软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007661] [CSTR: 32375.14.jos.007661]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

FedProbe: 代理模型引导的可解释联邦学习投毒攻击防御框架*

曹益皓[1,2], 张建标[1,2], 赵亚茹[3], 韩宇飞[1,2], 张承昱[1,2], 叶 涛[4]

[1](北京工业大学 计算机学院, 北京 100124)
[2](可信计算北京市重点实验室 (北京工业大学), 北京 100124)
[3](湖南科技大学 计算机科学与工程学院, 湖南 湘潭, 411201)
[4](青海民族大学 计算机学院, 青海 西宁, 810007)
通信作者: 张建标, E-mail: zjb@bjut.edu.cn

摘 要: 联邦学习允许多个参与方在不共享本地私有数据的前提下协同训练一个共享的深度学习模型. 然而, 该范式在实际应用中极易受到日益复杂的投毒攻击威胁. 现有的防御方法在检测效率、对多样化攻击的泛化能力以及在非独立同分布 (non-IID) 数据环境下的性能稳定性方面仍存在局限. 为应对上述挑战, 提出一种名为 FedProbe 的、由代理模型引导的可解释联邦学习投毒攻击防御框架. FedProbe 采用两阶段机制: 首先, 在代理模型引导阶段, 利用服务器端的可信样本集计算各本地模型间的 KL 散度以度量其行为相似性, 并确定最优聚类数量. 在剔除离群更新后, 各簇被聚合成代理模型, 此举旨在有效划分由不同数据分布训练的模型并提升后续检测效率. 其次, 在可解释分析阶段, FedProbe 利用 SHAP 可解释技术对代理模型进行深度分析, 计算其关键特征归因, 并最终通过跨类别可疑度分数来精准识别并剔除潜在的恶意模型. 实验结果表明, FedProbe 在多种基准数据集上展现出卓越的性能与鲁棒性. 在良性环境中, 其收敛时间仅为标准聚合算法的 1.11–1.21 倍, 性能开销极小. 在安全性方面, 即使在恶意用户占比高达 40% 的极端场景下, FedProbe 在抵御无目标攻击和后门攻击时仍能保持优越的收敛性, 并将攻击成功率控制在 25% 以下; 而在面对自适应攻击时, 其在防御有效性与任务准确率方面均具有显著优势.

关键词: 联邦学习; 投毒攻击; 防御策略; 可解释性; 代理模型
中图法分类号: TP18

中文引用格式: 曹益皓, 张建标, 赵亚茹, 韩宇飞, 张承昱, 叶涛. FedProbe: 代理模型引导的可解释联邦学习投毒攻击防御框架. 软件学报. http://www.jos.org.cn/1000-9825/7661.htm

英文引用格式: Cao YH, Zhang JB, Zhao YR, Han YF, Zhang CY, Ye T. FedProbe: Proxy-model-guided Explainable Defense Framework for Poisoning Attacks in Federated Learning. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7661.htm

FedProbe: Proxy-model-guided Explainable Defense Framework for Poisoning Attacks in Federated Learning

CAO Yi-Hao[1,2], ZHANG Jian-Biao[1,2], ZHAO Ya-Ru[3], HAN Yu-Fei[1,2], ZHANG Cheng-Yu[1,2], YE Tao[4]

[1](College of Computer Science, Beijing University of Technology, Beijing 100124, China)
[2](Beijing Key Laboratory of Trusted Computing (Beijing University of Technology), Beijing 100124, China)
[3](College of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China)
[4](College of Computer, Qinghai Minzu University, Xining 810007, China)

Abstract: Federated Learning enables collaborative model training among multiple clients without exchanging their local private data. However, this paradigm is vulnerable to increasingly sophisticated poisoning attacks. Existing defense mechanisms often exhibit

* 基金项目: 北京市自然科学基金 (M21039)
收稿时间: 2025-08-06; 修改时间: 2025-12-30, 2026-01-22, 2026-02-23; 采用时间: 2026-03-04; jos 在线出版时间: 2026-06-03

shortcomings in detection efficiency, generalization to diverse attacks, and performance stability under non-independent and identically distributed (non-IID) data. To address these challenges, this study proposes FedProbe, an interpretable federated poisoning attack defense framework guided by proxy models. FedProbe adopts a two-stage mechanism. In the first stage, under proxy model guidance, the framework computes the KL divergence among local models on a trusted server-side dataset to measure their behavioral similarity and determine the optimal number of clusters. After removing outlier updates, each cluster is aggregated into a proxy model, aiming to effectively partition models trained on different data distributions and improve subsequent detection efficiency. In the second stage, an interpretable analysis is conducted. FedProbe leverages the SHAP technique to perform an in-depth analysis of the proxy models, computing key feature attributions and ultimately using cross-category suspicion scores to identify and remove potentially malicious models accurately. Experimental results demonstrate that FedProbe exhibits superior performance and robustness across various benchmark datasets. In benign settings, the convergence time is only 1.11 to 1.21 times that of standard aggregation algorithms, indicating minimal overhead. In terms of security, even in extreme scenarios with up to 40% malicious clients, FedProbe maintains superior convergence when defending against untargeted attacks and backdoor attacks, while keeping the attack success rate below 25%. Moreover, when facing adaptive attacks, it shows significant advantages in both defense effectiveness and task accuracy.

Key words: federated learning (FL); poisoning attack; defense strategy; explainable; proxy model

随着人工智能技术已深度融入社会生活的方方面面, 其发展高度依赖海量数据的支撑. 在传统的集中式机器学习范式中, 用户数据需被收集并上传至中心服务器进行模型训练. 尽管此模式为服务提供了便利, 但原始数据的集中传输与存储引发了严重的隐私问题, 并带来了显著的泄露风险.

为应对此挑战, 联邦学习 (federated learning, FL) 作为一种变革性的协作式机器学习框架应运而生[1]. 该框架允许模型在分布式设备网络 (如移动电话、物联网设备) 上进行协同训练, 而无需交换用户的本地原始数据. 在 FL 中, 各客户端利用本地数据独立训练模型, 仅将模型更新发送至中央服务器. 服务器则聚合这些更新, 以构建一个性能与泛化能力更强的全局模型. 由于敏感数据始终留存于用户本地, FL 在物联网、金融、医疗保健等数据高度敏感的领域展现出广阔的应用前景.

然而, 联邦学习固有的安全脆弱性对其广泛的产业化应用构成了严峻挑战[2]. 研究表明, 攻击者可通过分析本地模型更新或聚合后的全局模型, 实施成员推理[3]、数据重构等隐私攻击, 窃取用户敏感信息[4]. 与隐私攻击不同, 投毒攻击是另一类更为直接的威胁[5]. 在此类攻击中, 恶意客户端向全局模型注入精心构造的恶意更新, 意图破坏模型的完整性[6] (如降低整体精度) 或可用性[7,8] (如植入后门, 使其在特定输入下产生预设的错误输出). 在自动驾驶等高安全需求场景中, 此类攻击可能导致灾难性后果. 因此, 如何有效检测并防御投毒攻击, 确保全局模型的可靠性与完整性, 已成为当前联邦学习安全领域的研究焦点.

近年来, 针对投毒攻击已涌现出多种防御策略, 大致可归为两大类: 影响缓解与异常检测[9]. 影响缓解策略旨在通过修正模型参数或引入差分隐私噪声等手段, 削弱恶意更新对全局模型的负面影响[10,11]. 异常检测策略则通常采用统计或机器学习技术, 对客户端上传的模型更新进行甄别, 并直接剔除被判定为恶意的更新[12−16]. 尽管如此, 现有防御技术仍面临普遍的局限性. 首先, 许多防御方法仅针对特定攻击类型设计, 普适性不足. 其次, 部分方法在提升安全性的同时, 引入了高昂的计算开销, 牺牲了训练效率. 最后, 现有防御大多依赖简单的统计或距离度量, 而攻击者可采取自适应策略对其进行规避, 这使得防御的鲁棒性面临严峻考验.

针对现有防御策略在普适性、效率及对抗自适应攻击方面的局限性, 本文提出一种名为 FedProbe、代理模型引导的可解释联邦投毒攻击防御框架. FedProbe 的核心机制在于: 服务器端持有一个小规模的、可信的样本集, 并以 SHAP (Shapley additive explanations)[17]可解释性分析为核心度量工具, 对客户端上传的模型更新进行深度探测. 在此过程中, FedProbe 通过一种跨类别可疑度分数来量化模型更新对关键特征的归因是否存在异常偏移. 这一基于 SHAP 的分析旨在精准识别包括隐蔽后门在内的多样化投毒攻击, 并对攻击者的自适应策略展现出优越的鲁棒性.

然而, 单纯依赖 SHAP 值分析的方案在实际应用中面临两大挑战. 其一, 若服务器对每个本地模型逐一进行 SHAP 分析, 将引入巨大的计算开销, 严重影响检测效率. 其二, 联邦场景中普遍存在的非独立同分布 (non-independent and identically distributed, non-IID)[9]的数据异构场景可能导致正常的模型差异被误判为恶意行为, 从而影响检测准确性.

为应对上述挑战, FedProbe 集成了一种代理模型引导机制. 该机制首先计算各模型更新在服务器样本集上的

输出概率分布, 并利用 Kullback-Leibler (KL) 散度[18]度量它们之间的行为相似性. 接着, 基于此相似性度量对模型更新进行聚类, 并为每个聚类生成一个代表性的代理更新, 而计算成本高昂的可解释性分析仅需在这些代理之上执行. 该机制通过聚类平滑个体差异, 显著抑制数据异构性带来的噪声, 并在 FL 任务场景下将计算复杂度从 $\mathcal{O}(n)$ 降至 $\mathcal{O}(k)(k \ll n)$, 从而兼顾检测精度与系统可扩展性.

