

基于风险权值聚合联邦学习的金融欺诈检测模型

孙元林^{1,2}, 王楠³, 张焘³, 刘娟¹, 赵沁馨⁴

(1. 北京交通大学软件学院, 北京 100044; 2. 吉林大学符号计算与知识工程教育部重点实验室, 吉林 长春 130012;
3. 北京交通大学网络空间安全学院, 北京 100044; 4. 南京大学软件学院, 江苏 南京 210008)

摘要: 金融欺诈给金融机构和消费者带来了巨大的损失, 如何能够快速、有效、准确地识别并阻止金融欺诈行为成为人们普遍关注的问题。机器学习、深度学习等技术由于能够有效地检测动态多变的欺诈模式, 得到了广泛研究与关注。然而, 金融数据的隐私性、保密性等特点, 使得深度学习等数据驱动的智能分析技术在金融领域的实际应用中受到限制。因此, 研究如何在保障数据隐私安全的条件下进行协同训练并建立有效的欺诈检测模型具有重要的价值。针对上述问题, 提出风险权值聚合联邦学习 (risk weight aggregation federated learning, FedRWA) 算法进行多模型融合, 在保证数据隐私的前提下构建中央异常检测模型, 实现多源敏感数据模型融合学习, 使中央联合模型能够精准获取各机构私有数据集的欺诈风险特征, 以提升欺诈检测效果。为了验证所提方法的有效性, 使用信用卡交易数据集开展了实验, 实验结果表明, FedRWA 算法的性能优于多种前沿的联邦学习方法, 能够检测出更多更大额的欺诈样本, 同时进一步提升了隐私数据的安全性。

关键词: 联邦学习; 金融欺诈; 风险权值; 欺诈检测

中图分类号: TP311.1

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025056

Financial fraud detection model based on risk weight aggregation federated learning

SUN Yuanlin^{1,2}, WANG Nan³, ZHANG Tao³, LIU Juan¹, ZHAO Qinxin⁴

1. School of Software, Beijing Jiaotong University, Beijing 100044, China
2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
3. School of Cyber Space Security, Beijing Jiaotong University, Beijing 100044, China
4. School of Software, Nanjing University, Nanjing 210008, China

Abstract: Financial fraud had caused significant losses to financial institutions and consumers, making the rapid,

收稿日期: 2025-01-21; 修回日期: 2025-05-07

通信作者: 王楠, wangnanbjtu@bjtu.edu.cn

基金项目: 国家自然科学基金 (62202042); 中央高校基本科研业务费专项基金 (2024JBMC031); 航空科学基金 (ASFC-2024Z0710M5002); 先进密码技术与系统安全四川省重点实验室开放课题资助项目 (SKLACSS-202312)

Foundation Items: The Natural Science Foundation of China (62202042), Fundamental Research Funds for the Central Universities (2024JBMC031), Aeronautical Science Foundation of China (ASFC-2024Z0710M5002), OpenFund of Advanced Cryptography and System Security Key Laboratory of Sichuan Province (SKLACSS-202312)

引用格式: 孙元林, 王楠, 张焘, 等. 基于风险权值聚合联邦学习的金融欺诈检测模型[J]. 网络与信息安全学报, 2025, 11(5): 88-100.

Citation Format: SUN Y L, WANG N, ZHANG T, et al. Financial fraud detection model based on risk weight aggregation federated learning[J]. Chinese Journal of Network and Information Security, 2025, 11(5): 88-100.

effective, and accurate identification and prevention of financial fraud a widely concerned issue. Machine learning and deep learning technologies had received extensive research attention due to their capability to effectively detect dynamic and changing fraud patterns. However, the privacy and confidentiality requirements of financial data were found to limit the practical application of data-driven intelligent analysis technologies, including deep learning, in the financial field. Therefore, research on how to conduct collaborative training and establish effective fraud detection models while ensuring data privacy and security was considered to be of great value. To address these challenges, a risk weight aggregation federated learning (FedRWA) method was proposed for multi-model fusion. A central anomaly detection model was constructed to achieve multi-source sensitive data model fusion learning while ensuring data privacy. This approach enabled the central joint model to accurately capture the fraud risk characteristics of each institution's private dataset, thereby improving fraud detection effectiveness. To verify the effectiveness of the proposed method, experiments were conducted using a credit card transaction dataset. The experimental results demonstrate that the FedRWA method performs better than numerous cutting-edge federated learning methods, could detect more and larger fraud samples, and further enhances the security of private data.

Keywords: federated learning, financial fraud, risk weight, fraud detection

0 引言

近年来,网贷、消费贷款等新的互联网金融产品不断涌现,极大地丰富了金融服务类型,给大众日常生活带来诸多便利,但同时滋生出更加复杂、隐蔽的欺诈风险。金融欺诈涵盖信用卡欺诈、保险欺诈、洗钱,以及证券与商品欺诈等多种形式,其恶劣影响广泛蔓延,已引起人们的高度关注。金融欺诈不仅会造成经济损失,其衍生的各类犯罪形态(如有组织的犯罪和贩毒等非法活动^[1])还会在社会层面产生广泛而负面的影响。鉴于金融欺诈现象的日益复杂与高发,欺诈防御系统的研究与开发已成为当下的焦点。

传统的欺诈防御系统借助规则引擎进行欺诈行为检测,即将行业专家积累的经验转化为机器能够识别并执行的规则。此类基于规则的系统会对所有传入的交易展开扫描,寻找可疑模式。以一条典型规则为例:若某持卡人在某个国家/地区完成某笔金额的交易,并且次日又在另一规则给定的国家/地区进行另一笔规则给定金额的交易,系统便判定欺诈行为发生^[2]。这类规则可以避免公式化的欺诈行为,但极易给正常用户的合理操作造成不必要的困扰。与此同时,梳理并制定各类风控规则也是一项复杂且耗时费力的任务。当下,金融领域的业务场景在不断丰富,交易模式更是千变万化,所产生的交易数据呈现出海量增长的态势。在此情形下,单纯依靠人工方式进行分析,已无法设计出切实有效的判断规

则。综上所述,传统基于规则引擎的欺诈检测手段已难以满足当下金融行业对欺诈防御的严苛要求。

机器学习和深度学习等技术的蓬勃发展,为金融业的风控工作提供了新思路。这些技术能够有效解决传统风控方式效率低、精度差、不能预测未知风险等难题,因此被广泛应用于欺诈检测。逻辑回归、支持向量机、决策树、朴素贝叶斯、随机森林和XGBoost等都是常用的模型。深度学习等相关算法的出色效果,高度依赖大量且全面的数据信息。然而,金融数据具有敏感、隐私的特性,导致不同金融机构之间的数据无法互通。各机构只能利用自身内部观测到的数据训练模型,分散在不同金融机构的欺诈交易数据无法被整合起来用于集中式的模型训练。这极大地制约了数据驱动的机器学习模型的效果^[3]。

联邦学习作为一种创新性的分布式机器学习模型训练范式,能够在确保数据保留在本地、不外泄的情况下,联合多个数据节点协作建模,构建隐私保护与知识共享兼具的模型,为多金融机构联合建模提供了全新的解决思路。

