learning[J]. Sensors, 2024,24(15): 4967.

[29] TASTAN N, FARES S, AREMU T, et al. Redefining Contributions: Shapley-Driven Federated Learning[C]//. International Joint Conference on Artificial Intelligence, 2024.

[30] SHLENS J. Notes on Kullback-Leibler Divergence and Likelihood[J]. ArXiv, 2014,abs/1404.2000.

[31] YOU K. Semantics at an Angle: When Cosine Similarity Works Until It Doesn't[J]. Arxiv, 2025.

[32] Z. X, S. S. FedKL: Tackling Data Heterogeneity in Federated Reinforcement Learning by Penalizing KL Divergence[J]. IEEE Journal on Selected Areas in Communications, 2023,41(4): 1227-1242.

[33] NASIM M A A, SOSHI F T J, BISWAS P, et al. Principles and Components of Federated Learning Architectures[J]. ArXiv, 2025,abs/2502.05273.

[34] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]//. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research: PMLR, 2017: 1273-1282.

[35] BO Y, SONI A K, SRIVASTAVA S, et al. Evaluating Representational Similarity Measures from the Lens of Functional Correspondence[J]. ArXiv, 2024,abs/2411.14633.

[36] HOFMANN T, SCHOLKOPF B, SMOLA A. Kernel methods in machine learning[J]. Annals of Statistics, 2007,36: 1171-1220.

[37] KORNBLITH S, NOROUZI M, LEE H, et al. Similarity of Neural Network Representations Revisited[C]//. Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research: PMLR, 2019: 3519-3529.

[38] GRETTON A, BOUSQUET O, SMOLA A, et al. Measuring Statistical Dependence with Hilbert-Schmidt Norms[C]//, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 63-77.

[39] CORTES C, MOHRI M, ROSTAMIZADEH A. Algorithms for learning kernels based on centered alignment[J]. J. Mach. Learn. Res., 2012,13(1): 795-828.

[40] MURPHY A, ZYLBERBERG J, FYSHE A. Correcting Biased Centered Kernel Alignment Measures in Biological and Artificial Neural Networks[J]. ArXiv, 2024,abs/2405.01012.

[41] STECK H, EKANADHAM C, KALLUS N. Is Cosine-Similarity of Embeddings Really About Similarity?[J]. Companion Proceedings of the ACM on Web Conference 2024, 2024.

[42] CAO X, FANG M, LIU J, et al. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping[J]. ArXiv, 2020,abs/2012.13995.

[43] KULLBACK S, LEIBLER R A. On Information and Sufficiency[J]. The Annals of Mathematical Statistics, 1951,22(1): 79-86.

[44] GAO D, YAO X, YANG Q. A Survey on Heterogeneous Federated Learning[J]. ArXiv, 2022.

[45] T. L, A. K S, A. T, et al. Federated Learning: Challenges, Methods, and Future Directions[J]. IEEE Signal Processing Magazine, 2020,37(3): 50-60.

致　谢

本文的顺利完成，离不开我的指导教师周璐教授的悉心指导与无私支持。在论文的选题、研究设计与实验验证阶段，周教授以深厚的学术造诣和严谨的治学态度，为我指明研究方向，并在关键节点提出建设性意见和研究资源支持。

最后，感谢本文参考文献中的科研工作者，正是他们的研究为本文实验的顺利完成奠定了理论基础。