具体来说, 本文主要贡献以及解决的挑战如下.

(1) 提出一种代理模型引导的可解释联邦投毒攻击防御框架 FedProbe. 该框架利用模型更新在可信样本集上的可解释性来识别恶意行为, 旨在提升对多样化及自适应投毒攻击的检测鲁棒性.

(2) 设计并集成了一种代理模型引导机制, 以应对模型可解释性方法在联邦场景下的效率瓶颈及异质性干扰问题. 通过基于 KL 散度的相似性度量对本地更新进行聚类并构建代理更新, 该机制在显著降低可解释性分析计算开销的同时, 也提高了框架在异质数据环境下的检测稳定性, 增强了整体可扩展性.

(3) 通过在多种基准数据集上的广泛实验验证了所提框架的有效性. 实验结果表明, 即使在数据 non-IID 且攻击者比例高达 40% 的条件下, 该框架仍能高效检测并防御多种攻击, 包括无目标投毒攻击和自适应后门攻击, 展现出优越的性能和鲁棒性.

本文第 1 节介绍联邦学习投毒攻击的背景知识和相关工作. 第 2 节定义本文所针对的威胁模型与核心防御目标. 第 3 节详细阐述所提检测框架的设计与实现. 第 4 节通过一系列实验, 对框架的性能与精度进行综合评估. 第 5 节探讨本方法的潜在局限性与未来的改进方向. 第 6 节对全文进行总结.

1 背景知识和相关工作

本节主要论述理解本工作的前置背景知识以及联邦学习投毒攻击防御的相关工作.

1.1 联邦学习投毒攻击

在联邦学习框架中, 客户端利用本地数据训练模型, 并与服务器协作以聚合生成全局模型. 然而, 恶意客户端可能利用其本地数据或模型参数对全局模型进行投毒, 进而发起无目标攻击 (亦称拜占庭攻击)[19]或目标攻击 (亦称后门攻击)[5].

无目标攻击旨在通过攻击者篡改的数据或模型参数, 显著降低全局模型的性能, 甚至使其完全失效. 典型的无目标攻击包括噪声攻击、符号翻转攻击和自适应无目标攻击. 噪声攻击[6]的攻击者在模型参数或数据中注入高斯噪声, 或直接替换正常数据/参数, 从而阻碍全局模型的收敛. 符号翻转攻击[20]通过反转梯度更新的符号, 能够更隐蔽地污染全局模型, 进而规避某些基于参数欧氏距离的鲁棒性检测算法. 更为高级的自适应无目标攻击[21]则能够在每一轮迭代中, 通过试探并学习检测算法的响应, 动态调整其污染策略, 例如改变梯度污染的维度或层数, 使得此类攻击更难被检测和防御.

与无目标攻击旨在普遍降低模型的整体性能不同, 目标攻击通常表现出更强的隐蔽性. 在此类攻击中, 恶意客户端通过精心构造其本地数据或模型更新, 在全局模型中植入特定的后门[5]. 该后门在处理常规, 无触发器的输入样本时通常保持静默, 全局模型行为正常, 因此难以被察觉. 然而, 一旦攻击者提供包含特定触发器的输入, 后门即被激活, 迫使全局模型按照攻击者预设的方式产生错误输出. 例如, 将某一特定类别的图像错误地分类为攻击者指定的目标类别[22].

实现目标攻击的手段呈现多样性. 一种常见的方法是数据投毒, 即恶意客户端在其本地训练数据中混入少量携带特定触发器模式 (如图像中的微小像素块) 的样本, 并将这些样本的标签篡改为攻击者预期的目标标签. 通过此类数据训练得到的本地模型在上传至服务器并参与聚合后, 会将此后门知识传播至全局模型[23]. 另一种策略是模型投毒或直接操控模型参数, 恶意客户端可直接修改其本地模型的参数, 使其在遇到特定触发器时展现预设行为.

自适应目标攻击将威胁提升至新的层面, 其核心在于动态性与环境感知能力[24]. 恶意客户端会持续监控联邦学习过程的动态, 全局模型的演进以及现有防御机制, 并据此调整攻击策略, 如优化触发器模式, 改变投毒数据比例或调整模型更新的递交时机, 以规避检测并最大化攻击效果.

1.2 现有防御方法分析

本节从影响缓解策略和异常检测过滤策略两个主要方面对现有的联邦学习防御方法进行梳理与评析.

● 影响缓解策略. 该类方法并不直接剔除疑似的恶意模型, 而是旨在通过应用机器学习理论或统计学习方法来减轻中毒模型对全局模型聚合的负面影响. 例如, Median[25]和 Repeated Median[26]利用中值定理对各本地模型参数进行聚合, 以削弱极端参数的影响. Krum[27]和 Multi-Krum[27] (及其变种如 SEAR[28]) 则基于欧氏距离度量模型间的相似性, 通过对本地模型间的参数距离进行排序或评估, 选择距离中心或多数模型较近的 $n-f$ 个模型参与聚合 (其中 n 为客户端总数, f 为预估的恶意客户端数量上限). 上述基于距离或统计排序的聚合方法在抵御一些较为简单的无目标攻击时表现出较好的有效性, 但对于复杂的自适应攻击和隐蔽性强的后门攻击, 其防御能力相对有限.

一些研究工作从其他角度探索影响缓解策略. LeadFL[10]在客户端本地训练阶段引入损失函数惩罚项, 旨在减轻中毒数据对本地模型训练过程的污染程度. 类似地, FreqFed[29]提出将模型参数转换至频域进行分析, 通过在频域中识别并保留那些携带足够权重信息的所谓“核心频率分量”, 期望在客户端本地训练过程中有效削弱恶意更新的影响. 然而, 尽管这类方法在某些针对后门攻击的强恶意假设下能够展现出一定的抵抗能力, 但它们也会显著增加训练开销或降低训练效率.

此外, 还有在服务器端进行参数修正的尝试. Residual[30]采用鲁棒回归技术, 在服务器端对聚合前的模型参数进行高维线性回归建模, 并将那些偏离预设或动态计算的置信区间的参数值矫正回置信区间内部. ADFL[31]则利用对抗蒸馏的思想, 借助一个良性模型在可信样本集上的输出来指导和重构可能已中毒的全局模型, 以期矫正其恶意行为. 然而, 这类矫正方法在 non-IID 的数据设定下, 可能会错误地将那些因数据分布与主流不同而表现“异常”但本身是良性的客户端模型也纳入矫正范围, 从而不当调整其参数, 反而导致全局模型准确率的下降.

● 异常检测过滤策略. 这类方法通常利用多样化的检测指标对客户端上传的模型进行分析与验证, 旨在识别并直接排除可疑的恶意模型. 例如, FoolsGold[32]和 FLTrust[11]利用各本地模型更新与聚合模型 (或彼此之间) 的余弦相似度作为核心度量指标来进行检测. 然而, 这类依赖单一或少数相似性度量的指标在面对精心设计的自适应攻击时可能失效, 攻击者相对容易构造出能规避此类检测的恶意更新. 为应对此挑战, CrowdGuard[33]提出一种基于隐藏层激活模式的后门检测指标, 并结合客户端投票机制来识别和排除恶意模型. 尽管如此, 该方法对参与投票的客户端的诚实性和可靠性提出了较高的假设. FLAME[15]则通过向模型更新中注入微小噪声, 并结合层次聚类离群点检测技术来识别和排除潜在的恶意模型.

另一类检测方法利用模型可解释性进行深入分析. 例如, SHERPA[12]和蒋伟进等人[34]利用基于 SHAP 的模型可解释性技术, 通过分析模型在服务器端持有的少量可信样本上的行为特征 (如关键特征贡献度) 来进行恶意模型检测与排除, 其思路与本文提出的部分方法有相似之处. 然而, 精确计算 SHAP 值通常计算成本极高, 尤其是在参与联邦学习的客户端数量众多时, 会显著影响防御机制的整体效率. 此外, 若仅依赖基于 SHAP 值的单一分析维度, 在 non-IID 环境下, 同样可能导致对那些数据分布独特但行为良性的客户端模型产生误判, 进而影响全局模型的性能.

综上所述, 影响缓解策略能够在不直接丢弃客户端模型的前提下, 一定程度减轻全局模型遭受投毒攻击的损害. 但其代价往往是训练效率的降低, 或可能对全局模型的最终性能 (尤其在复杂攻击和异构数据场景下) 造成不利影响. 另一方面, 前述的异常检测过滤方法虽然致力于直接移除威胁, 但普遍面临误检, 对高级自适应攻击的防御能力不足以及整体鲁棒性欠佳等挑战. 本文针对现有方法的不足, 提出一种代理模型引导的模型可解释性的新型防御框架, 旨在有效抵抗多样化投毒攻击的同时, 显著提升防御机制的检测效率与准确性.

2 威胁建模

本节阐述本文研究的 FL 系统设定和典型的对抗模型, 并明确防御目标与面临的挑战.

2.1 系统设置

本文考虑一个典型的 FL 场景. 在该场景中, 多个客户端在各自的本地数据上进行模型训练, 而一个中央服务

器负责聚合各客户端上传的本地模型. 具体而言, 各参与方的定义如下.

