

联邦学习框架下基于 Logistic 回归的金融欺诈检测模型优化研究

谭纪新

中南大学数学与统计学院, 湖南 长沙 410083

摘要

Logistic 回归作为金融风控领域广泛采用的分类算法, 在传统集中式训练模式下面临数据隐私泄露风险与处理效率瓶颈。为应对这些挑战, 该研究构建基于 Logistic 回归的联邦学习框架, 创新性提出集成 SMOTE、SOS 与 HT 的优化模型, 在保持联邦学习隐私保护优势的前提下, 实现模型性能与效率的双重提升。实验表明, 该研究提出的预测模型在不平衡数据处理中展现出更优越的预测性能, 该模型受试者工作特征曲线下面积 (AUC) 值达 0.9971, 较传统联邦 Logistic 模型提升约 7%, 模型改进前后的召回率从 32.25% 提升至 97.57%, 且降低了总通信开销, 实现了识别能力与通信效率的协同优化。

关键词

联邦 Logistic 回归模型; 横向联邦学习; SMOTE; 不平衡数据; 分层最优子抽样方法

中图分类号: TP181

文献标志码: A

doi:10.11959/j.issn.2096-0271.2026043

Research on optimization of financial fraud detection model based on logistic regression under federated learning framework

Tan Jixin

School of Mathematics and Statistics, Central South University, Changsha 410083, China

Abstract

Logistic regression, as a widely adopted classification algorithm in the financial risk control field, faces the risks of data privacy leakage and bottlenecks in processing efficiency under the traditional centralized training mode. To address these challenges, this study constructs a federated learning framework based on logistic regression, innovatively proposes an optimized model that integrates SMOTE, SOS, and HT. While maintaining the privacy protection advantages of federated learning, the framework achieves dual improvements in model performance and efficiency. Experimental results show that the proposed predictive model exhibits superior predictive performance in handling imbalanced data: the model's AUC value reaches 0.9971, which is 7% higher than that of the traditional federated

logistic model; the recall rate increases from 32.25% to 97.57%. Additionally, the model reduces the total communication overhead, realizing the coordinated optimization of recognition capability and communication efficiency.

Key words

federated logistic regression model, horizontal federated learning, SMOTE, imbalanced data, stratified optimal subsampling method

1 引言

在数字化时代，数据已经成为推动各行各业创新与发展的核心要素^[1]。随着数据隐私问题日益凸显及跨组织数据合作需求不断增长，Google提出了联邦学习这一开创性理念^[2]，该领域引发了众多学者的深入探索与拓展。Yang等^[3]学者进一步将其定义为涵盖所有致力于保护隐私的去中心化协作机器学习技术，明确了联邦学习的核心特征与本质内涵。但现有联邦学习框架在金融欺诈检测场景中仍存在对极端不平衡数据处理能力不足和性能与效率的协同优化缺失两大未解决的关键挑战。

针对不平衡数据的Logistic回归建模始终是一个重要课题，已有诸多相关研究。例如，合成少数类过采样技术（synthetic minority oversampling technique, SMOTE）为解决数据分布不平衡问题提供了有效途径^[4]。然而，关于分布式环境下不平衡数据处理的研究仍处于初步探索阶段^[5-7]。为提升大数据场景中Logistic回归的联邦学习效率，学者分别从不同角度提出了解决方案，包括采用打包编码技术^[8]、分布式交替方向乘子法算法（Distributed Alternating Direction Method of Multipliers）^[9-10]、模型融合技术以及自适应参数调整机制^[11]等。

本研究针对联邦学习场景下极端不平

衡数据与训练效率的双重问题，设计客户端本地数据优化流程：先通过SMOTE合成少数类样本，解决数据分布不平衡问题；再基于 L_2 范数距离筛选多数类代表性样本，在缩减数据规模的同时保留关键特征。该设计突破传统“过采样仅平衡数据、抽样仅提升效率”的单一功能局限，实现“数据平衡-效率提升-特征保留”的协同优化，为联邦学习框架下不平衡数据建模提供新的技术范式。

2 预备知识

2.1 Logistic模型

Logistic模型作为经典的二分类概率模型，基于线性回归基础构建，通过逻辑函数将线性组合 $\mathbf{w}^T \mathbf{x} + b$ 的输出映射到0到1之间的概率值，其数学表达式为：

$$P(Y=1|\mathbf{x}; \mathbf{w}, b) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \quad (1)$$

其中， $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ 为输入特征向量， $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$ 为对应特征的权重向量， b 为偏置项， $y \in \{0, 1\}$ 表示样本类别标签。

2.2 横向联邦学习

横向联邦学习是一种去中心化的分布

式建模方法。在一个横向联邦学习系统中，假设存在 K 个数据结构相同的参与者，各参与者拥有的数据集为 D_k ， $k \in V = \{1, 2, \dots, K\}$ 。横向联邦学习过程主要包括以下 5 个步骤。

步骤 1：服务器初始化一个全局 Logistic 模型 M' ，服务器将全局模型 M' 发送给所有参与联邦学习的 K 个参与者。

步骤 2：各参与者接收到全局模型 M' 后，使用本地数据集 D_k 训练模型 M'_k ，将 M'_k 的参数发送给服务器。

步骤 3：服务器安全聚合来自各参与者的模型 M'_k ，根据各参与者数据集大小 $|D_k|$ 进行加权聚合，得到一个新的全局模型 M^{t+1} 。模型参数聚合方式如下：

$$b^{t+1} = \frac{\sum_{k=1}^K |D_k| \cdot b'_k}{\sum_{k=1}^K |D_k|}, \theta^{t+1} = \frac{\sum_{k=1}^K |D_k| \cdot \theta'_k}{\sum_{k=1}^K |D_k|} \quad (2)$$