不过,在实际应用场景中,联邦学习方法暴露出诸多短板。例如,在欺诈检测环节,正负类样本比例严重失衡,极大地干扰了本地模型与联合模型的性能表现^[4-5]。同时,各金融机构间欺诈样本数据的分布也存在不均衡现象,具体体现为欺诈样本数量有多寡之分,欺诈情况的严重程度参差不齐。传统的联邦学习方法往往直接依据

各节点数据集规模对多个模型进行加权聚合。但在金融欺诈检测这类异常检测场景中,这种方式无法取得最优效果。

本文针对金融欺诈检测领域中数据集类别严重失衡,以及多金融机构间欺诈样本分布不均衡问题展开深度剖析,并创新性地提出基于数据集样本的欺诈特征统计模型与风险权值聚合联邦策略,简称为风险权值聚合联邦学习(risk weight aggregation federated learning, FedRWA),将其应用于欺诈检测场景。该方法能够在强化金融机构数据隐私安全保障的同时,进一步提升联邦学习理论技术在欺诈检测这类异常检测领域的应用效果。具体而言, FedRWA 在模型参数更新环节融合了每个本地数据集中的欺诈统计特征。与此同时,在模型训练过程中,通过对数据集进行动态类别平衡处理,有效引导模型学习到更加丰富、全面的欺诈特征,提升模型识别复杂欺诈模式的能力。

为了验证 FedRWA 方法的有效性,本文选用欧洲信用卡交易数据集 Credit Card Fraud Detection 展开实验。实验结果表明, FedRWA 在性能表现上优于市面上常见的联邦学习算法。

1 相关工作

1.1 金融欺诈检测

在全球范围内,金融欺诈事件层出不穷,近年来更是呈现出持续攀升的态势。加拿大反欺诈中心发布的报告显示,2022年,该国记录在案的欺诈事件多达 91 190 余起,致使 57 055 名受害者遭受损失,金额累计超 5.31 亿美元^[6]。

Abdallah 等^[7]的研究指出,当前主流的欺诈防范机制涵盖欺诈防御系统(fraud prevention systems, FPS)与欺诈检测系统(fraud detection systems, FDS)。FPS 作为金融机构抵御欺诈的首道防线,主要运用限制、销毁、管控、删除等手段,防范外部势力针对金融业务系统的硬件与软件发起网络攻击。其核心手段包括部署防火墙、应用加密算法和使用电子签名等。尽管 FPS 发挥了一定的防御功效,然而在许多情形下,欺诈者凭借多样的技术手段与伪装操作,依然能够突破防线,侵入金融业务系统的操作流程之中。而 FDS 则对那些已越过 FPS、成功进入业务系统的

欺诈交易行为,展开检测、报告与拦截操作。该系统进一步保障了金融机构与消费者的财产安全,已然成为金融机构实施欺诈风险防控的关键举措。

以往,银行主要借助统计分析和制定规则等手段来规避风险^[8]。具体来说,银行依靠业务专家对欺诈模式展开深入分析,构建包含欺诈交易特征的特征库,并在此基础上建立特征模型,进而制定出相应的判断规则,以此决定是拒绝还是放行金融交易。然而,随着移动互联网时代的来临,各类金融产品和消费场景不断涌现,用户产生的交易数据呈现出海量增长的态势。在这样的情况下,仅仅依靠人力根本无法对如此庞大的数据信息进行有效的分析处理,更难以精准捕捉那些复杂多变的欺诈模式,从而实现对金融欺诈的有效防控。

近年来,机器学习和深度学习等先进技术的兴起,为金融业的风控工作带来了全新的思路,智能风控被广泛应用于金融科技领域。智能风控体系以用户海量的历史数据为基础,经过数据清洗、特征工程、算法模型构建、模型训练评测与优化等一系列流程,最终生成能够有效判别欺诈模式、预测欺诈风险的智能模型。

在相关研究方面, Sahin 等^[9]指出,传统的欺诈预防机制并不能有效地预防常见类型的欺诈,并提出使用成本敏感的决策树技术来加以改进。Sahin 等^[10]对信用卡欺诈检测所面临的挑战进行了深入分析,同时比较了人工神经网络和逻辑回归在信用卡诈骗检测方面的性能。Gyamfi 等^[11]运用支持向量机技术来检测和预防银行欺诈行为,表明了大数据和机器学习技术在银行欺诈检测与预防工作中的重要性。Alghofaili 等^[12]开发出了一种基于 LSTM 技术的金融欺诈检测模型。尽管人工智能模型在金融欺诈检测领域得到了广泛的认可,并且众多机构也进行了大量的尝试,但金融数据本身具有敏感和隐私的天然特性,导致不同金融机构之间的数据无法实现互通共享。因此,各金融机构只能利用自身组织内部所观测到的数据样本进行模型训练。这种数据获取的局限性,使得那些依赖大量且全面训练数据的机器学习模型的实际效果大打折扣。

1.2 联邦学习

2016年, Google的研究人员 McMahan 等^[13]提出了联邦学习的概念, 其核心宗旨在于解决不同节点数据集在严格遵循数据隐私保护原则的前提下, 实现联合训练以提升模型性能的问题。联邦学习采用联邦平均 (federated averaging, FedAvg) 算法, 其致力于降低通信成本。具体做法是增加本地模型训练的迭代次数, 让本地模型在性能得到显著提升后, 再将相关参数上传至中央服务器进行聚合。自联邦学习理论诞生以来, 迅速引起了学术界和工业界众多学者的高度关注。此后, 围绕联邦学习算法架构的优化, 以及其在各个领域的应用, 一直都是研究的热点。

在相关研究成果方面, Li 等^[14]针对异质网络中各节点数据的联邦学习问题, 提出了 FedProx 算法。该算法通过在模型损失计算过程中引入近似项, 从理论层面保障了模型的收敛性。相较于传统的联邦平均算法 (FedAvg), FedProx 算法具有更强的鲁棒性。应作斌等^[15]提出了一个具备隐私保护功能的联邦学习框架, 该框架采用动态权重聚合 (DP-DFL) 的方式, 有效解决了现有联邦学习框架在面对不可信中央服务器时, 存在的数据非独立同分布问题以及隐私泄露隐患。余锋等^[16]提出了一种基于差分隐私生成对抗网络 (DP-enabled GAN) 模型的隐私增强型联邦学习方案。该方案的优势在于, 能够在不牺牲模型准确性的前提下提升数据的隐私保护能力。Bonawitz 等^[17]提出了模型对比损失的概念。在本地模型的训练过程中, 通过增加模型对比损失项, 使模型在迭代过程中能够更紧密地趋近于全局模型, 同时远离上一轮次的本地模型, 进而实现更好的全局模型性能。Wang 等^[18]提出的 FedNova 模型解决了目标不一致性问题, 并且在保持快速误差收敛方面表现出色。