● 客户端. 假设存在 n 个客户端, 记为集合 $\mathbb{C}=\{C_1,C_2,\ldots,C_n\}$. 每个客户端 $C_i\in\mathbb{C}$ 持有其本地私有数据集 D_i. 这些数据集的分布通常是异构的, 即 non-IID. 在每轮训练开始时, 被服务器选中的客户端 C_i 利用其私有数据 D_i 训练本地模型, 并将训练完成的模型参数上传至服务器 $\mathcal{S}$. 恶意客户端为了达到污染全局模型的目的, 也会遵循联邦学习的协议格式, 上传其精心构造的恶意模型更新.

● 服务器. 假设服务器是“诚实但好奇”的. 这代表服务器会诚实地执行预设的良性操作 (例如模型聚合, 参数分发等), 但可能试图从收集到的客户端模型参数中推断客户端的私有信息. 此外, 借鉴 FLTrust[11]工作的设定, 本文假设 FL 服务部署方在服务器端预先收集了一小部分可信的样本集合 $\mathcal{X}$, 用于辅助验证或检测客户端上传模型的行为. 此外, 鉴于本文聚焦于防范来自客户端的恶意投毒攻击, 服务器自身的潜在系统安全漏洞或其可能发起的其他隐私攻击行为不在讨论范围之内.

2.2 对手建模

本文主要考虑两个恶意对手 $\mathcal{A}_1$ 和 $\mathcal{A}_2$, 它们的定义分别如下.

定义 1. $\mathcal{A}_1$. 恶意对手 $\mathcal{A}_1$ 具备控制联邦学习中若干客户端的能力, 包括其本地数据样本及模型参数. $\mathcal{A}_1$ 旨在通过无目标攻击或后门攻击手段污染全局模型或植入特定后门[30]. 具体而言, $\mathcal{A}_1$ 在联邦学习训练过程中的每一轮次或选定若干轮次, 通过篡改本地数据样本或直接修改本地模型参数的方式实施投毒. 此外, 本文做出如下假设: $\mathcal{A}_1$ 所控制的各恶意客户端之间无法相互勾结; 同时, $\mathcal{A}_1$ 对服务器端可能部署的防御机制完全无知.

● 目标 1 (可用性破坏). $\mathcal{A}_1$ 通过对所控客户端的本地模型进行投毒, 旨在显著降低全局模型的整体性能, 极端情况下甚至导致全局模型完全失效.

● 目标 2 (后门注入). $\mathcal{A}_1$ 致力于在全局模型中植入后门, 使得模型在接收到包含特定触发器的输入时, 能够产生攻击者预期的输出. 具体实现上, $\mathcal{A}_1$ 将精心构造携带触发器的样本 x^* 并赋予其目标标签 t_a, 用于训练其控制的本地模型. 经过模型聚合后, 当全局模型接收到包含该触发器的输入时, 便会将其错误分类为目标标签 t_a.

此外, 鉴于破坏模型可用性与植入后门这两类攻击意图通常被认为是相互独立的, 我们假设在单次联邦学习任务执行过程中, $\mathcal{A}_1$ 仅会选择实施其中一种攻击.

定义 2. $\mathcal{A}_2$. 本文进一步考虑一个更高级的恶意对手 $\mathcal{A}_2$. 与 $\mathcal{A}_1$ 不同, $\mathcal{A}_2$ 在执行攻击时, 能够感知并适应服务器端可能存在的防御策略以及本地模型的结构特性, 从而对其投毒行为进行自适应调整, 以最大化攻击效果并规避检测[21]. 在追求达成目标 1 或目标 2 的同时, $\mathcal{A}_2$ 还具备以下额外目标.

● 目标 3 (隐蔽性). $\mathcal{A}_2$ 会采用更为复杂和精密的投毒机制. 例如, 它可能针对已知的服务器防御措施, 对模型参数或数据样本进行更具策略性的篡改, 以确保在不激活后门触发器的情况下, 受污染模型在主任务及其他相关评估指标上均表现得尽可能与良性模型一致, 从而有效规避各类异常检测机制.

2.3 防御目标与挑战

根据以上建模和系统设置, FedProbe 所需要达成的防御目标如下.

● 目标 1 (抵御常规投毒攻击). FedProbe 需能在不过分牺牲模型在主要任务上的准确率的前提下, 有效抵御各类 $\mathcal{A}_1$ 所发动的常规投毒攻击, 并在此过程中展现出较强的鲁棒性, 尤其是在 non-IID 的数据环境下.

● 目标 2 (防御自适应攻击). FedProbe 应能有效应对由具备学习与适应能力的隐蔽的对手 $\mathcal{A}_2$ 所发起的自适应投毒攻击, 特别是要能准确识别并拦截此类对手上传至服务器的、经过精心伪装的恶意模型.

为了达成以上目标, 本文方法面临着以下挑战.

● 挑战 1 (non-IID 下检测鲁棒性). 在 non-IID 场景中, FedProbe 要求能够最大限度区分由局部数据质量不高或数据分布独特所导致的表现异常的良性客户端模型与真正的恶意中毒模型, 从而在高效过滤中毒模型的同时, 最大限度地降低对全局模型最终性能的潜在负面影响.

● 挑战 2 (利用良性样本检测自适应攻击). 鉴于自适应投毒模型在处理无特定触发器的常规输入时, 其行为往往与良性模型高度相似, FedProbe 需要在仅依赖服务器端持有的可信样本集的情况下, 依然能够高效地检测出这

类经过自适应策略强化的投毒模型.

3 FedProbe 方法

本节旨在具体叙述 FedProbe 方法以应对第 2 节所述的威胁模型与防御挑战. FedProbe 的整体架构如图 1 所示, 其核心在于利用代理模型引导的 SHAP 值分析, 以有效抵御来自恶意客户端的投毒攻击. 此架构的运行流程主要包括以下 3 个阶段.

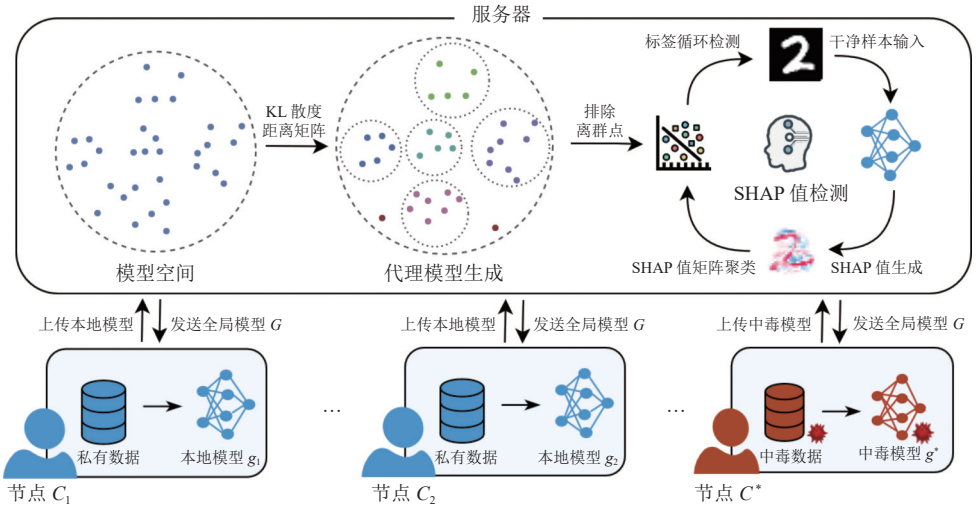


图 1 FedProbe 架构图

● 初始化阶段: 在 FL 任务启动时, 服务器将 $\mathcal{S}$ 从客户端集合中随机选取 n 个参与方. 随后, $\mathcal{S}$ 对全局模型 G_0 及相关本地训练参数进行初始化, 并将这些信息分发给所有被选中的客户端. 为确保通信过程的机密性与完整性, 服务器会与每个客户端协商建立独立的会话密钥.

● 本地训练阶段: 客户端 C_i 利用其本地私有数据集, 并依据服务器下发的全局模型和参数, 独立训练本地模型 g_{r+1}. 与此同时, 恶意客户端会使用其中毒数据集进行训练, 生成并上传含有后门或恶意更新的本地模型 g^*_{r+1} 至服务器.

● 检测聚合阶段: 服务器 $\mathcal{S}$ 在接收到各客户端上传的本地模型后, 将启动检测与聚合流程. 首先, $\mathcal{S}$ 根据各模型在可信样本集上的输出 (通过 KL 散度衡量相似性), 并结合轮廓系数来确定最优聚类数量, 将相似的模型分组, 并将每组聚合为一个代理模型 (详见第 3.1 节). 随后, 为了检测 $\mathcal{A}_2$ 发动的隐蔽攻击, 系统会裁剪仅包含单个成员模型的簇, 并对其余的代理模型进行基于 SHAP 值的深入检测. 该检测通过分析模型在分类良性样本时对不同特征的归因模式, 构建 SHAP 值矩阵, 并利用层次聚类算法来跨类别的识别和排除行为异常的代理模型 (详见第 3.2 节). 最终, 只有通过检测的代理模型所对应的原始本地模型才会被聚合, 形成新一轮的全局模型 G_{r+1}. 服务器 $\mathcal{S}$ 将 G_{r+1} 分发给下一轮随机选择的客户端, 循环往复直至达到 $\mathcal{S}$ 规定的任务轮次要求.

综上所述, 本方法以代理模型引导机制带来效率提升和 SHAP 值对模型内在行为的精确刻画为两大核心策略, 构建了一种针对模型投毒攻击的鲁棒检测与防御机制.