其中， b'_k 表示第 k 个参与者在第 t 轮本地模型的偏置项， θ'_k 表示第 k 个参与者在第 t 轮本地模型的权重参数向量。

步骤 4：服务器将聚合后的新全局模型 M^{t+1} 重新发送给所有参与者。

步骤 5：参与者更新各自的本地模型。

通过上述迭代过程，直至损失函数收敛或达到预先设定的迭代轮次要求，最终完成整个训练过程。

在金融欺诈检测中，本研究选用横向联邦学习的 FedAVG（联邦平均）方法，其在处理不平衡、异质性数据时具备显著优势：FedAVG 通过客户端多轮本地训练，反复拟合少数类样本，避免 FedSGD（联邦随机梯度下降）单次梯度更新导致的特征学习不足问题；FedAVG 仅传模型参数，降低敏感数据泄露风险，且支持部分节点聚合，机构故障不影响全局训练；FedSGD 因单次更新、高频通信，易受多数类主导、开销高、隐私风

险大，而 FedAVG 通过“本地多轮+周期聚合”，在性能、效率与隐私间实现更优平衡。

2.3 SMOTE

过采样技术是处理不平衡数据集的常用技术，通过增加少数类样本的数量，使数据集的类别分布更加均衡。SMOTE 通过插值的方式合成新的少数类样本^[12]。对于每个少数类样本，SMOTE 会在其近邻中随机选择一个近邻，然后在该样本与其近邻之间的连线上随机选取一点作为新的合成样本。这样能有效增加少数类样本的多样性，减少过拟合的风险^[13]。

2.4 分层最优子抽样

分层最优子抽样（stratified optimal subsampling, SOS）是一种从原始数据集中选择最优子集进行模型训练的抽样方法。具体步骤如下。

步骤 1：类别分层。设原始数据集为 $D' = (X', y')$ ，其中 X' 为特征矩阵， y' 为标签向量，将 X' 按 y' 划分为两个样本集：多数类样本集 X_{maj} 和少数类样本集 X_{min} ；

步骤 2：信息价值量化。通过 L_2 范数距离矩阵衡量多数类样本的信息价值，计算 X_{maj} 中每个样本 x_i 与 X_{min} 中所有样本 x_j 的距离 $D_{ij} = \|x_i - x_j\|_2$ ， $\|\cdot\|_2$ 为 L_2 范数，并计算 X_{maj} 中每个样本的平均距离 $d_i = \frac{1}{|X_{\text{min}}|} \sum_{j=1}^{|X_{\text{min}}|} D_{ij}$ ；

步骤 3：最优子集筛选。设定缩减比例 α ，选择 X_{maj} 中 d_i 最小的 $\alpha \cdot |X_{\text{maj}}|$ 个样本与 X_{min} 中所有样本组合，得到分层最优子抽样最终获得的数据集。

3 本研究提出的联邦 Logistic 模型

3.1 模型选择

金融风控领域要求模型决策具备透明性，监管机构与业务人员需通过特征权重直接理解欺诈预测逻辑。相较于 LightGBM、XGBoost 等树模型，Logistic 回归的线性系数可直接量化单个特征对欺诈概率的边际影响，且复杂模型的参数规模较大，在跨机构低带宽场景中易导致通信瓶颈。因此本研究选择对联邦 Logistic 回归模型进行进一步优化。

3.2 符号说明

伪代码中符号及含义见表1。

3.3 联邦 Logistic 模型+SMOTE

金融数据集具有类别高度不平衡的特点，直接使用其建模会导致少数类样本的特征难以被有效学习和识别。因此本研究在各客户端采用 SMOTE 合成少数类样本平衡数据分布，联合构建联邦 Logistic 模型。合成过程在客户端本地完成，各客户端仅向服务器传输本地训练后的模型参数，合成样本不参与跨机构传输，不会发生隐私泄露。

算法 1 联邦 Logistic 模型+SMOTE

- 1: 输入:
- 2: K 个客户端持有数据 $\{D_k\}_{k=1}^K$ 其中 $D_k = (X_k, y_k)$
- 3: 训练轮数 T
- 4: 学习率 $\eta > 0$
- 5: SMOTE 算法参数
- 6: 输出:
- 7: 全局逻辑回归模型 $M_G = (\omega^*, b^*)$

表1 伪代码中符号及含义

符号	核心含义
$D_k = (X_k, y_k)$	第 k 个客户端的本地数据集,其中 X_k 为特征矩阵, y_k 为标签向量
V	参与本轮训练的有效客户端索引集合, $V = \{1, \dots, K\}$
ω_k	第 k 个客户端本地更新后的权重向量
b_k	第 k 个客户端本地更新后的偏置项
N	本轮参与训练的所有客户端数据总量
n_k	第 k 个客户端的本地数据量
k_{smote}	SMOTE 中,合成少数类样本时的近邻数
α	分层最优子抽样的样本缩减比例
k_{sub}	分层最优子抽样中,计算样本距离时的近邻数
c	类别标签的通用占位符,遍历所有可能的类别, $c \in \{0, 1\}$
c_{maj}	数据集中样本数量最多的类别标签
c_{min}	数据集中样本数量最少的类别标签
Θ	超参数调优(hyperparameter tuning, HT)的参数集合
X', y'	SMOTE 平衡后的客户端特征矩阵(X')与标签向量(y')
X^*, y^*	分层最优子抽样筛选后的客户端特征矩阵(X^*)与标签向量(y^*)
D_{ij}	分层最优子抽样中,多数类样本 x_i 与少数类样本 x_j 的距离,基于 L_2 范数计算
∇L	Logistic 回归模型的损失函数梯度,损失函数为交叉熵损失