尽管上述这些模型在提高模型鲁棒性、保护客户端数据隐私、加快收敛速度和提升中央模型性能等方面取得了显著的成果。然而, 它们普遍存在一个共同的缺陷, 即没有充分考虑到不同客户端模型之间的差异, 以及不同数据分布所带来的问题。这一不足导致全局模型在对部分群体进行预测和应用时, 泛化能力较差, 无法达到理想

的效果。

为了有效应对不同数据分布和计算能力存在差异问题, 诸多新方法采用了加权聚合策略。其中, Li 等^[19]提出的 FEDLAW 方法引入了可学习的聚合权重, 显著提高了全局模型的泛化性能。Deng 等^[20]提出的 FAIR 框架, 基于梯度下降方法设计自动学习最优聚合权重算法, 显著提高了全局学习模型的质量。张连福等^[21]提出的 DP-FedAWA 方法, 依据各模型的归一化得分, 按照加权平均的方式进行聚合得到全局模型, 确保那些表现欠佳的异常更新被赋予较低的权重。

然而, 在金融欺诈这一特定领域中, 由于数据集普遍存在类别严重不平衡的问题, 同时各机构之间的欺诈样本分布也不均衡, 上述这些通用模型往往难以发挥出最佳的效果。

尽管联邦学习技术能够在保障数据隐私安全的前提下打破数据孤岛, 协作多家银行机构利用各自的本地数据进行联合建模。然而, 在实际的应用过程中, 联邦学习仍然面临着诸多严峻的挑战。例如, 欺诈检测中的数据类别严重不平衡问题、多金融机构合作激励问题、联合模型效果提升问题、通信开销问题^[22]和不同客户端算力问题^[23]等。

2 基于风险权值聚合联邦学习的金融欺诈检测方法

风险权值聚合联邦学习算法面向金融欺诈检测等需要保证数据隐私安全的异常检测领域。它首先统计各机构本地数据集的欺诈特征, 利用欺诈特征进行风险权值计算, 结合风险权值更新模型参数, 通过风险特征与模型信息的聚合迭代, 从而更好地捕获数据集中的欺诈样本特征, 提升欺诈检测效果。模型迭代流程如图1所示。下面结合图1对风险权值聚合联邦学习算法模型各模块进行详细介绍。

2.1 本地客户端训练

基于联邦学习理论, 搭建起中央服务器与多节点联合训练架构。将各个本地机构视为客户端, 构建相同的网络模型结构。所有本地模型的初始参数均由中央服务器统一分发, 随即开启模型训练的迭代流程。在每一轮通信过程中, 中央

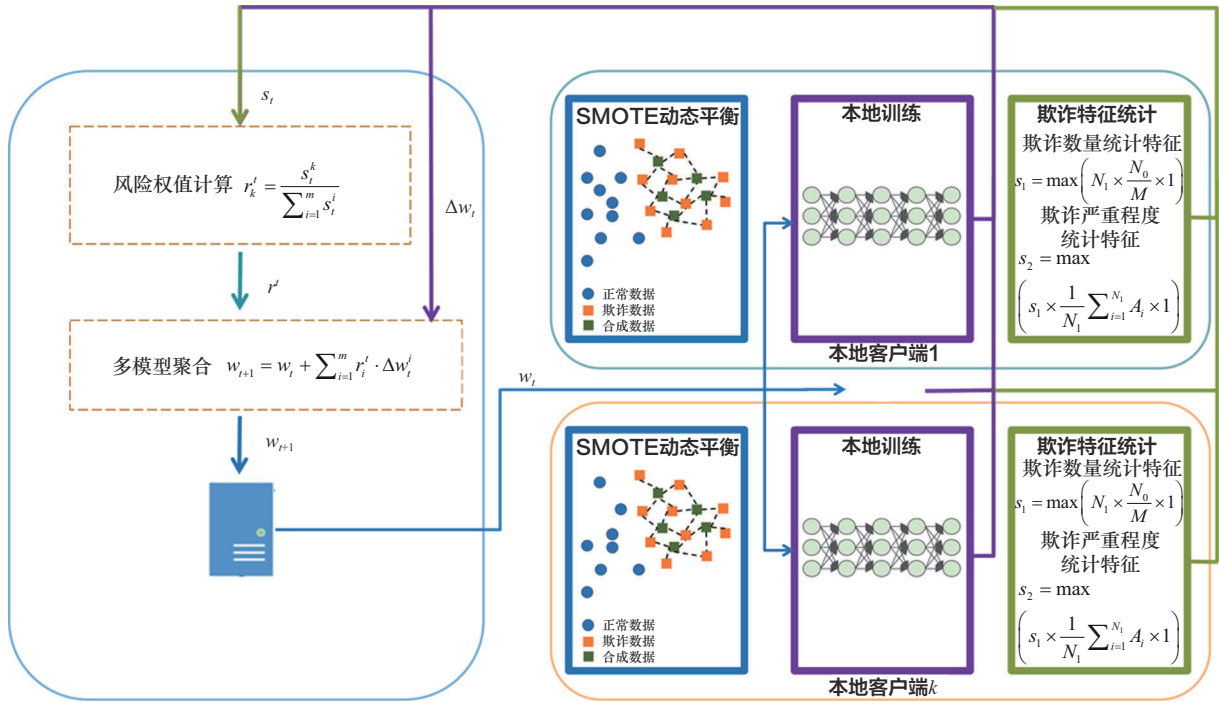


图1 风险权值聚合联邦学习模型迭代流程

Figure 1 Iteration process of risk weight federated aggregation learning model

服务器会随机挑选多家金融机构参与联邦模型的训练工作，并将当前联合模型的参数 w_t 同步下发至各选定机构。

2.1.1 SMOTE 动态平衡

为契合风险权值聚合联邦学习算法的运用需求，同时有效规避模型训练过程中对某一次平衡后的数据集过度拟合问题，本文创新性地提出客户端本地数据集动态平衡策略。该策略在每一轮通信训练开启前实施，针对参与本轮训练的客户端节点私有数据，通过改变 SMOTE 算法的随机种子，重新进行类别平衡处理。这样的动态平衡处理机制能够让本地模型从更多角度学习欺诈样本的特征模式，提升聚合后的中央联合模型的欺诈检测效果，使模型拥有更好的泛化能力。经过 SMOTE 算法平衡后，将对平衡后的数据集进行训练。

2.1.2 本地训练

对于第 k 个机构的本地私有数据集 D_k ,

$$D_k = \left\{ \{x_1, y_1\}, \dots, \{x_{n_k}, y_{n_k}\} \right\} \quad (k = 1, 2, \dots, K) \quad (1)$$

其中， n_k 表示第 k 个金融机构拥有的数据样本量； x_i 表示特征向量， y_i 表示标签， $i = 1, \dots, n_k$ ； K 表示联邦学习参与者（金融机构）的数量。

中央联合模型参数 w_{t+1} 与本地模型参数 w_{t+1}^k 的关系为

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (2)$$

其中， n 表示所有机构节点数据样本量的总和。中央联合模型的训练目标是

$$\min_{w \in \mathbb{R}^d} l(x, y; w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n l(x_i, y_i; w) \quad (3)$$