3.1 代理模型引导机制

为了应对挑战, 本文引入了代理模型作为引导 SHAP 值可解释分析的高效方法. 其核心动机在于, 当客户端数量庞大且其数据分布存在显著异构性时, 对每个本地模型进行独立的深度分析会引发难以承受的计算开销. 通过将行为相似的本地模型聚类并聚合为代理模型, 可以显著降低需要进行精细化恶意检测的单元总数, 从而大幅提升防御机制的整体效率. 代理模型的具体生成过程详见算法 1.

算法 1. 代理模型引导机制.

输入: 客户端模型集合 $\{g_1, g_2, \ldots, g_n\}$, 服务器 C 类验证样本集合 $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_c\}$, 轮廓系数搜索范围 $[K_{\min}, K_{\max}]$, *Softmax* 温度 τ;

输出: 代理模型集合 $g' = \{g'_1, g'_2, \ldots, g'_m\}$, 原始模型映射 I_p.

1. 初始化 $g_p \leftarrow 0$, $I_p \leftarrow 0$;
2. **for** $i = 1$ **to** n **do**
3. \\ 计算模型间的输出分布差异
3 **for** $c = 1$ **to** C **do**
5. $p_i \leftarrow Softmax(g_i(\mathcal{X}_c; \tau))$;
6. $v_i \leftarrow \oplus_{j=1}^{n} \oplus_{c=1}^{C} \frac{1}{2}[D_{\mathrm{KL}}(p_{j,c} \| p_{i,c}) + D_{\mathrm{KL}}(p_{i,c} \| p_{j,c})]$;
7. \\ 轮廓系数分析最优聚类数
8. 初始化 $K_{\text{best}} \leftarrow K_{\min}$, $S_{\text{best}} \leftarrow -1$;
9. **for** $k = K_{\min}$ **to** $K_{\max}$ **do**
10. $L_k \leftarrow KMeans(\{v_1, v_2, \ldots, v_n\}, k)$;
11. $S_k \leftarrow SilhouetteScore(\{v_1, v_2, \ldots, v_n\}, L_k)$;
12. **if** $S_{\text{best}} < S_k$ **then**
13. $S_{\text{best}} \leftarrow S_k$, $K_{\text{best}} \leftarrow k$;
14. \\ 使用 K_{best} 聚类并生成代理模型
15. $L_{\text{best}} \leftarrow KMeans(\{v_1, v_2, \ldots, v_n\}, K_{\text{best}})$;
16. **for each** cluster k **in** L_{best} **do**
17. **if** $|k| > 1$ **then**
18. $g'_k \leftarrow \frac{1}{|k|} \sum_{g_i \in k} g_i$;
19. $g' \leftarrow g' \cup \{g'_k\}$;
20. **return** g', I_p

该过程首先需要计算模型间的行为相似度. 具体而言, 服务器在接收到所有本地模型 g_i 后, 利用可信样本集 $\mathcal{X}$ 对各模型进行评估, 并计算其预测概率分布之间的 KL 散度, 如公式 (1) 所示:

$$D_{\mathrm{KL}}(p_j \| p_i) = \sum_{c=1}^{C} p_{j,c} \log \frac{p_{j,c}}{p_{i,c}} \tag{1}$$

其中, p_j 和 p_i 为本地模型输出的概率分布. KL 散度能够有效量化不同模型在处理相同数据时其预测结果的差异性. 这种基于模型外在行为的度量方式, 能够直观地反映模型所学到的决策边界与知识特征.

与直接计算模型参数的欧氏距离或余弦相似度等方法相比, 采用 KL 散度衡量模型行为具有两大优势. 首先, 在 non-IID 场景下, 参数空间中的距离未必能准确反映模型功能上的差异, 而基于模型输出行为的分析则能更灵敏地捕捉到这种差异. 其次, 参数的微小扰动与模型行为的显著变化之间不存在稳定的对应关系, 这使得基于参数的度量方法在区分恶意模型与正常模型时缺乏鲁棒性. 相比之下, 本方法直接在模型的输出空间进行度量, 不依赖于模型内部参数的具体数值, 而是聚焦于模型的最终功能表现. 因此, 它为在 non-IID 场景下识别行为异常的模型提供了更为鲁棒和直接的依据.

在每个 g_i 都生成 KL 散度后, 将对所得散度值进行平均, 并将其拼接成距离矩阵 (算法 1 第 6 行). 随后, 需要确定用于聚类的代理模型数量. 若在每一轮通信中均采用固定的聚类数量, 将难以适应模型分布的动态变化, 从而可能导致漏报或误报风险的增加. 为此, 本方法采用轮廓系数来自适应地确定当前轮次中的最优聚类数量, 以提升

聚类结果的有效性 (算法 1 第 7–13 行). 轮廓系数通过计算聚类簇之间的分离度以及簇内的紧密程度来判断聚类的质量, 对于每一个本地模型 g_i, 聚类总的轮廓系数 S 如公式 (2) 所示:

$$S = \sum_{i=1}^{n} \frac{b(g_i) - a(g_i)}{\max\{a(g_i), b(g_i)\}} \tag{2}$$

其中, $a(g_i)$ 为 g_i 所属簇内其他模型的平均距离, $a(g_i)$ 越小, 说明簇的内聚性越高. $b(g_i)$ 为 g_i 与其他簇中模型的平均距离最小值, $b(g_i)$ 越大, 则说明簇间分离度越高. 需要指出的是, 轮廓系数在评估凸显球状结构的簇时效果最佳, 而对于 HDBSCAN 等旨在发现非规则形状簇的算法则适用性较差. 基于此特性, 本方法采用 K-means 算法对 KL 散度距离矩阵进行聚类.

确定最优聚类数量 (即代理模型的数量) 后, 方法会将任何仅包含单个成员的簇识别为离群模型. 此判断基于如下解释: 在 KL 散度构成的输出行为空间中, 即使是 non-IID 环境下具有特定数据偏斜的客户端, 也会与拥有相似主导类别的其他客户端形成簇. 若某个模型在所有验证样本上的输出分布与其余所有客户端均不为同一簇内, 这通常标志着本地训练失败或伪装失败的中毒行为. 余下的本地模型则根据其所属的簇被聚合为相应的代理模型, 并以此引导下一阶段的 SHAP 值分析流程.

3.2 SHAP 值分析

尽管基于 KL 散度的代理模型引导机制的外部行为度量能够识别出攻击者占比较少且行为显著异常的模型, 但其难以抵御攻击者占比较高时发动的更为隐蔽的投毒攻击. 例如, 自适应攻击者 $\mathcal{A}_2$ 可在每轮精心构造中毒模型, 使其在可信样本集上的行为与良性模型高度一致, 从而成功规避初步检测. 这种局限性的根源在于, 聚类仅完成了模型间行为相似性的表象归类, 却无法探究模型做出特定决策的内在依据, 使防御体系面临着无法甄别此类模型的风险: 其外在行为看似正常, 内部决策逻辑却已被恶意篡改.

为弥补这一防御短板, 实现对模型内部决策逻辑的深度审查, FedProbe 的代理模型主要用于引导基于 SHAP 值的可解释性分析. SHAP 是一种基于合作博弈论且具有坚实理论基础的模型无关可解释性框架. 它借鉴了合作博弈论中的夏普利值 (Shapley value) 思想, 旨在为复杂的机器学习模型对任意单一样本的预测, 提供具有一致性和局部准确性的归因解释. 其核心目标是公平地量化在一次具体的预测任务中, 每一个输入特征变量所做出的边际贡献, 从而揭示模型在该决策上的内部行为范式.

具体而言, 对于一个给定的预测样本 $\mathcal{X}_c$, SHAP 为 g_i' 中每个特征计算一个归因值 (即 SHAP 值 ϕ_i), 该值代表此特征在所有可能的特征组合中, 对模型输出的平均边际效应, 如公式 (3) 所示:

$$\phi_i(g_i', \mathcal{X}_c) = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[g'_{i,\mathcal{X}_c}(S \cup \{i\}) - g'_{i,\mathcal{X}_c}(S) \right] \tag{3}$$

其中, F 为全部特征集合, S 为 F 的一个子集, $g'_{i,\mathcal{X}_c}(S)$ 代表当代理模型 g_i' 只得知子集 S 中的特征时, 对样本 $\mathcal{X}_c$ 做出的预测.

图 2 以一个遭受后门攻击的中毒代理模型为例, 直观地展示了如何通过 SHAP 值识别其内在逻辑异常, 其中, 红色像素代表该特征对预测类别的正向贡献, 蓝色代表负向贡献. 由图 2 中对第 2 和 3 行的样本分析可见, 该中毒模型在分类良性样本时呈现出明显反常: 尽管样本的真实标签为类别 2, 但与该类别相关的特征归因却呈现显著的负向贡献; 与此同时, 与攻击目标类别 8 相关的特征归因则表现出非正常的强正向贡献.

这种反常的归因模式与既定的后门攻击任务高度吻合, 即攻击者意图将携带触发器的类别 2 样本误导为类别 8. 即便在不含触发器的良性样本上进行测试, 中毒模型的内部决策逻辑也已遭到扰动. SHAP 值所揭示的这种归因模式的系统性偏移, 正是模型已被后门污染的可靠凭证.

算法 2 详细阐述了基于 SHAP 值的深度分析流程. 在当前训练轮次中, 代理模型一经生成, 服务器便会利用其本地的验证样本集 $\mathcal{X}$ 为每个代理模型 g_i', 针对每个数据类别 $\mathcal{X}_c$ 计算 SHAP 值 (算法 2 第 1–5 行). 在通过公式 (3) 计算后, 会得出每个特征对于代理模型 g_i' 的 SHAP 值, 此过程输出的 SHAP 值 ϕ_i 为一个高维归因张量. 以图像数据集为例, g_i' 对每个样本 $\mathcal{X}_c$ 计算的 SHAP 值维度为 $\phi_i \in \mathbb{R}^{W \times H \times C \times |\mathcal{X}|}$, 其中 W、H、C 为样本的长宽以及颜色通道.