8: **服务端执行:**

9: 初始化 $\omega \leftarrow 0^d, b \leftarrow 0$

10: **对于轮次 $t = 1$ 到 T 执行:**

11: 将当前全局模型 (ω, b) 传输至所有客户端

12: **客户端并行更新:**

13: **对每个客户端 $k = 1$ 到 K 并行执行:**

14: $(\omega_k, b_k) \leftarrow$ 客户端更新函数 (k, ω, b)

15: 将客户端更新后的模型 (ω_k, b_k) 和本地数据量 n_k 发送到服务端

16: **结束客户端循环**

17: **参数聚合更新:**

18: $V \leftarrow \{1, \dots, K\}$

19: $N \leftarrow \sum_{k \in V} n_k$

20: $\omega \leftarrow \sum_{k \in V} n_k \omega_k / N$

21: $b \leftarrow \sum_{k \in V} n_k b_k / N$

22: **结束训练轮次**

23: **返回全局模型 $M_G \leftarrow (\omega, b)$**

24: **客户端更新函数 (k, ω, b)**

25: $(X_k, y_k) \leftarrow D_k$

26: $(X'_k, y'_k) \leftarrow \text{SMOTE}(X_k, y_k)$

27: $(\omega_k, b_k) \leftarrow (\omega, b)$

28: **对本地训练轮次执行:**

29: $\nabla L = \nabla_{\omega, b} l(\omega_k, b_k; X'_k, y'_k)$

30: $\omega_k \leftarrow \omega_k - \eta \nabla_{\omega} L$

31: $b_k \leftarrow b_k - \eta \nabla_b L$

32: **结束本地训练轮次**

33: **返回客户端更新后的模型参数 (ω_k, b_k)**

34: **结束函数**

3.4 联邦 Logistic 模型 +SMOTE+SOS

为解决用 SMOTE 过采样后数据量增加导致的训练耗时问题, 本研究引入分层

最优子抽样方法对模型进行针对性改进。相较于传统随机抽样, 该方法能在缩减数据规模的同时, 保留关键特征与分布特性, 在提升训练效率的同时避免信息损失导致的模型性能大幅下降。

算法 2 联邦 Logistic 模型 +SMOTE+SOS

1: **输入:**

2: K 个客户端持有数据 $\{D_k\}_{k=1}^K$, 训练轮数 T

3: SMOTE 参数 $k_{\text{smote}}, \alpha, k_{\text{sub}}$

4: **输出:**

5: 全局模型 (ω^*, b^*)

6: **服务端执行:**

7: $(\omega^*, b^*) \leftarrow$ 联邦训练 $(\{(X_k, y_k)\}, \theta^*)$

8: **客户端更新函数 (D_k, ω, b)**

9: $(X_k, y_k) \leftarrow D_k$

10: **执行 SMOTE 过采样:**

11: $(X'_k, y'_k) \leftarrow \text{SMOTE}(X_k, y_k, k_{\text{smote}})$

12: **执行分层最优子抽样:**

13: $(X^*, y^*) \leftarrow$ 分层最优子抽样 $(X', y', \alpha, k_{\text{sub}})$

14: **本地模型训练:**

15: **更新本地模型参数 (ω, b)**

16: **返回更新后的本地模型参数 (ω_k, b_k)**

17: **结束函数**

18: **分层最优子抽样函数 (X, y, α, k)**

19: 确定多数类标签 $c_{\text{maj}} \leftarrow \arg \max_c |\{y_i = c\}|$, 少数类标签 $c_{\text{min}} \leftarrow \arg \min_c |\{y_i = c\}|$

20: $X_{\text{maj}} \leftarrow \{x_i | y_i = c_{\text{maj}}\}$

21: $X_{\text{min}} \leftarrow \{x_i | y_i \neq c_{\text{min}}\}$

22: 计算距离矩阵 $D \in R^{|X_{\text{maj}}| \times |X_{\text{min}}|}$

23: $D_{ij} \leftarrow \|x_i - x_j\|^2, x_i \in X_{\text{maj}}, x_j \in X_{\text{min}}$

#计算 L_2 范数

24: 在 X_{maj} 中选取行和最小的 $\alpha \cdot$

$|X_{\text{maj}}|$ 个样本

25: 返回抽样后的数据集 (X^*, y^*)

26: 结束函数

3.5 联邦 Logistic 模型 + SMOTE + SOS + HT

由于 Logistic 模型性能很大程度上依赖于超参数设置, 本研究引入超参数调优方法对模型进行改进。采用随机搜索结合并行评估的超参数调优策略, 从预设参数空间中随机选取参数组合进行评估。由于 Fraud 数据集欺诈样本很少, 准确率易被多数类主导, 而 AUC 值能客观反映模型对欺诈交易的区分能力, 因此本研究选择 AUC 表现最优的参数组合进行模型训练。