根据数据集大小在各本地机构的分布，中央联合模型优化目标函数为

$$l(x, y; w) = \sum_{k=1}^K \frac{n_k}{n} L_k \quad (4)$$

其中， L_k 表示各机构本地模型训练的损失函数。 L_k 定义为

$$L_k = \frac{1}{n_k} \sum_{i=1}^{n_k} l(x_i, y_i; w) \quad (5)$$

在所有客户端都参与训练且本地模型采用随机梯度下降（stochastic gradient descent, SGD）更新的情况下，令第 t 轮通信 k 个机构节点本地模型训练更新的梯度 $g_k = \nabla L_k(x_k, y_k, w_t)$ ，总的联合模型更新梯度 $\nabla l = \sum_{k=1}^K \frac{n_k}{n} g_k$ ，则一轮更新后联合模型参数表示为

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^K \frac{n_k}{n} g_k \quad (6)$$

则当前轮次每个客户端本地更新的模型参数可表示为

$$w_{t+1}^k \leftarrow w_t - \eta g_k \quad (7)$$

结合式(6)和式(7), 同样可以得到中央联合模型参数与本地模型参数的关系, 即式(2)。

在联邦平均聚合算法中, 为了提高通信效率和保护数据的隐私安全, 引入 C 、 B 、 E 3 个关键参数。在每轮通信中随机选择比例为 C , 即个数为 $\max(K \cdot C, 1)$ 的客户端进行本地训练, 且每个客户端在本地训练过程中采用小批量梯度下降法 (MBGD) 每次使用批次大小为 B 的数据进行训练。当训练完本地所有数据时为 1 轮, 客户端本地进行 E 轮迭代训练。FedRWA 算法沿用这里的 3 个参数进行训练, 训练完毕, 提取更新后的模型参数 Δw_k 。

$$\Delta w_k = w_{t+1}^k - w_t \quad (8)$$

2.1.3 计算欺诈统计特征

本文根据数据集中欺诈样本的数量及金额等统计特征进行更深一层的分析, 以提升联邦学习模型对异常样本的学习效果。本文提出以下两个模型用来计算数据集的欺诈统计特征。

(1) 欺诈数量统计特征

在数据集中, 正常交易样本数量充足。而在欺诈检测情境下, 正常样本与异常欺诈样本的重要程度存在差异。用 $I = 1:M$ 表示 1 个欺诈样本的重要程度与 M 个正常样本相当。这里, M 可以根据正负类样本不平衡比例, 或是结合不同样本错分所产生的代价差异等实际业务场景灵活确定。若数据集中包含 N 个样本, 其中有 N_0 个正常样本, N_1 个欺诈样本, 则该数据集的欺诈数量统计特征值 s_1 为

$$s_1 = \max \left(N_1 + \frac{N_0}{M}, 1 \right), M \in [1, +\infty] \quad (9)$$

由式(9)可知, 当 M 取值为 1 时, 意味着正常样本与欺诈样本的重要程度等同, 此时欺诈数量统计特征值就等于数据集的样本总数。而当 M 取值为正无穷大时, 相当于只聚焦数据集中的欺诈样本, 完全忽略正常样本。鉴于实际中正常样本数量极为充裕, 这种处理方式在该情境下具备合理性。

在一些特殊情形下, 若某个金融机构的数据集里不存在欺诈样本, 仅观测到正常交易样本, 即面临单类别学习问题^[24]。为保证此机构节点的数据集能够正常参与模型训练与聚合流程, 令 $s_1=1$, 给予它较低的权重。在此过程中所提取的特征值 s_1 标记为 FedRWA- S_1 。

(2) 欺诈严重程度统计特征

为了更深入地探究数据集中样本的特征, 本文对数据集中样本的欺诈严重程度统计特征展开分析。为了方便计算和理解, 本文以欺诈金额作为评定欺诈样本严重程度的基准。欺诈样本的金额越大, 意味着该欺诈行为的严重程度越高。用 A 表示欺诈样本的金额。首先需要计算出机构所观测到的所有欺诈样本的平均金额, 然后结合前面所得到的欺诈数量统计特征, 便可以进一步得到数据集中欺诈严重程度统计特征值 s_2 , 其具体的计算方式如下。

$$s_2 = \max \left(s_1 \times \frac{1}{N_1} A, 1 \right) \quad (10)$$

结合式(9)进行深入分析, 可以发现当 M 取值为正无穷大时, 式(10)中的欺诈严重程度统计特征值就等同于数据集中所有欺诈样本的金额总和。同样地, 为了确保那些仅包含单类样本的数据集能够正常参与模型的训练, 将欺诈严重程度统计特征值 s_2 的最小值设为 1。此阈值可以根据实际应用场景灵活调节。此处提取的特征值 s_2 记为 FedRWA- S_2 。

客户端将计算得到的欺诈统计特征 S (FedRWA- S_1 或 FedRWA- S_2) 与模型参数更新信息 Δw 传递给中央服务器。

2.2 中央服务器聚合

2.2.1 风险权值计算

从客户端得到欺诈统计特征后, 计算风险权值, 具体计算方式如下。

$$r_k^t = \frac{s_t^k}{\sum_{i=1}^m s_t^i} \quad (11)$$

其中, m 表示本轮次联邦学习训练过程中用来聚合模型的节点数量, s_t^k 表示每个客户端节点私有数据集的欺诈统计特征值, 其计算方法可从第 2.1.3 节所阐述的两种模型中任选其一。 r_k^t 满足密度分布 ($\sum_{k=1}^m r_k^t = 1$)。当完成风险权值的

计算后, 所得到的风险权值将被应用于中央模型参数的更新操作。

2.2.2 多模型聚合

当风险权值计算工作结束之后, 要依据各节点本地数据集的欺诈特征值 s_i^k 对应的权重 r_k^t 进行模型聚合, 如式(12)所示。

$$w_{t+1} = \sum_{k=1}^m r_k^t \cdot w_{t+1}^k \quad (12)$$

其中, r_k^t 表示在第 t 轮通信中参与该轮次训练的 k 个客户端节点数据集的欺诈风险权值。

第 k 个客户端在第 t 轮通信中经过本地训练更新后的模型参数 w_{t+1}^k 如下。

$$w_{t+1}^k = w_t + \Delta w_t^k \quad (13)$$

其中, Δw_t^k 表示在第 t 轮通信中参与训练的第 k 个客户端的模型更新信息。结合式(12)和式(13), 可以得到式(14)。

$$w_{t+1} = w_t + \sum_{k=1}^m r_k^t \cdot \Delta w_t^k \quad (14)$$

式(14)表示中央服务器在进行多模型融合时不再直接聚合模型参数, 而是收集各模型更新信息, 并依据风险权值策略进行聚合。这样, 在通信过程中, 不再需要传输庞大的模型参数, 仅需传递当前轮次的模型更新信息即可, 从而避免模型参数在传输过程中被第三方恶意截取, 进一步保障了金融机构数据和模型的安全。