随后, 为便于度量, ϕ_i 被看作长度为 $|\mathcal{X}|$ 的向量 $\{\phi_{i,1},\phi_{i,2},\ldots,\phi_{i,c}\}$, 通过计算不同代理模型 g'_i 对相同类别样本的 SHAP 张量之间的成对余弦距离, 可以构建长度为 $|\mathcal{X}|$ 的多个距离矩阵. 最后, FedProbe 采用层次聚类算法对每个距离矩阵进行分析 (算法 2 第 6–11 行).

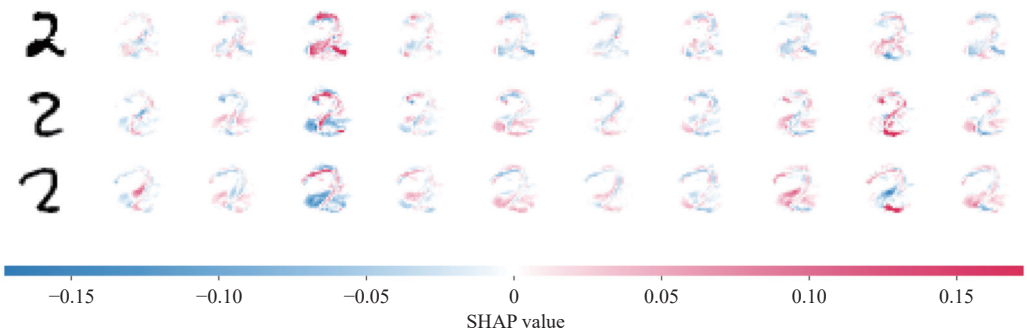


图 2　中毒代理模型在 MNIST 数据集中预测样本类别为 2 时的 SHAP 值 (共 3 个样本)

算法 2. SHAP 值分析.

输入: 代理模型集合 g', 原始模型映射 I_p, 服务器 C 类验证样本集合 $\mathcal{X}=\{\mathcal{X}_1,\mathcal{X}_2,\ldots,\mathcal{X}_C\}$;
输出: 全局模型 G.

1. \\ 计算模型 SHAP 值
2. **for each** $g'_i \in g'$ **do**
3. **for** $c \leftarrow 1$ **to** C **do**
4. $\phi_{i,c} \leftarrow SHAP(g'_i, \mathcal{X}_c)$;
5. \\ 逐类计算相似度聚类
6. **for** $c \leftarrow 1$ **to** C **do**
7. 初始化距离矩阵 $\mathcal{D}_c \in \mathbb{R}^{|g'|\times|g'|}$;
8. **for** $i \leftarrow 1$ **to** $|g'|$ **do**
9. **for** $j \leftarrow i+1$ **to** $|g'|$ **do**
10. $\mathcal{D}_c(i,j) \leftarrow 1-\cos(\phi_{i,c},\phi_{j,c})$;
11. $L_c \leftarrow AggCluster(\mathcal{D}_c, 2)$;
12. \\ 计算可疑分数
13. 初始化可疑代理模型分数 $s \in \mathbb{R}^{|g'|}$ 零向量;
14. 根据 I_p 得出聚类 L_c 中各簇所含原始模型平均规模 $\bar{n}_c$;
15. **for** $c \leftarrow 1$ **to** C **do**
16. **for each** cluster k' **in** L_c **do**
17. **if** $|k'| < \bar{n}_c$ **then**
18. **for each** g'_i **in** k' **do**
19. $s_i \leftarrow s_i + 1$;
20. \\ 识别可疑模型
21. $\eta \leftarrow mean(s)$;
22. $g^* \leftarrow \{g'_i | s_i > \eta\}$;

23. $g' \leftarrow \cup_{g'_i \notin g^*} \{g'_i\}$;
24. $G \leftarrow \frac{1}{|g'|} \sum_{g'_i \in g'} g'_i$;
25. **return** G

按照基于“大多数诚实”这一先验假设, FedProbe 应将规模较小的簇识别为行为异常的潜在恶意簇直接排除. 然而, 该策略存在两大内在脆弱性: 首先, 判定簇规模的标准应基于其包含的原始客户端数量, 而非代理模型的数量; 其次, 部分数据分布独特的良性模型可能因非恶意原因表现异常, 草率地将其排除会损害全局模型的性能与收敛效率. 为此, 本方法提出一种基于证据累积的跨类别可疑度评分机制, 以实现更精细的异常判定. 首先, 方法通过原始模型索引 I_p 确定小簇, 若一个代理模型落入基于原始客户端数量定义的小簇, 其对应的可疑度分数便加 1. 遍历所有数据类别后, 每个代理模型都将获得一个累积可疑度分数 s_i, 该分数汇总了其在所有类别分析中的表现 (算法 2 第 12–19 行). 最后, 系统将所有模型累积可疑度的平均值动态地设定为阈值 η. 任何分数超过 η 的代理模型将被最终判定为恶意并予以剔除, 余下的则被视为良性模型进入下一轮聚合 (算法 2 第 20–25 行). 值得注意的是, FedProbe 遵循安全优先的防御原则. 在 non-IID 的良性环境下, 直接剔除大于阈值的模型虽然可能导致极少数数据分布极端的良性客户端被排除, 但在面对高破坏性的投毒更新时, 这种策略是确保全局模型完整性的必要权衡. 这种“证据累积”的决策范式, 避免了因模型在单一类别分析中表现异常而被立即错判, 通过全面的跨类别评估显著降低了对良性模型的误伤率.

4 实验评估

为全面评估 FedProbe 性能, 本节展开详细的实验验证. 实验平台配置为 Intel i7-13700KF CPU, 64 GB DDR5 RAM 以及 NVIDIA GeForce RTX 4070Ti GPU. 所有实验均基于 PyTorch[35]框架实现.

4.1 实验设置

● 数据集以及数据划分. 为保证实验结果的可复现性与可比性, 实验采用 CIFAR-10[36]、MNIST[37]及 Fashion-MNIST[38]这 3 个公开基准数据集. 在服务器端, 我们预留了一个由 30 个样本构成的可信样本集 (每个类别包含 3 个样本), 该验证集在整个联邦训练周期内保持固定, 以提供一致的评估基准. 为模拟联邦学习中普遍存在的 non-IID 场景, 我们采用 $\alpha = 0.5$ 的 Dirichlet 分布进行数据划分.

● 模型和训练参数. 针对图像分类任务, 实验采用两种卷积神经网络 (convolutional neural network, CNN)[39]结构: 一个 5 层 CNN 用于 MNIST 和 Fashion-MNIST 数据集, VGG-9 网络则用于 CIFAR-10. 联邦学习的全局通信轮次设为 50 轮.

● 攻击设置. 实验主要考虑两类威胁模型: 常规投毒攻击与自适应攻击. 基于此, 实验设定了以下 3 种具体攻击方式.

(1) 无目标攻击. 实验采用 IPM (inner product manipulation)[20]攻击作为无目标攻击的实现. 在 IPM 攻击中, 攻击者旨在操纵其上传的模型更新, 使其与其余良性更新的聚合方向呈负相关, 从而阻碍全局模型的收敛或降低其最终性能.

(2) 后门攻击. 本文复现了 Chu 等人[30]使用的像素后门攻击. 在 CIFAR-10 数据集中, 通过修改“猫”类别图像的特定 12 个像素, 植入一个将目标误导向“狗”的后门; 在 MNIST 和 Fashion-MNIST 数据集中, 则通过修改第 3 类别图像的 12 个像素, 植入针对第 9 类别的后门. 此外, 我们还应用 constrain-and-scale[30]技术来增强后门攻击的隐蔽性与有效性.

(3) 自适应攻击. 本文采用了 Zhuang 等人[24]近期提出的基于层重要性的自适应攻击. 该方法在每轮本地训练中动态评估各层参数对模型性能的影响, 并据此调整投毒策略, 以实现更具针对性和隐蔽性的攻击.

此外, 若无特别说明, 实验均在以下默认配置下进行: 总客户端数量为 100, 每轮通信中随机选择 30% 的客户端参与本地训练, 其中恶意客户端占被选中客户端总数的 40%. 为保证实验的严格可复现性, 本研究对本地模型训

练时的随机种子进行控制, 使用的种子为 2048. 自适应攻击和像素后门攻击等实验均在此随机种子序列下进行 3 次独立运行, 实验中所报告的平均值及误差棒均基于这 3 次实验的结果计算得出.

● 超参数设置. 各客户端本地训练的超参数设置为: 本地训练轮次为 10, batch size 为 64, 学习率为 0.1. 此外, 对于检测算法, 本次实验将聚类的轮廓系数搜索范围 $[K_{\min}, K_{\max}]$ 设置为 (5, 10), 计算 KL 散度所需要的 *Softmax* 函数温度 τ 设置为 5.

● 评估指标. 实验采用以下指标来评估所提方法的性能与防御效果.

(1) 任务准确度 (main task accuracy, MTA). 该指标主要用于评估模型在常规投毒攻击下的良性任务性能, 定义为全局模型在可信样本集上的分类准确率. 其计算公式为: $Accuracy = \dfrac{TP+TN}{TP+TN+FP+FN}$. 其中, TP、TN、FP 和 FN 分别代表真阳性、真阴性、假阳性和假阴性的样本数量.