算法 3 联邦 Logistic 模型 + SMOTE + SOS + HT

1: 输入:

2: K 个客户端持有数据 $\{D_k\}_{k=1}^K$ 其中 $D_k = (X_k, y_k)$

3: SMOTE 近邻数 k_{smote} , 子抽样缩减比例 α

4: 超参数集合 Θ

5: 输出:

6: 全局逻辑回归模型 $M_G = (\omega^*, b^*)$

7: 主过程 M_{AIN}

8: $\theta^* \leftarrow$ 超参数调优 $(\{(X_k, y_k)\}, \Theta)$

9: $(\omega^*, b^*) \leftarrow$ 联邦训练 $(\{(X_k, y_k)\}, \theta^*)$

10: 结束主过程

11: 超参数调优函数 $(\{(X_k, y_k)\}, \Theta)$

12: 对于每个 $\theta \in \Theta$ 执行:

13: 初始化平均 AUC $\text{avg_auc} \leftarrow 0$

14: 对每个客户端 k 执行:

15: $(X'_k, y'_k) \leftarrow$ SMOTE 过采样

$(X_k, y_k, k_{\text{smote}})$

16: $(X_k^*, y_k^*) \leftarrow$ 分层最优子抽样

(X'_k, y'_k, α)

17: $(\omega_k, b_k) \leftarrow$ 训练本地模型

(X_k^*, y_k^*, θ)

18: $\text{auc}_k \leftarrow$ 计算本地模型的 AUC

$(\omega_k, b_k, X_k, y_k)$

19: 按客户端数量计算平均

$\text{avg_auc} \leftarrow \text{avg_auc} + \text{auc}_k / K$

20: 结束客户端循环

21: 结束超参数循环

22: 返回平均 AUC 最大的最优超参数组合 $\arg \max_{\theta} \text{avg_auc}$

23: 结束函数

4 实验评估

本研究采用的数据集源自 Kaggle 平台的“Financial Datasets For Fraud”, 为简便起见, 将其命名为 Fraud 数据集。该数据集包含 11 个特征和 6 362 620 个样本, 其中欺诈样本仅占 0.13%。根据金融风控行业报告, 全球范围内金融交易欺诈率普遍低于 0.5%, 本研究选取的数据集的极端不平衡分布符合金融欺诈检测场景的真实特征。

4.1 数据预处理

4.1.1 分类特征编码

本研究针对分类特征“type”(交易类型)采用独热编码(One-Hot Encoding)进行处理, 将 6 类交易类型转换为 5 个二进制编码。编码过程在各客户端本地完成。

4.1.2 数据标准化

本研究采用 Z-Score 标准化方法对所

有数值特征进行处理，计算式为：

$$x_{scaled} = \frac{x-\mu}{\sigma} \tag{3}$$

其中， μ 为特征均值， σ 为特征标准差。

4.2 对比模型介绍

本研究针对金融欺诈检测场景，构建联邦 Logistic 回归模型，并与其他集中式及联邦学习模型进行对比，下面对有关对比模型进行介绍。

4.2.1 联邦单神经元神经网络模型

本研究建立的联邦单神经元神经网络（SNNN）模型仅包含一个输出层，无隐藏层，采用 Tanh 激活函数。参数设置如下：模型采用 Adam 优化器，学习率是 0.001，训练轮次为 10 轮。

4.2.2 联邦多层感知器模型

本研究建立的联邦多层感知器（MLP）模型包含两个隐藏层和一个输出层。两个隐藏层分别包含 64 个神经元和 32 个神经元，均以 ReLU 为激活函数，输出层只有一个神经元，激活函数是 Sigmoid。参数设置如下：模型采用 Adam 优化器，学习率是 0.001，训练轮次为 10 轮。

4.2.3 联邦深度信念网络模型

本研究建立的联邦深度信念网络（DBN）模型包含两个受限玻尔兹曼机层（RBM）和一个输出层，均使用 Sigmoid 激活函数。两个 RBM 层分别包含 128 个神经元和 64 个神经元，RBM 层通过对比散度算法进行无监督预训练，使用 Gibbs 采样近似梯度。训练过程中参数设置如下：

模型采用 Adam 优化器，学习率是 0.001，训练轮次为 10 轮。

4.2.4 联邦长短期记忆网络模型

本研究所建立的联邦长短期记忆网络（LSTM）模型包含一个 LSTM 层和一个全连接层，LSTM 层包含 16 个神经元，用于学习输入序列中的长期依赖关系，使用 Tanh 和 Sigmoid 联合学习时序特征。全连接层包含 1 个神经元，使用 Sigmoid 激活函数，输出样本属于正类的概率。模型参数设置如下：模型采用 Adam 优化器，学习率是 0.001，训练轮次为 10 轮。

4.2.5 联邦随机森林模型

本研究建立的联邦随机森林（RF）模型包含 100 棵决策树，训练轮次为 10 轮，最终采用概率平均策略聚合预测结果。

4.2.6 联邦高斯混合模型

本研究构建的联邦高斯混合模型（GMM）包含 3 个高斯分布成分，训练轮次为 10 轮。

4.2.7 联邦风险权重聚合模型

本研究构建的联邦风险权重聚合模型（RWAM）各客户端分别训练包含 50 棵决策树的本地模型，根据归一化风险权重对各客户端的数据进行分层抽样，用聚合数据训练包含 50 棵决策树的全局模型，训练轮次为 10 轮。