在完成权值聚合操作后, 中央服务器会把更新后的模型参数下发至各本地节点。这些参数将在各节点持续下放、迭代, 直至模型达到收敛条件。

FedRWA 算法的详细流程见算法 1。

算法 1 FedRWA 算法

输入 各金融机构独立保存的私有数据集

参数: C, B, E 参数含义同联邦平均算法, M 为欺诈样本与良性样本重要性比值。欺诈特征值 s 计算以式(9)为例。

输出 联合欺诈检测模型参数 w

ServerUpdate() 中央服务器 Server 端执行:

初始化模型参数 w_0

FOR each 通信轮次 $t = 1, 2, \dots, T$ DO

随机从 K 个客户端中选取 m 个客户端, 该集合记作 C_t

FOR each C_t 中的客户端 DO

$s_k^t, \Delta w_t^k \leftarrow \text{ClientUpdate}(k, w_t)$

计算风险权值: $r_k^t \leftarrow \frac{s_k^t}{\sum_{i=1}^m s_i^t}$

更新联合模型: $w_{t+1} \leftarrow w_t + \sum_{k=1}^m r_k^t \cdot$

Δw_t^k ClientUpdate(i, w_t) 客户端执行:

接收中央服务器下发的参数 w_t , 更新本地模型: $w \leftarrow w_t$

基于数据集信息计算欺诈统计特征值:

$$s_t^k = \max(N_1^k + \frac{N_0^k}{M}, 1)$$

改变随机种子, SMOTE 算法平衡数据集

将本地数据集拆分为 B 大小的批次, 记

为集合 U

FOR each 本地训练迭代轮次 $t' = 1, 2, \dots$ DO

FOR each 集合 U 中批次数据 DO

$w \leftarrow w - \eta \nabla L(x, y; w)$

提取模型参数更新信息: $\Delta w_t^k \leftarrow w - w_t$

返回 $s_t^k, \Delta w_t^k$

3 实验设计与分析

3.1 数据集介绍与预处理

在联邦学习研究中, 从不同机构实际采集数据集非常困难。根据联邦学习的思想架构, Kairouz 等^[25]提出把现有的公开数据集, 按不同的样本分布切割成大小各异的子集, 将这些子集当作各节点的数据, 以此模拟仿真多节点联邦学习的现实应用场景。

在本文的实验分析中, 所采用的数据集是由 Machine Learning Group - ULB 提供的欧洲信用卡交易数据集 Credit Card Fraud Detection。该数据集包含 492 个欺诈样本, 仅占数据集总量的 0.17%, 正常样本与欺诈样本的数量比例为 578:1, 是一个典型的类别高度不平衡数据集, 这与金融领域异常欺诈检测任务所涉及数据的一般特征高度契合。

基于此数据集, 在实验环节设置了不同数量的客户端及其数据集大小比例, 进行随机样本选择分配, 模拟实际的多金融机构节点信用卡交易和联邦学习的应用场景。具体而言, 实验时将该数据集按照 80% 和 20% 的比例拆分为训练集和

测试集。数据拆分后的训练集与测试集的样本详细情况已统计在表1中。

表1 训练集和测试集信息

Table 1 Information for training sets and testing sets

对比项	总样本数量/个	欺诈样本数量/个	欺诈样本金额总计/元
总数据集	284 807	492	60 127.97
训练集	227 845	394	47 781.35
测试集	56 962	98	12 346.62

接下来,将训练集根据实验条件设置拆分为 K 个子集,用于模拟实际应用场景中多个金融机构各自独立持有私有数据的情形。

3.2 模型评价指标

实验分析中混淆矩阵见表2。

表2 混淆矩阵

Table 2 Confusion matrix

真实值	预测值	正类0	负类1
正类0		TN	FP
负类1		FN	TP

基于混淆矩阵,本文评估模型性能使用的指标如下。

1) G -mean: 为正类和负类的分类准确率的几何平均值。对于类别不平衡数据分类任务,该指标能够很好地体现模型预测的准确率^[26]。

$$G\text{-mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (15)$$

2) Recall: 欺诈样本的召回率,即检测出的欺诈样本数量在所有欺诈样本数量中的占比。

$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN} \quad (16)$$

3) AUC: ROC曲线下的面积。ROC曲线由模型在不同阈值分类下得到的一系列模型的(FPR,TPR)点组成,AUC取值范围为[0,1],AUC值越接近1意味着模型的分类效果越好。

4) Amount-pred (模型检出欺诈样本金额): 该指标由本文提出,用于衡量模型在欺诈检测中检出的欺诈样本的合计金额。该指标能够体现模型为金融机构挽回的资金损失,是在金融欺诈检测场景下机器学习模型召回率指标的有效补充。Amount-pred的计算方式如下。

$$\text{Amount-pred} = \sum_{i \in U_{TP}} A_i \quad (17)$$

其中, U_{TP} 表示模型预测出的正类样本(即欺诈样本)集合, A_i 表示某个欺诈样本的交易金额。

3.3 对比模型

为了评估FedRWA算法及本文提出的欺诈特征统计模型的性能,实验时分别采用两种风险计算方式,并调整重要性比值 M ,分别记为FedRWA- S_1 -M578、FedRWA- S_1 -MInf、FedRWA- S_2 -M578和FedRWA- S_2 -MInf, S_1 和 S_2 表示欺诈数量统计特征计算方式,即第2.1.3节所阐述的内容。M578表示式(9)中 M 取值为578,MInf表示 M 为正无穷。实验条件设置见表3。

表3 FedRWA算法的实验条件设置

Table 3 Experimental setting for FedRWA

模型	欺诈特征统计模型	M
FedRWA- S_1 -M578	S_1	578
FedRWA- S_1 -MInf	S_1	$+\infty$
FedRWA- S_2 -M578	S_2	578
FedRWA- S_2 -MInf	S_2	$+\infty$

为了验证FedRWA算法的优越性,实验共采用以下几种模型来评估。

1) Center (数据集中式模型): 直接训练一个数据集中式的中心模型。在该场景下,无须考虑数据隐私特性,模型能够获取并处理所有的数据信息,并且可以基于训练集中的全部数据,先进行采样平衡处理,通过调整各类数据样本的比例,使得数据分布更加合理,然后再对模型进行充分的训练。

2) FedAvg^[13]: FedAvg是一种用于联邦学习的优化算法,它通过在每轮通信中对参与方的局部梯度进行加权平均来更新全局模型参数。

3) FedProx^[14]: FedProx是专门应用于联邦学习领域的一种框架,它能够有效应对联邦学习过程所面临的诸多挑战,如数据隐私保护问题、数据与设备的异构性问题,以及大规模分布式网络环境下的复杂状况等。该框架允许依据不同设备的资源状况,在各个设备上执行数量各异的本地计算任务,以此来适配训练过程中因设备和数据差异而产生的异构性。与经典的FedAvg框架相