(2) 收敛速度. 针对 IPM 无目标攻击, 我们根据 Chu 等人[30]计算的全局模型的收敛速度进行评估, 即计算最后通信轮次本方法所达到的精度与其他方法所达到的这个精度的轮次之比.

(3) 任务精确率 (*Precision*) 和召回率 (*Recall*). 为了更细致地评估本方法在自适应攻击场景下的性能, 引入任务精确率和召回率. 公式分别为 $Precision = \dfrac{TP}{TP+FP}$, $Recall = \dfrac{TP}{TP+FN}$.

(4) 攻击成功率 (attack success rate, ASR). 针对后门攻击和自适应攻击, 其定义为全局模型在面对携带触发器的测试样本时, 将其错误分类为攻击者预设目标类别的比率.

4.2 性能评估

本节主要测试不同场景下 FedProbe 的性能, 直观的输出可视化以及消融研究.

(1) 样本数量选择

在评估开始前, 有必要对 FedProbe 服务器中可信样本集的数量对性能的影响进行测试. 首先, 在实际实践中, 要求服务器持有大量可信样本集显然不符合数据最小化原则. 此外, SHAP 值的计算复杂度较高, 计算开销与样本数量呈线性正相关. 因此, 本节选择在服务器端每个数据集的每 1 类分别存储 1、3、5 个样本, 通过攻击成功率 (ASR) 和任务准确度 (MTA) 对性能进行评估, 如图 3 所示, 其中柱状图为 ASR, 折线图为 MTA.

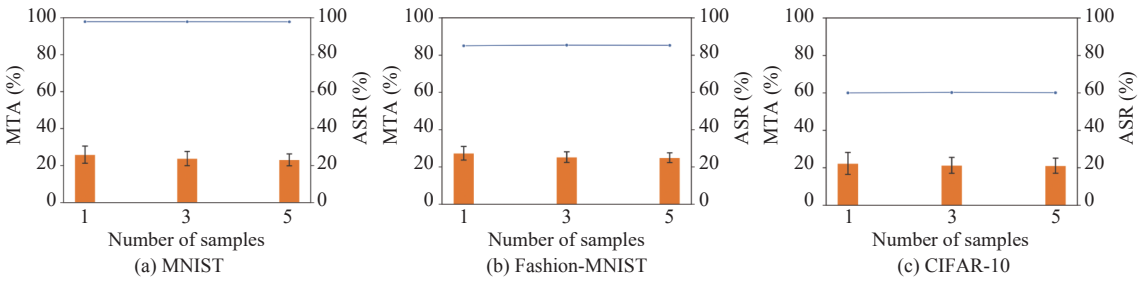
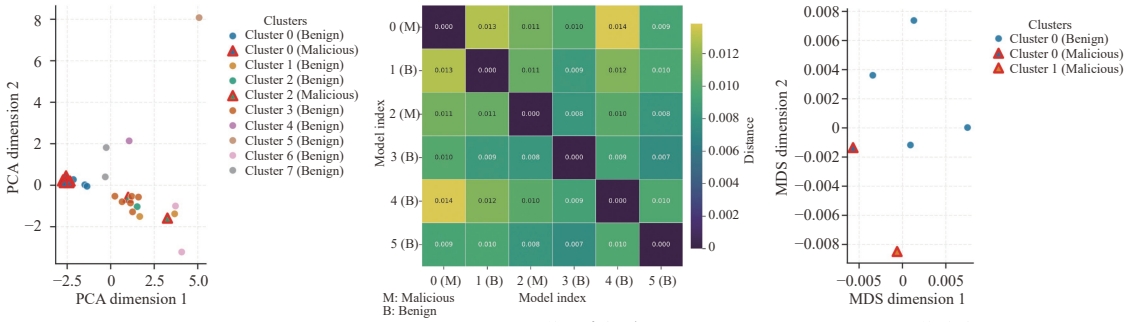


图 3 不同可信样本数量抵抗后门攻击时的性能

从图 3 中可以看到, 在不同的可信样本数量上, 各个数据集的攻击成功率和准确度几乎是相同的. 然而, 为了更深入地分析其影响, 我们在柱状图中引入了误差棒. 结果显示, 当样本数量为 1 时, CIFAR-10 数据集上的检测结果表现出明显的高方差. 这意味着虽然平均 ASR 尚可, 但在单次检测中极易受到随机噪声的干扰, 导致防御性能不稳定. 相比之下, 当样本数增加至 3 时, 误差棒收窄, 检测波动降低. 这表明, 虽然样本数量对平均防御成功率的边际影响较小, 但对于防御的稳定性较为重要. 因此, 为了平衡检测效率和检测的稳定性, 最终方法选择使用每类 3 个样本进行后续实验评估.

(2) 输出可视化

为直观展示本方法的有效性, 图 4 对防御机制在特定攻击场景下的各关键阶段输出进行了可视化. 该可视化实验的设定如下: 数据集为 CIFAR-10 (non-IID), 攻击类型为像素后门攻击, 恶意客户端占比为 40%.



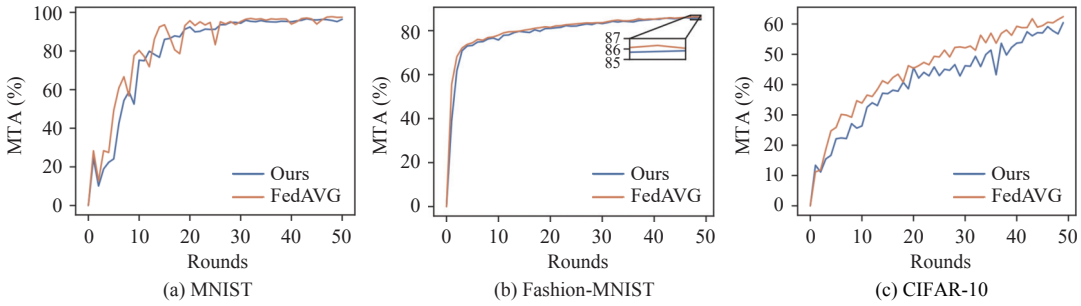
(a) 代理模型聚类 (b) SHAP 值距离矩阵 (c) SHAP 值聚类

图 4 方法各阶段在像素后门攻击下的可视化 (聚类结果经过 PCA 与 MDS 降维处理)

从图 4(a) 的初始聚类结果可见, 在通过 KL 散度完成模型行为相似性度量后, 大部分恶意模型被成功聚集到同一簇 (Cluster 0). 然而, 由于数据异质性及部分攻击的隐蔽性, 仍有少数中毒模型与良性模型混杂在同一簇中 (如 Cluster 2). 随后, 各簇被聚合成代理模型并进入 SHAP 值分析阶段. 鉴于本次攻击旨在将带有触发器类别 2 的样本误导至类别 8, 本文在此重点考察针对类别 2 的 SHAP 值归因模式. 图 4(b) 的热度图展示了各代理模型在该类别上的 SHAP 值距离矩阵. 可以清晰地看到, 由恶意簇 (Cluster 0) 生成的代理模型 0 与其他部分模型的 SHAP 值距离显著偏高, 表现出异常. 相比之下, 最初混入良性簇 (Cluster 2) 的中毒模型, 其生成的代理模型 2 则因聚合效应而表现正常. 因此, 在图 4(c) 的最终 SHAP 值聚类中, 行为异常的代理模型 0 被轻易识别并隔离. 而对于初步漏检的中毒模型, 由于其恶意影响在代理模型聚合阶段已被簇内大量的良性模型中和, 其最终对全局模型的威胁也得到了有效控制, 从而显著抑制了后门攻击的成功率.

(3) 无攻击场景下方法性能

本节旨在评估在无攻击的良性场景下, FedProbe 对联邦学习收敛性的影响. 图 5 展示了 FedProbe 与标准 FedAVG 在 3 个数据集上的收敛速度对比.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

图 5 方法在无攻击场景下与 FedAVG 的收敛速度对比

实验结果表明, 在无攻击场景下, FedProbe 的收敛速度与 FedAVG 基线相当, 仅引入了轻微的性能开销. 在 MNIST 数据集上, 为达到 96.43% 的准确率, FedProbe 需 50 个通信轮次, 而 FedAVG 仅需 41 个通信轮次, 所需轮次约为 FedAVG 的 1.21 倍; 在 Fashion-MNIST 数据集上, 为达到 85.83% 的准确率, FedProbe 所需轮次约为 FedAVG 的 1.11 倍; 在更为复杂的 CIFAR-10 数据集上, 为达到 60.44% 的准确率, 所需轮次约为 FedAVG 的 1.19 倍.

此外, 从图 5(a) 和 (c) 中可以观察到, FedProbe 的准确率曲线相较于 FedAVG 存在更明显的波动. 我们推测, 这是由于防御机制在无攻击时, 仍可能将部分数据分布极为独特的良性客户端识别为暂时性异常. 在 non-IID 设定下, 这种筛选机制不可避免地会带来轻微的准确率波动与收敛速度下降, 为了验证猜想, 我们选择了在中间轮次时各个代理模型的可疑度分数, 如图 6 所示.