4.2.8 联邦 Logistic 模型

本研究构建了联邦 Logistic 欺诈检测模型，模型结果见表 2。

表2 联邦Logistic回归模型系数

变量	step	type	amount	OldOrg	newOrig	oldDest	newDest
系数	0.36	-11.30	152.10	-178.33	35.00	-37.99	3.73

联邦 Logistic 模型的可解释性更强，能够清晰解读特征的贡献程度，更适用于实际应用场景。在几个变量中，交易金额（amount）的系数高达 152.10，交易金额与欺诈风险呈强正相关，交易金额越高，越容易被判别存在欺诈行为；汇款前账户原有余额的系数为-178.33，与欺诈行为发生呈强负相关，账户原有余额越高，用户发生欺诈行为的可能性越小。这种高可解释性使该模型适用于需要高信任度的应用场景，确保模型的预测结果能够更容易被相关人员理解和接受。

4.2.9 模型对比

通过对比各联邦学习模型和经典非联邦学习模型的运行时间、准确率、AUC 值

等评价指标和通信开销情况，突出联邦 Logistic 模型相较于其他模型所表现出的优越性能，模型评价指标对比及联邦学习模型通信开销对比结果见表3和表4。其中通信开销计算方法如下：每轮联邦学习训练的通信开销=服务器到客户端的下行通信量+客户端到服务器的上行通信量，多轮训练后累加得到总通信开销。

为更直观对比不同联邦学习模型的分类区分能力，图 1 展示了各模型的 ROC（受试者工作特征曲线）图。

在本研究所构建的几种联邦学习模型和集中式模型中，联邦 Logistic 回归模型展现出显著优势。模型效率较高，准确率和精确率与联邦长短期记忆网络模型、联邦随机森林模型、联邦风险权重聚合模型

表3 模型评价指标对比(五折交叉验证)

模型	运行时间/s	准确率	精确率	F1分数	AUC 值
联邦 Logistic 回归模型	262.35	99.91%	98.34%	0.485 7	0.931 8
联邦单神经元模型	2 237.25	87.00%	10.80%	0.104 3	0.578 4
联邦多层感知器模型	4 133.81	99.92%	89.14%	0.532 2	0.809 3
联邦深度信念网络模型	1 493.75	97.87%	0.08%	0.001 9	0.498 7
联邦长短期记忆网络模型	46 310.53	99.94%	94.59%	0.682 1	0.988 4
联邦随机森林模型	13 752.12	99.95%	97.72%	0.755 1	0.995 5
联邦高斯混合模型	1 198.42	77.53%	0.32%	0.000 6	0.584 4
联邦风险权重聚合模型	12 375.76	99.94%	95.80%	0.794 4	0.969 1
集中式 Logistic 回归模型	254.67	96.42%	3.07%	0.059 4	0.973 6
集中式多层感知器模型	3 264.36	99.93%	95.61%	0.631 3	0.960 6
集中式随机森林模型	6 039.76	97.82%	5.46%	0.103 4	0.992 3
集中式长短期记忆网络模型	25 539.36	99.88%	32.23%	0.194 0	0.943 0

表4 联邦学习模型通信开销对比

模型	平均每轮参数传输量/KB	通信轮次/次	总通信开销/KB
联邦Logistic回归模型	0.37	10	3.73
联邦单神经元神经网络模型	0.19	10	1.89
联邦多层感知器模型	62.11	10	621.21
联邦深度信念网络模型	215.23	10	2 152.32
联邦长短期记忆网络模型	26.82	10	268.24
联邦随机森林模型	48 856.39	10	488 563.87
联邦高斯混合模型	8.13	10	81.31
联邦风险权重聚合模型	26 493.15	10	264 931.53

和集中式多层感知器模型接近，对正样本的预测可靠性强。而且其F1分数与AUC值均显著高于联邦深度信念网络模型、联邦高斯混合模型等，实现了精确率和召回率的良好平衡。且在通信开销方面，联邦Logistic回归模型的总通信开销仅为3.73KB，远低于联邦多层感知器模型和联邦高斯混合模型，其在数据传输过程中对网络资源的占用更少，更适用于网络环境受限的场景。

4.3 联邦学习不平衡数据的主流处理模型

本研究聚焦联邦学习领域的不平衡数据问题，梳理了联邦自适应合成采样(ADASYN)模型、联邦焦点损失模型、联邦SMOTE-Boost模型、联邦元不平衡模型4种主流处理模型，并与本研究构建的联邦Logistic回归模型进行对比，下面对有关模型进行介绍。

4.3.1 联邦ADASYN模型

联邦ADASYN模型通过自适应过采样技术为各客户端的少数类样本生成合成数

据，且对难以学习的少数类样本分配更高的合成权重。本研究构建的联邦ADASYN模型设置K近邻数为5，自适应采样比例系数为0.8，训练轮次为10轮。

4.3.2 联邦焦点损失模型

联邦焦点损失模型通过降低多数类样本的损失权重，改善少数类样本的错分情况。各客户端采用Focal Loss（焦点损失）替代传统交叉熵损失，全局模型通过加权聚合客户端参数实现协同优化。本研究构建的联邦焦点损失模型设置聚焦参数 $\gamma=2$ ，平衡因子 $\alpha=0.25$ ，训练轮次为10轮。