比, FedProx 在设计上更加灵活, 是一个普适性更强的联邦学习框架。

4) MOON^[27]: MOON 是一种基于 FedAvg 的简单而有效的算法。它通过在本地训练阶段引入轻量级但新颖的修改来解决联邦学习中非独立同分布数据的挑战。该算法使用一个基础编码器、一个投影头和一个输出层来构建网络架构, 并通过最小化本地目标函数来更新本地模型。

5) FedBN^[28]: FedBN 是一种新型的联邦学习算法。本文用它解决了特征偏移这一关键问题, 即本地客户端所拥有的样本示例呈现出与其他客户端截然不同的特征分布。FedBN 算法的核心思路在于, 先运用本地批量归一化技术来缓解特征偏移现象, 使得不同客户端间的特征分布差异得以缩小, 在此基础上再对模型进行平均处理。FedBN 算法优于先前的各类联邦学习算法, 并且收敛速度更快。

6) FedRN^[29]: FedRN 借助可靠邻居机制来缓解数据异质性和标签噪声导致的低性能和高性能客户端之间的性能差异。FedRN 通过与可靠邻居的协同合作来构建高斯混合模型, 以此从目标客户端的噪声数据中识别干净的样本。低性能模型从其可靠邻居处获益更多, 有助于提高全局模型的整体鲁棒性。

3.4 实验设置

为了验证联邦学习训练生成模型的实际效果, 所有模型均基于同一 MLP 模型架构开展训练。在训练起始阶段, 将 MLP 模型的初始化参数进行持久化保存, 确保后续所有模型启动训练时, 均采用这一套已保存的初始化参数, 以此保证实验初始条件的一致性。在实验过程中, 固定

参数 $B=128$, $E=2$, $C=0.4$ 。

3.5 实验过程与结果分析

为了全面对比模型的性能, 共开展了两次实验。在实验 1 中, 所有参与训练的数据集大小相等。而在实验 2 中, 将数据集划分为大小不等的子集, 以此模拟现实场景中不同机构所拥有的数据量存在差异的状况。

3.5.1 等量数据实验

在实验 1 中, 将训练集拆分为 30 个样本数量和大小相等的子集, 用于模型训练。图 2 统计了拆分后的 30 个数据子集中的欺诈样本数量分布情况, 可以看到, 虽然各节点的总数据样本量相等, 但由于欺诈样本的出现具有随机性, 每个节点所观测到的欺诈样本数量并不相同, 部分机构节点私有数据集中包含的欺诈样本数量相差较大。图 3 则显示不同节点所观测到的欺诈样本合计金额分布更加不均衡, 各机构节点间欺诈样本金额差别较大, 分布无明显规律。

各模型均经过 200 轮次训练迭代收敛得到。FedRWA 算法与各联邦学习算法的性能数据统计在表 4 中。从表 4 中数据对比可以看出, FedRWA 的性能整体优于 Center。FedRWA 由于更精准地捕捉了各节点数据集的欺诈风险特征, 在 G -mean、Recall 和 Amount-pred 指标上相比于 FedAvg 都有进一步的提升, 部分实验条件下 AUC 值也得到显著提升。FedProx 在 AUC 指标上较有优势, 但是在 Recall、Amount-pred 指标上与 FedAvg 效果相当。MOON 在 Recall 指标上有进一步的提升, 然而在 Amount-pred 指标上偏低, 这体现了传统算法仅考虑样本数量方面的评价指标所带来的不

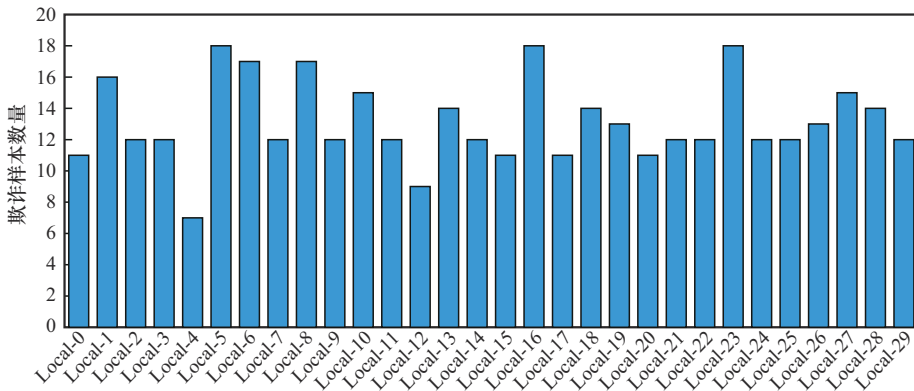


图 2 欺诈样本数量分布(等量数据)

Figure 2 Distribution of fraud samples (equal amounts of data)

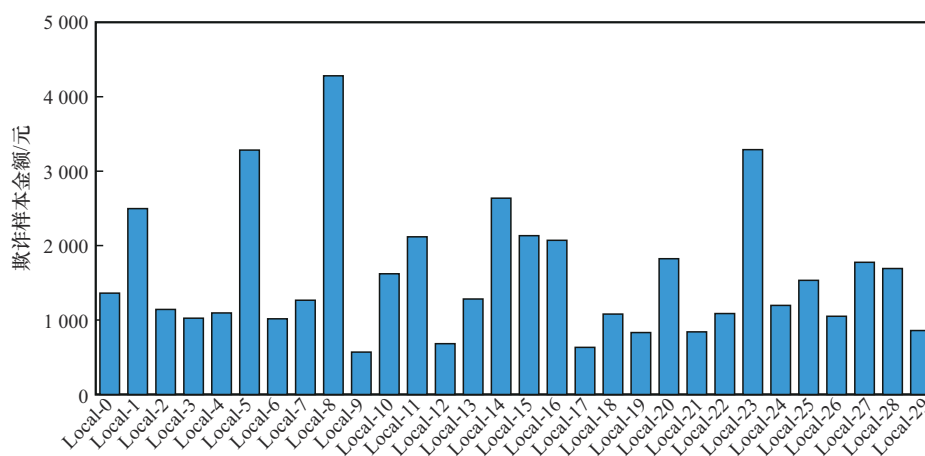


图3 欺诈样本金额分布(等量数据)

Figure 3 Distribution of fraudulent amounts (equal amounts of data)

足。与FedBN相比，FedRWA在大多数情况下的G-mean、AUC和Recall指标占有微小优势或与之相当，其Amount-pred指标在所有情况下均占据明显优势。相比于FedRN，FedRWA表现出了更高的优越性，在4个指标上都大幅领先。

表4 FedRWA与各联邦学习算法性能对比(等量数据)
Table 4 Performance comparison between FedRWA and various federated learning algorithms (equal amounts of data)

算法	G-mean	AUC	Recall	Amount-pred
Center	0.925	0.950	0.857	9 965.4
FedAvg	0.930	0.948	0.867	9 973.4
FedProx	0.931	0.956	0.867	9 973.4
MOON	0.936	0.946	<u>0.878</u>	8 621.2
FedBN	0.936	0.953	<u>0.878</u>	8621.1
FedRN	0.742	0.870	0.551	1659.5
FedRWA-S ₁ -M578	0.936	0.946	<u>0.878</u>	<u>9 974.4</u>
FedRWA-S ₁ -MInf	<u>0.941</u>	0.953	0.888	9 975.4
FedRWA-S ₂ -M578	0.942	0.938	0.888	9 975.4
FedRWA-S ₂ -MInf	0.931	<u>0.954</u>	0.867	<u>9 974.4</u>