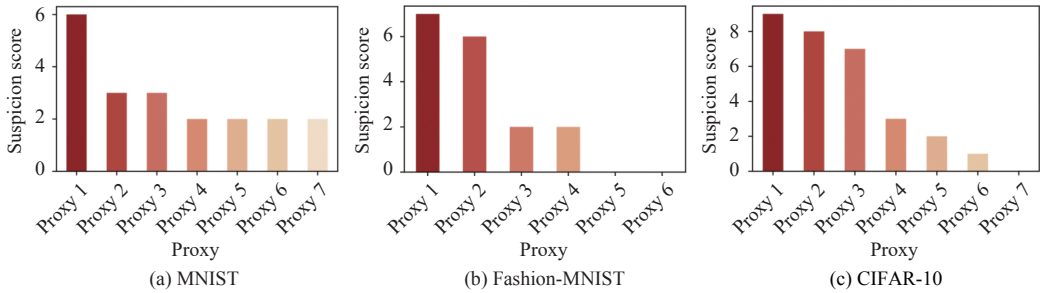


图 6　方法在无攻击场景下第 25 轮次的可疑度分数

从图 6 中看到, 在良性环境中, 3 个数据集的可疑度分数呈现出类似于长尾分布的特征: 大多数代理模型的可疑度分数较低, 但存在部分表现出暂时性异常的极值. 在这种分布下, 计算得出的平均值阈值会被右侧的极值显著拉高, 发生向右偏移. 因此, 算法实际上仅过滤了少数位于长尾端的离群代理模型. 此外, FedProbe 在最后实际上过滤的为聚合成代理模型的原始模型. 为了更进一步验证, 在上述完全良性的环境下, 本文在 3 个数据集上测试了每一轮所误检的模型数量, 如图 7 所示.

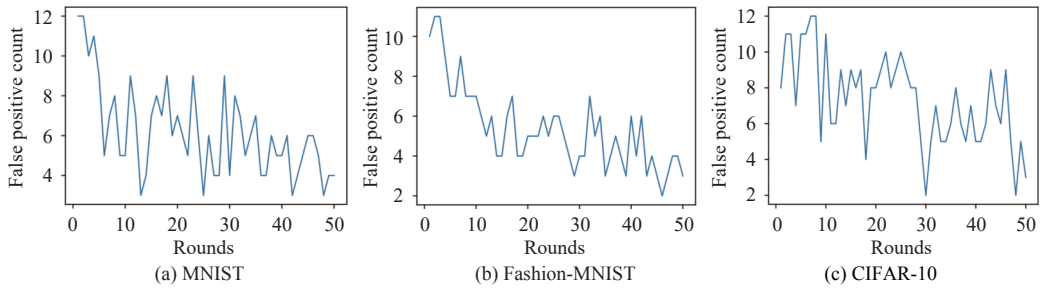


图 7　方法在无攻击场景下每轮误检的模型数量

图 7 展示了 FedProbe 在无攻击环境下每轮的误过滤模型数量. 观察可知, 实际的剔除数量随模型收敛迅速下降至极低水平, 这一现象在收敛速度较快的 MNIST 和 Fashion-MNIST 数据集上尤为显著. 这表明前期被过滤的模型是训练初期不稳定的低质量模型, 因此, 配合跨类别可疑度评估的证据累积机制, FedProbe 在有效剔除恶意离群点的同时, 并未对特定的边缘模型造成系统性的覆盖缺失, 保障了异构环境下的特征多样性. 值得注意的是, 上述结果基于良性环境下的评估; 而在存在攻击的场景中, 植入后门或篡改参数的模型将表现出更高的可疑度分数, 从而进一步降低针对良性模型的误检率.

(4) 检测计算代价

为评估本方法在检测阶段的计算开销, 表 1 记录了其各关键步骤的耗时. 此外, 我们额外在 CIFAR-10 数据集上增加了 ResNet-18 模型, 以便测试方法对较深层模型的处理时间.

表 1　方法检测时间开销

模型	本地训练 (s)	代理模型生成 (s)	SHAP值分析 (s)	SHAP值相似度计算 (s)	聚类 (ms)	总时间 (s)
CNN	1.2	10.12	6.33	2.24	9.34	21.89
VGG-9	2.1	10.72	9.46	3.25	9.23	26.53
ResNet-18	2.8	11.52	12.33	3.65	9.31	30.30

注: SHAP值分析以及SHAP值相似度计算为处理单个代理模型的耗时

从表 1 可以看出, FedProbe 的时间开销主要集中在 SHAP 值分析与 SHAP 值相似度计算阶段. 以 VGG-9 模型为例, 这两个步骤的耗时合计约为 12.71 s. 其主要原因是, 为实现精细化的异常检测, FedProbe 需要针对每个数据类别独立进行 SHAP 值分析, 导致计算复杂度与类别总数成正比, 从而增加了时间开销. 代理模型的生成是次要

的时间开销来源, 其耗时同样与类别数量相关.

考虑到 FedProbe 采用轮廓系数动态生成的代理模型数量上限设为 10, 可以估算出在最坏情况下, VGG-9、CNN 和 ResNet-18 的单轮 SHAP 检测总耗时分别约为 127.1 s、85.73 s 和 159.5 s. 尽管代理模型引导机制已大幅降低了原本分析所有客户端所需的高昂计算量, 但逐类别进行深度检测的策略, 决定了数据集的类别总数是影响 FedProbe 时间效率的关键因素.

(5) 消融研究

为深入剖析本框架中两阶段防御机制 (代理模型引导与 SHAP 值分析) 各自的贡献, 本节进行了一项全面的消融研究. 该研究针对 IPM 与后门两类典型攻击, 系统性地评估了 3 种防御配置的性能: 仅代理模型聚类 (Proxy)、仅 SHAP 值分析 (SHAP) 以及 FedProbe 完整方法 (Ours). 图 8 展示了此次消融研究的核心结果.

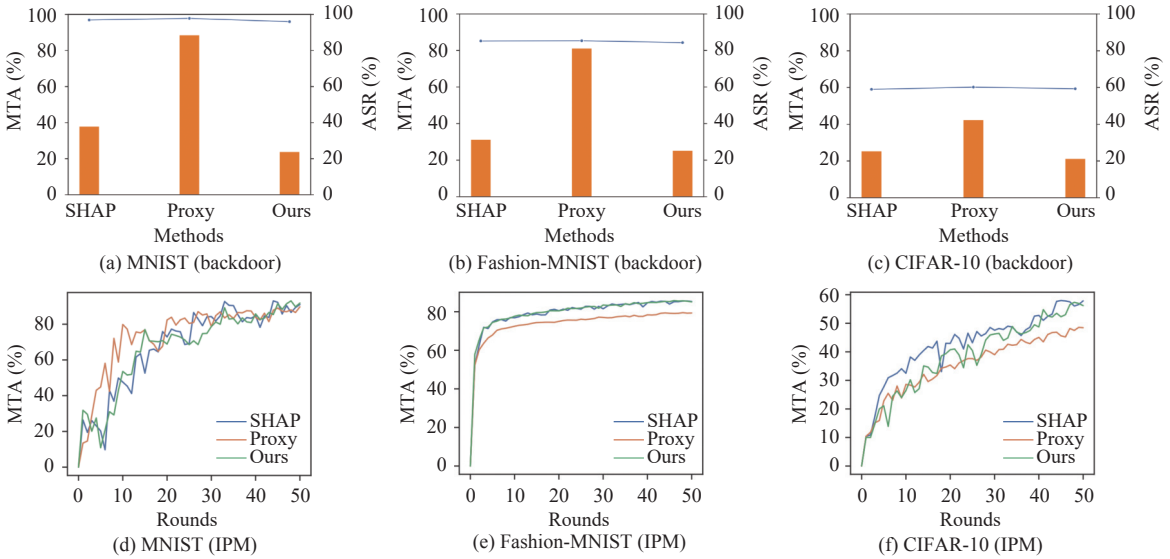


图 8 不同防御组件在后门攻击与 IPM 攻击下的性能对比

从图 8(a)–(c) 的结果来看, 仅依赖代理模型聚类的防御策略几乎无法抵御后门攻击, 其攻击成功率 (ASR) 接近未设防状态. 此结果验证了基于 KL 散度的宏观行为度量难以察觉后门攻击的细微伪装. 相比之下, 仅 SHAP 值分析与 FedProbe 均能将 ASR 降至极低水平, 证明了基于 SHAP 的微观行为分析是防御此类隐蔽攻击的核心.

在图 8(d)–(f) 所展示的 IPM 攻击场景下, 仅代理模型聚类策略同样表现不佳, 模型收敛性受到严重破坏. 而仅 SHAP 值分析与完整方法均能有效抵御攻击, 其收敛曲线与良性场景接近. 然而还观察到, 在 MNIST 和 CIFAR10-DVS 数据集上, 这两种方法的准确率曲线出现了比无攻击时更显著的波动. 我们推测, 这是因为在防御过程中, 对恶意模型的正确排除或对良性模型的偶然误判, 都可能导致全局模型在特定轮次的准确率瞬时下降. 这种波动在收敛速度较快的数据集上 (如 MNIST) 表现得尤为敏感.

综合来看消融研究证实了 FedProbe 两个阶段设计的必要性与互补性: 第 1 阶段的代理模型引导机制对于提升效率至关重要, 而第 2 阶段的 SHAP 分析则是确保高强度安全防御的核心. 因此, 本文所提出的两阶段框架, 通过有机地结合这两种不同维度的分析手段, 最终构建了一个更具广谱性与鲁棒性的防御体系.

4.3 对比实验

为进一步评估 FedProbe 的防御性能, 本节将其与多种经典的及先进的防御方法, 在 IPM 无目标攻击、像素后门攻击和自适应攻击场景下进行了系统性对比.