4.3.3 联邦SMOTE-Boost模型

联邦SMOTE-Boost模型融合了SMOTE与Boosting集成学习思想，在联邦学习场景下实现不平衡数据的分层优化。本研究构建的联邦SMOTE-Boost模型设置SMOTE近邻数为5，弱分类器数量为10，训练轮次为10轮。

4.3.4 联邦元不平衡模型

联邦元不平衡模型通过元知识迁移指

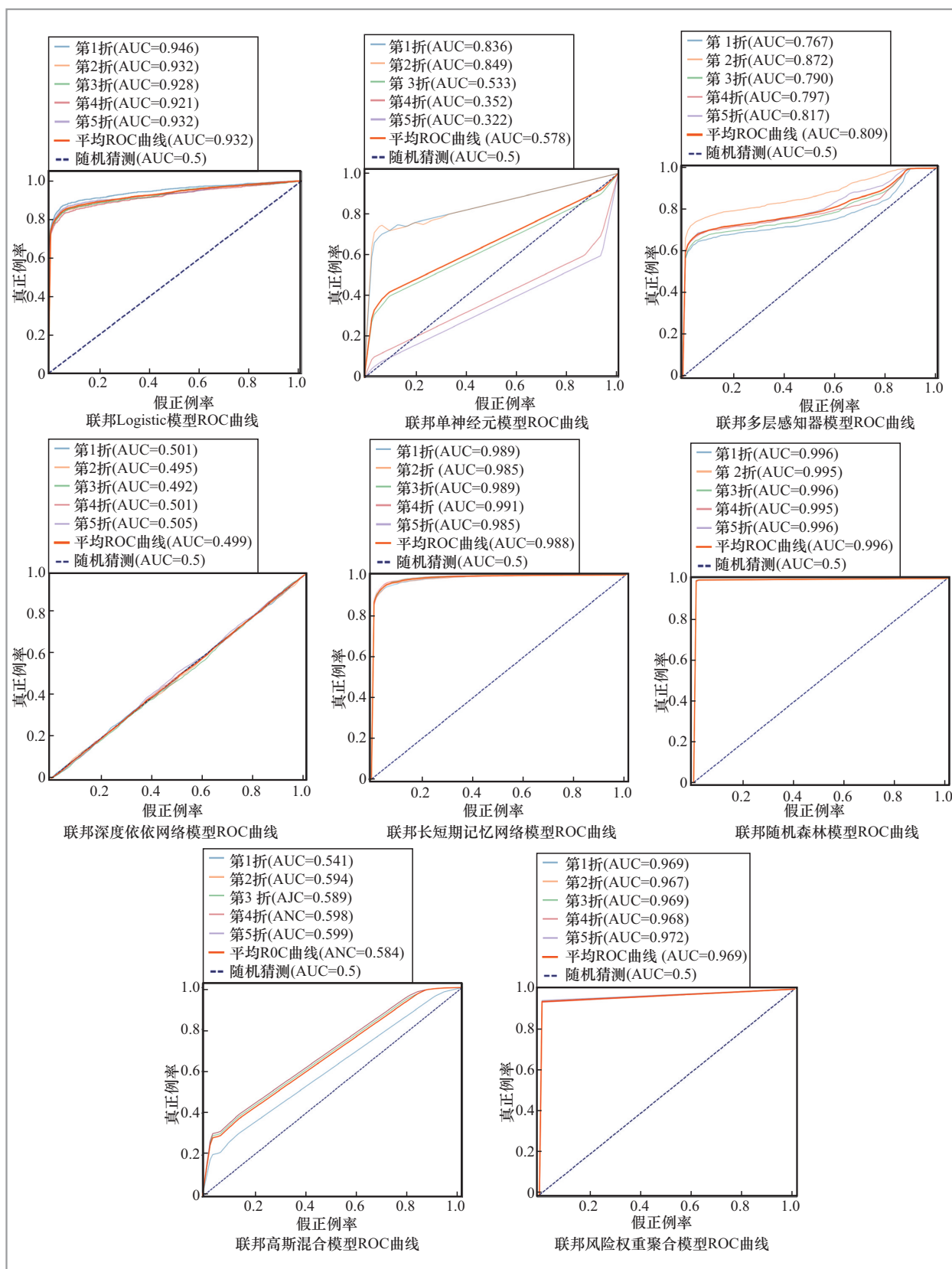


图1 模型ROC曲线对比

导各客户端的本地训练，提升模型对跨客户端少数类特征的泛化能力。该模型在服务器端维护元参数，客户端通过元参数初始化本地模型，结合本地少数类样本微调后将参数反馈至服务器聚合。本研究构建的联邦元不平衡模型设置元学习支持集比例为 20%，查询集比例为 20%，训练轮次为 10 轮。

4.3.5 模型结果对比

本研究针对 Fraud 数据集构建了目前联邦学习领域不平衡数据的主流处理模型，通过对比突出本研究所构建的联邦 Logistic 回归模型相较于其他模型在处理不平衡数据方面所表现出的优越性能，模型评价指标对比及模型通信开销对比结果见表 5 和表 6。