注：最优值用粗体标记、次优值用下划线标记。

值得注意的是，在FedRWA-S₂-MInf实验条件下，以各节点的欺诈金额统计值为权值进行聚合，虽然其在Recall指标上与FedAvg算法相等，但是在Amount-pred指标上获得了提升，对比MOON算法与FedRWA-S₁-M578实验条件下所得模型也可以得出类似结论。这说明本文所提Fe-

dRWA算法能够给予金额较大的欺诈样本更多的关注，从而为金融机构检测出更多大额欺诈样本，提升欺诈检测的实际应用价值。

3.5.2 不等量数据实验

这里进一步模拟实际生活中不同的金融机构拥有不同大小数据集的场景，将训练集划分为20个样本数量和大小不等的子集，用于模型训练。20个金融机构节点样本的拆分数量比例设置为1:2:2:3:3:3:4:4:4:1:2:2:3:1:2:2:3:3:4:3:1:2:3，划分后的数据集中欺诈样本的数量分布如图4所示、欺诈样本金额分布如图5所示。其中折线表示数据集划分的比例，可以看出，在各金融机构的数据集大小不同的情况下，欺诈样本的数量和金额分布与数据集大小比例是相关的。

各模型同样经过200轮次训练迭代收敛得到。FedRWA算法与各联邦学习算法的性能对比见表5，可以看出，相比实验1等量划分数据集场景，FedRWA在Recall和Amount-pred指标上取得了更优的结果，尤其是在Amount-pred指标上提升比较明显。从表5中数据可以看出，FedProx和MOON算法对应的模型在部分指标上相比FedAvg算法所对应的模型获得了小幅提升，但是总体性能不如本文提出的FedRWA算法。与FedBN相比，FedRWA在各种条件下G-mean、Recall和Amount-pred指标都表现出明显优势，在大多数条件下AUC值表现相当，总体而言，FedRWA优于FedBN。相比于FedRN，FedRWA的优势依然明显，在各项指标上均有大幅提升。

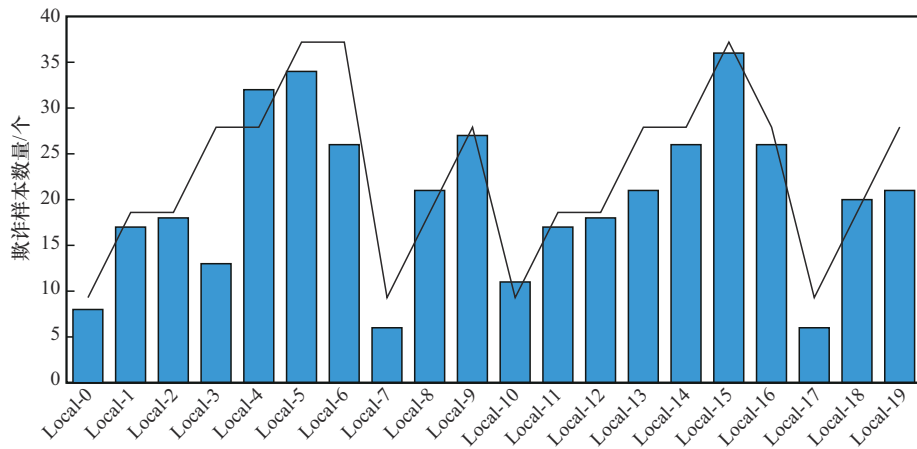


图4 欺诈样本数量分布(不等量数据)

Figure 4 Distribution of fraudulent sample (unequal amounts of data)

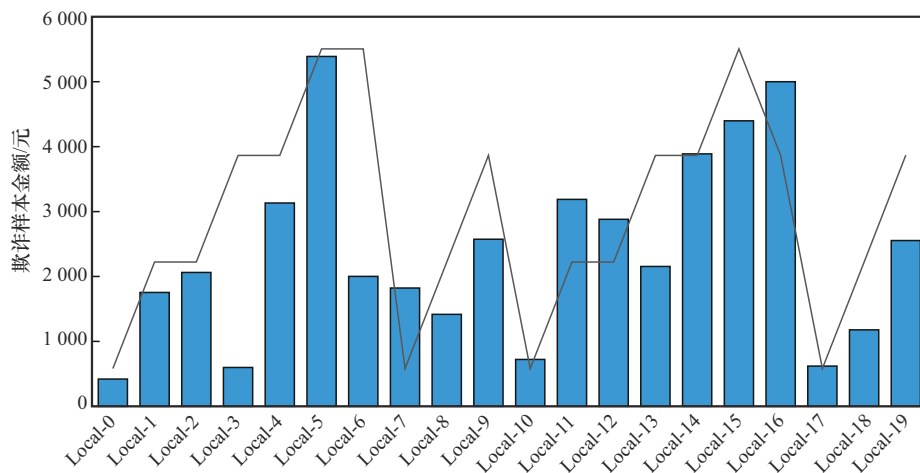


图5 欺诈样本金额分布(不等量数据)

Figure 5 Distribution of fraudulent amounts (unequal amounts of data)

表5 FedRWA与各联邦学习算法性能对比(不等量数据)

Table 5 Performance comparison between FedRWA and various federated learning algorithms (unequal amounts of data)

算法	G-mean	AUC	Recall	Amount-pred
Center	0.925	0.950	0.857	9 965.4
FedAvg	0.936	<u>0.953</u>	0.878	9 975.4
FedProx	0.936	<u>0.953</u>	0.888	10 747.14
MOON	<u>0.940</u>	0.929	0.888	9 975.4
FedBN	0.936	0.954	0.878	9974.4
FedRN	0.700	0.490	0.793	1229.55
FedRWA-S ₁ -M578	0.937	0.954	0.888	10 747.14
FedRWA-S ₁ -MInf	<u>0.940</u>	0.954	0.888	10 747.14
FedRWA-S ₂ -M578	0.937	0.947	<u>0.908</u>	<u>12 298.75</u>
FedRWA-S ₂ -MInf	0.941	<u>0.953</u>	0.918	12 299.75

注:最优值用粗体标记、次优值用下划线标记。

4 结束语

对实验结果数据进行分析发现,在金融欺诈交易检测这一应用场景中,本文提出的FedRWA算法展现出显著优势。

FedRWA依据数据集的欺诈风险特征模型,精确计算风险权值,得以更加精准地捕捉各机构本地数据集中的欺诈样本特征。从关键指标来看,无论是AUC、Recall,还是G-mean, FedRWA的表现均超越了FedAvg;在新引入的Amount-pred指标上, FedRWA同样优于FedAvg,进一步证实了该算法在欺诈检测场景下的优越性。