本研究构建的联邦 Logistic 回归模型，在处理不平衡数据集时表现出显著的性能

优势。该模型的准确率达 97.41%，对正样本的预测可靠性强，且在召回率和 AUC 值等评价指标上远超其他主流模型，在捕捉少数类欺诈样本上的优势显著。该模型的训练效率较高，且其通信开销均为所有对比模型中最低，在联邦学习场景下的通信效率优势极为突出。

4.4 联邦 Logistic 改进模型的实验结果

本研究从多个方面对联邦 Logistic 回归模型（FL+LR）进行改进，改进后的模型在处理类别不平衡的大规模数据集时表现出更优越的性能，实验对比结果见表 7 和表 8。

首先针对数据集类别分布极不平衡的问题，本研究引入了 SMOTE 合成少数类样本，平衡各客户端的数据分布，改进后的模型召回率及 AUC 指标都有所提升，能

表 5 模型评价指标对比(5折交叉验证)

模型	运行时间/s	准确率	召回率	AUC 值
联邦 ADASYN 模型	271.23	66.21%	24.82%	0.802 6
联邦焦点损失模型	2 793.25	99.87%	19.68%	0.769 3
联邦 SMOTE–Boost 模型	9 145.64	59.13%	80.70%	0.809 5
联邦元不平衡模型	1 356.32	74.93%	55.51%	0.683 6
FL+LR+SMOTE+SOS+HT	285.38	97.41%	97.57%	0.997 1

表 6 模型通信开销对比

模型	平均每轮参数传输量/KB	通信轮次/次	总通信开销/KB
联邦 ADASYN 模型	0.37	10	3.69
联邦焦点损失模型	0.41	10	4.14
联邦 SMOTE–Boost 模型	1.33	10	13.32
联邦元不平衡模型	1.05	10	10.52
FL+LR+SMOTE+SOS+HT	0.35	6	2.12

表7 改进前后的联邦Logistic回归模型评价指标对比(五折交叉验证)

模型	运行时间/s	准确率	召回率	AUC 值
FL+LR	262.35	99.91%	32.25%	0.931 8
FL+LR+SMOTE	478.23	97.26%	90.36%	0.984 6
FL+LR+SMOTE+SOS	335.37	83.31%	97.13%	0.978 0
FL+LR+SMOTE+SOS+HT	285.38	97.41%	97.57%	0.997 1

表8 模型通信开销对比

模型	平均每轮参数传输量	通信轮次/次	总通信开销
FL+LR	0.37	10	3.73
FL+LR+SMOTE	0.40	10	4.02
FL+LR+SMOTE+SOS	0.44	10	4.41
FL+LR+SMOTE+SOS+HT	0.35	6	2.12

表9 Wilcoxon符号秩检验结果

指标	(FL+LR) 均值±标准差	(FL+LR+SMOTE+SOS+ HT)均值±标准差	差异值	Wilcoxon符号秩检验 p值
AUC 值	0.931 8±0.0080	0.997 1±0.0018	0.065 3	0.012
召回率	0.322 5±0.0099	0.975 7±0.0297	0.653 2	0.008
总通信开销	3.73±0.21	2.12±0.17	-1.51	0.015

有效应对数据不平衡问题，改进后的模型相较于传统联邦 Logistic 回归模型相比召回率提升了 65.32%，在金融欺诈检测这类样本比例悬殊的数据场景中能够更有效地识别欺诈行为，减少漏判情况，有助于增强客户信任。

为解决 SMOTE 过采样后数据量增加导致的训练耗时问题，本研究进一步引入分层最优子抽样方法对各客户端经过 SMOTE 平衡后的数据进行优化处理，筛选出部分具有代表性的子样本进行模型训练，各客户端处理的数据规模显著降低，提升了训练效率。

最后，本研究对模型的超参数进行调优，提升了模型的准确率及 AUC 值。为了验证改进后的模型结果的显著性，本研究进行了 Wilcoxon 符号秩检验，以 95% 置信水平 ($\alpha=0.05$) 判断指标提升是否具有统计学意义。检验结果见表 9。

基于 Wilcoxon 符号秩检验结果，(FL+LR+SMOTE+SOS+HT) 改进模型较基准模型 (FL+LR) 实现 AUC 值、召回率、总通信开销三维指标的全面显著优化，改进后的模型能在类别分布极不平衡的数据环境下更有效地识别出潜在的欺诈行为。

5 结论

本研究提出的模型在金融风控、医疗诊断等对少数类识别敏感的领域具有重要实用价值：一方面，通过联邦学习框架保护数据隐私，为跨机构数据协作提供技术支撑^[14-17]；另一方面，不平衡数据处理策略提升了对欺诈行为等稀有事件的识别能力，减少漏判带来的经济损失。

未来可引入生成对抗网络（GAN）替代传统 SMOTE，通过学习真实欺诈样本的潜在分布生成高可信度合成数据；并结合注意力机制改进分层子抽样，聚焦对分类贡献度高的特征子集，降低关键信息丢失风险。可进一步探索联邦 Logistic 回归模型与轻量化深度学习模型的级联架构，利用深度学习模型提取交易数据的时序特征与空间特征，再通过 Logistic 回归进行概率化决策，提升对复杂欺诈模式的识别能力。

参考文献：

[1] 朱建明, 张沁楠, 高胜, 等. 基于区块链的隐私保护可信联邦学习模型[J]. 计算机学报, 2021, 44(12): 2464-2484.