本文紧密围绕金融欺诈检测的实际需求,创新性地提出了一种能够有效融合各机构节点数据

集欺诈风险特征的风险权值聚合联邦学习算法。该算法聚焦于各客户端欺诈样本的分布情况, 提高了欺诈样本的检测效果, 成功识别出更多大额欺诈样本, 并且增强了金融数据信息的隐私安全, 同时并未带来复杂的网络架构与通信计算成本。这一研究成果为联邦学习理论在金融欺诈检测以及其他异常检测领域的研究与应用提供了新的思路。

参考文献:

- [1] HILAL W, GADSDEN S A, YAWNEY J. Financial fraud: a review of anomaly detection techniques and recent advances[J]. *Expert Systems with Applications*, 2022, 193: 116429.
- [2] GIANINI G, FOSSI L G, MIO C, et al. Managing a pool of rules for credit card fraud detection by a game theory based approach[J]. *Future Generation Computer Systems*, 2020, 102: 549-561.
- [3] 杨强. AI与数据隐私保护: 联邦学习的破解之道[J]. *信息安全研究*, 2019, 5(11): 961-965.
YANG Q. AI and data privacy protection: the cracking method of federated learning[J]. *Information Security Research*, 2019, 5(11): 961-965.
- [4] MAKKI S, ASSAGHIR Z, TAHER Y, et al. An experimental study with imbalanced classification approaches for credit card fraud detection[J]. *IEEE Access*, 2019, 7: 93010-93022.
- [5] XIAO C, WANG S. An experimental study of class imbalance in federated learning[C]// *Proceedings of 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. Piscataway: IEEE Press, 2021: 1-7.
- [6] MOTIE S, RAAHEMI B. Financial fraud detection using graph neural networks: a systematic review[J]. *Expert Systems with Applications*, 2024, 240: 122156.
- [7] ABDALLAH A, MAAROF M A, ZAINAL A. Fraud detection system: a survey[J]. *Journal of Network and Computer Applications*, 2016, 68: 90-113.
- [8] YANG G, LIU X, LI B. Anti-money laundering supervision by intelligent algorithm[J]. *Computers & Security*, 2023, 132: 103344.
- [9] SAHIN Y, BULKAN S, DUMAN E. A cost-sensitive decision tree approach for fraud detection[J]. *Expert Systems with Applications*, 2013, 40(15): 5916-5923.
- [10] SAHIN Y, DUMAN E. Detecting credit card fraud by ANN and logistic regression[C]// *Proceedings of 2011 International Symposium on Innovations in Intelligent Systems and Applications*. Piscataway: IEEE Press, 2011: 315-319.
- [11] GYAMFI N K, ABDULAI J D. Bank fraud detection using support vector machine[C]// *Proceedings of 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. Piscataway: IEEE Press, 2018: 37-41.
- [12] ALGHOFALI Y, ALBATTAH A, RASSAM M A. A financial fraud detection model based on LSTM deep learning technique[J]. *Journal of Applied Security Research*, 2020, 15(4): 498-516.
- [13] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// *Proceedings of Artificial Intelligence and Statistics*. 2017: 1273-1282.
- [14] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks[J]. *Machine Learning and Systems*, 2020, 2: 429-450.
- [15] 应作斌, 方一晨, 张怡文. 动态聚合权重的隐私保护联邦学习框架[J]. *网络与信息安全学报*, 2022, 8(5): 56-65.
YING Z B, FANG Y C, ZHANG Y W. Privacy-preserving federated learning framework with dynamic weight aggregation[J]. *Chinese Journal of Network and Information Security*, 2022, 8(5): 56-65.
- [16] 余锋, 林庆新, 林晖, 等. 基于生成对抗网络的隐私增强联邦学习方案[J]. *网络与信息安全学报*, 2023, 9(3): 113-122.
YU F, LIN Q X, LIN H, et al. Privacy-enhanced federated learning scheme based on generative adversarial networks [J]. *Chinese Journal of Network and Information Security*, 2023, 9(3): 113-122.
- [17] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: system design[J]. *Machine Learning and Systems*, 2019, 1: 374-388.
- [18] WANG J, LIU Q, LIANG H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 7611-7623.
- [19] LI Z, LIN T, SHANG X, et al. Revisiting weighted aggregation in federated learning with neural networks[C]// *Proceedings of International Conference on Machine Learning*. 2023: 19767-19788.
- [20] DENG Y, LYU F, REN J, et al. Improving federated learning with quality-aware user incentive and auto-weighted model aggregation[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(12): 4515-4529.
- [21] 张连福, 谭作文. 一种基于自适应加权的鲁棒联邦学习算法[J]. *计算机科学*, 2023, 50(z1): 799-807.
ZHANG L F, TAN Z W. Robust federated learning algorithm based on adaptive weighting[J]. *Computer Science*, 2023, 50(z1): 799-807.
- [22] 杨智凯, 刘亚萍, 张硕, 等. 联邦学习通信优化方法综述[J]. *网络与信息安全学报*, 2024, 10(6): 1-23.
YANG Z K, LIU Y P, ZHANG S, et al. Review of communication optimization methods in federated learning[J]. *Chinese Journal of Network and Information Security*, 2024, 10(6): 1-23.
- [23] 刘天瑞, 王连海, 王启正, 等. 算力网络环境下基于区块链的自适应联邦学习[J]. *网络与信息安全学报*, 2024, 10(3): 130-142.
LIU T R, WANG L H, WANG Q Z, et al. Blockchain based adaptive federated learning in computational power network[J]. *Chinese Journal of Network and Information Security*, 2024, 10(3): 130-142.
- [24] YU F, RAWAT A S, MENON A, et al. Federated learning with only positive labels[C]// *Proceedings of International Conference on Machine Learning*. 2020: 10946-10956.
- [25] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. *Foundations and Trends in Machine Learning*, 2021, 14(1-2): 1-210.
- [26] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. *Machine Learning*, 1998, 30(2): 195-215.
- [27] LI Q, HE B, SONG D. Model-contrastive federated learning[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021: 10713-10722.

- [28] LI X, JIANG M, ZHANG X, et al. FedBN: federated learning on non-IID features via local batch normalization[J]. arXiv preprint arXiv:2102.07623, 2021.
- [29] KIM S M, SHIN W, JANG S, et al. FedRN: exploiting k-reliable neighbors towards robust federated learning[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 972-981.

[作者简介]



孙元林 (2000-), 男, 山东东营人, 北京交通大学硕士生, 主要研究方向为异常检测、网络安全。



王楠 (1993-), 女, 山东潍坊人, 北京交通大学副教授, 主要研究方向为网络安全、恶意攻击分析。



张焘 (1996-), 男, 安徽马鞍山人, 北京交通大学副教授, 主要研究方向为网络安全、联邦学习。



刘娟 (1987-), 女, 河北衡水人, 北京交通大学高级工程师, 主要研究方向为金融欺诈检测、工业异常检测、联邦学习。



赵沁馨 (2002-), 女, 江苏扬州人, 主要研究方向为网络安全、恶意攻击分析。