Zhu J M, Zhang Q N, Gao S, et al. Privacy preserving and trustworthy federated learning model based on blockchain[J]. Chinese Journal of Computers, 2021, 44(12): 2464-2484.

[2] Konečný J, McMahan B, Ramage D. Federated optimization: distributed optimization beyond the datacenter[PP]. arXiv preprint, 2015, arXiv: 1511.03575.

[3] Yang Q, Liu Y, Chen T J, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[4] 王丹阳, 梁伟红, 李玉萍, 等. 面向不平衡数据的木薯叶部病害图像识别方法[J]. 中国农机化学报, 2025, 46(3): 101-107, 123.

Wang D Y, Liang W H, Li Y P, et al. Cassava leaf disease image recognition method for imbalanced data[J]. Journal of Chinese Agricultural Mechanization, 2025, 46(3): 101-107, 123.

[5] 王璐, 程晓荣. 改进不平衡数据中的集成学习算法[J]. 计算机与数字工程, 2025, 53(1): 26-30.

Wang L, Cheng X R. Improve ensemble learning algorithms in unbalanced data[J]. Computer and Digital Engineering, 2025, 53(1): 26-30.

[6] Yang R, Jia S, Wang H L, et al. Dilated dynamic supervised contrastive learning framework for fault diagnosis under imbalanced dataset conditions[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2025, 239(7): 2626-2636.

[7] Salah Fathi K, Barakat S, Rezk A. An effective SQL injection detection model using LSTM for imbalanced datasets[J]. Computers & Security, 2025, 153: 104391.

[8] 汤殿华, 曹云飞, 黄云帆, 等. 基于CKKS同态加密的纵向联邦逻辑回归训练方案[J]. 通信技术, 2024, 57(11): 1173-1180.

Tang D H, Cao Y F, Huang Y F, et al. A vertical federated logistic regression

training scheme based on CKKS homomorphic encryption[J]. Communications Technology, 2024, 57(11): 1173–1180.

[9] 雷大江, 杜萌, 李智星, 等. 稀疏多元逻辑回归问题优化算法研究[J]. 重庆邮电大学学报(自然科学版), 2019, 31(3): 354–366.

Lei D J, Du M, Li Z X, et al. Research on optimization algorithm for sparse multinomial logistic regression problem[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31(3): 354–366.

[10] Ding W Y, Ng M K, Zhang W X. A generalized alternating direction implicit method for consensus optimization: application to distributed sparse logistic regression[J]. Journal of Global Optimization, 2024, 90(3): 727–753.

[11] 王俊豪. 联邦学习场景下的贡献评估和模型优化[D]. 中国科学技术大学, 2024.

Wang J H. Contribution Evaluation and Model Optimization in Federated Learning Scenarios[D]. University of Science and Technology of China, 2024.

[12] 王晓霞, 李雷孝, 林浩. SMOTE类算法研究综述[J]. 计算机科学与探索, 2024, 18(5): 1135–1159.

Wang X X, Li L X, Lin H. Survey of research on SMOTE type algorithms[J]. Journal of Frontiers of Computer Science & Technology, 2024, 18(5): 1135–1159.

[13] 杜睿山, 黄玉朋, 付晓飞, 等. 基于SMOTE和XGBoost的天然水合物与天然气储层识别[J]. 特种油气藏, 2024, 31(5): 11–19.

Du R S, Huang Y P, Fu X F, et al. Identification of natural gas hydrates and natural gas reservoirs based on SMOTE and XGBoost[J]. Special Oil & Gas Reservoirs, 2024, 31(5): 11–19.

[14] 王健宗, 孔令炜, 黄章成, 等. 联邦学习隐私保护研究进展[J]. 大数据, 2021, 7(3): 130–149.

Wang J Z, Kong L W, Huang Z C, et al. Research advances on privacy protection of federated learning[J]. Big Data Research, 2021, 7(3): 130–149.

[15] 张纪林, 顾小卫, 张亦钊, 等. 跨域数据授权运营研究及应用[J]. 大数据, 2023, 9(4): 83–97.

Zhang J L, Gu X W, Zhang Y Z, et al. Research and application of cross-domain data authorization and operation[J]. Big Data Research, 2023, 9(4): 83–97.

Qi X, Yang H B, Meng F, et al. Distributed photovoltaic short-term power prediction method based on horizontal federated learning[J]. Acta Energiae Solaris Sinica, 2025, 46(1): 686–695.

[16] 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.

Wu W T, Wu Y L, Lin W W, et al. Horizontal federated learning: research status, system applications and open challenges[J]. Chinese Journal of Computers, 2025, 48(1): 35–67.

[17] 孙爽, 李晓会, 刘妍, 等. 不同场景的联邦学习安全与隐私保护研究综述[J]. 计算机应用研究, 2021, 38(12): 3527–3534.

Sun S, Li X H, Liu Y, et al. Survey on security and privacy protection in different scenarios of federated learning[J]. Application Research of Computers, 2021, 38(12): 3527–3534.

作者简介



谭纪新（2003-），女，中南大学硕士生，主要研究方向为分布式学习。



王洪（1977-），男，副教授，博士后，主要研究方向为分布式学习、高维数据分析、大数据分析、生存分析。

收稿日期: XXXX-XX-XX
通信作者: 王洪, wh@csu.edu.cn
基金项目: 湖南省社科基金资助项目(No.23YBA009)
Foundation Items: The Social Science Fund Project of Hunan Province (No.23YBA009)