(1) IPM 攻击对比实验

本节首先评估各方法在抵御 IPM 攻击时的性能. 对比方法涵盖了基于距离度量的 Krum[27], 基于统计排序的

Median[25]和 Trimmed Mean[25], 以及先进的 FLTrust[11]和 FLAME[15].

如图 9 所示, 在本次实验设定的攻击强度下, Krum 与 Median 等传统方法基本失效, 模型准确率无法收敛. 这证明了依赖简单距离或值域排序的防御机制, 难以抵御旨在颠覆模型收敛方向的复杂模型中毒攻击. 基于统计的 Trimmed Mean 方法在相对简单的 MNIST 与 Fashion-MNIST 任务上展现了一定的防御能力, 但当任务复杂度上升至 CIFAR-10 时, 其性能也出现显著下滑, 暴露了纯统计方法在处理高维复杂数据分布时的局限性.

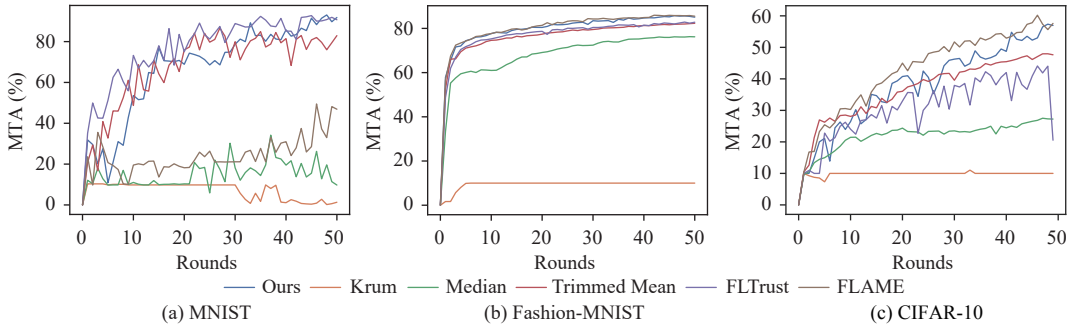


图 9　不同防御方法在 IPM 攻击下的收敛性能对比

在与更先进方法的对比中, 我们观察到现有防御方案对特定场景的特化现象, 这反而凸显出本方法在不同场景下的泛化优势. 例如, FLAME 在特征复杂的 Fashion-MNIST 与 CIFAR-10 上表现优异, 其性能与 FedProbe 不相上下. 然而, 该方法在看似最简单的 MNIST 任务上却意外失效. 我们推测, IPM 攻击在该数据集上可能诱导了一种独特的参数分散化模式, 恰好规避了 FLAME 基于参数相似性和裁剪聚类的检测逻辑. 与之相似, FLTrust 在 MNIST 与 Fashion-MNIST 上表现稳健, 但在 CIFAR-10 的高维复杂环境下, 其防御能力也显著下降.

相比之下, FedProbe 是唯一在所有 3 个数据集上均保持高水平防御性能收敛性的方法. 这一结果有力地证明了本框架的优越性: 通过将宏观的代理模型聚类分析引导微观的 SHAP 值行为洞察, FedProbe 不依赖于攻击在某一特定空间中必然呈现特定模式的单一假设, 从而构建了一个更具韧性和泛化能力的防御体系, 能够精准识别不同数据环境下由 IPM 攻击引发的各类异常.

(2) 像素后门攻击对比实验

本节旨在评估各防御方法在不同比例的恶意客户端参与下, 抵御像素后门攻击时的性能. 如图 10 所示, 实验结果同时展示了主任务准确率 (折线图) 与攻击成功率 (柱状图).

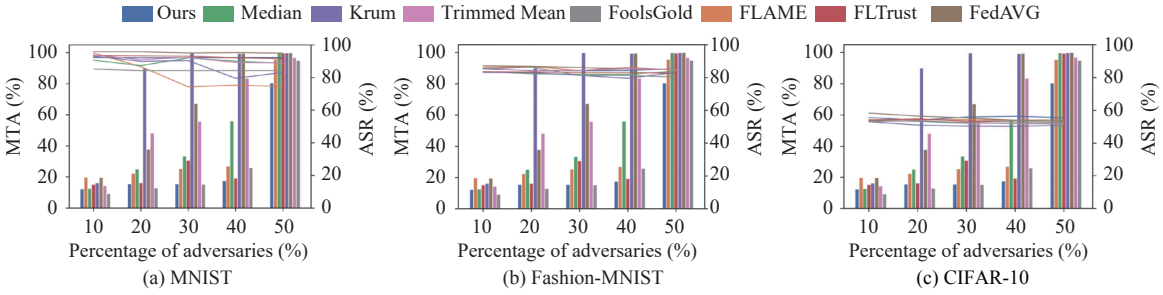


图 10　不同防御方法在像素后门攻击下的对比

实验结果显示, 当恶意客户端占比较低时, 后门攻击的威胁有限, 即便是未设防的 FedAVG, 其 ASR 也普遍低于 20%. 然而, 当该比例提升至 30% 以上时, Krum、Median 与 Trimmed Mean 等传统方法的防御便已崩溃, 其 ASR 在所有数据集上均急剧攀升. 在 40% 及以上的高比例攻击下, 这些传统方法已完全失效.

此外, 在与先进方法的对比中, FedProbe 展示了卓越的泛化性和稳定性. 例如, 在 MNIST 数据集上 (图 10(a)), 尽管 FLAME 在低比例攻击下表现良好, 但当恶意客户端占比超过 30% 后, 其主任务准确度便急剧下降至 80% 以

下. 而 FoolsGold 在所有占比下的任务准确度均低于 90%, 证明其防御机制牺牲了过多的模型效用. 在更为复杂的 CIFAR-10 数据集上 (图 10(c)), FLTrust 的主任务准确度也随着攻击比例的升高而出现明显下滑.

相比之下, FedProbe 是唯一在所有数据集下, 始终将 ASR 抑制在较低水平, 同时保持最稳定主任务准确度的防御策略. 即便在 50% 的极端恶意环境下, 本方法的任务准确度依然表现稳健, 充分证明了其在抵御高强度后门攻击时的优越性与鲁棒性.

(3) 自适应攻击对比实验

为评估 FedProbe 在最前沿自适应攻击下的防御能力, 本节采用基于层的自适应攻击 (LP attack). 鉴于传统防御方法已在先前实验中证实无效, 本节的对比对象聚焦于 FLAME、FLTrust、FLARE[40]及 FLIP[41]等先进的防御框架. 所有实验均在 non-IID 的 CIFAR-10 数据集上, 遭受连续 20 个通信轮次的不间断攻击, 各方法的综合性能由表 2 呈现.

表 2　不同防御方法在 LP attack 下的综合性能对比 (%)

防御框架	攻击成功率 (ASR)		任务准确度	精确率	召回率
FedAVG	Highest ASR	97.61 ± 2.23	39.81 ± 4.33	37.93 ± 5.22	39.81 ± 4.33
	Avg ASR	67.88 ± 7.25			
	Lowest ASR	48.20 ± 10.05			
FLAME	Highest ASR	52.08 ± 8.23	41.53 ± 1.55	41.74 ± 2.01	41.53 ± 1.55
	Avg ASR	26.88 ± 6.34			
	Lowest ASR	16.01 ± 5.88			
FLTrust	Highest ASR	48.78 ± 10.25	33.75 ± 1.37	33.86 ± 1.33	33.75 ± 1.37
	Avg ASR	32.22 ± 9.57			
	Lowest ASR	17.22 ± 5.39			
FLARE	Highest ASR	41.01 ± 4.88	43.03 ± 1.05	42.02 ± 1.25	43.03 ± 1.05
	Avg ASR	23.39 ± 7.65			
	Lowest ASR	13.39 ± 2.80			
FLIP	Highest ASR	39.82 ± 4.63	41.51 ± 2.22	**41.42 ± 1.88**	41.51 ± 2.22
	Avg ASR	**23.63 ± 3.76**			
	Lowest ASR	**11.33 ± 1.77**			
FedProbe	Highest ASR	**38.42 ± 5.33**	**43.18 ± 1.76**	**41.11 ± 1.22**	43.18 ± 1.76
	Avg ASR	**24.52 ± 3.85**			
	Lowest ASR	**12.83 ± 2.58**			

注: 在本实验的多类别设定下, 召回率采用加权宏平均计算, 其数值与任务准确度相同; 加粗数据表示最优结果

表 2 实验结果清晰地揭示了 FedProbe 在抵御自适应攻击时的综合性能. 首先, 在防御鲁棒性维度, FedProbe 在应对最极端攻击时表现突出. 以衡量最不利场景下防御能力的最高攻击成功率为核心指标, 本方法的数值仅为 38.42%, 在所有对比方法中表现最佳. 相较之下, FLAME 和 FLTrust 的最高 ASR 分别高达 52.08% 和 48.78%. 特别是 FLTrust, 其高达 10.25% 的标准差也反映了其防御性能的不稳定性. 值得一提的是, 尽管 FLIP 方法在平均攻击成功率和最低攻击成功率上略有优势, 但本方法在更关键的最高攻击成功率指标上表现更低, 证明了其在极限威胁下的防御能力更为可靠.

其次, 在防御与性能的权衡中, FedProbe 展现出更优的平衡性. FLTrust 导致模型准确率骤降至 33.75%, 是所有方法中的最低值, 体现了其高昂的性能成本. 反观本方法, 在提供稳健防御的同时, 其准确率与召回率均达到 43.18%, 位列所有对比方法之首, 也优于 FLIP 的 41.51%. 这一结果表明, FedProbe 所采用的防御策略能够精准识别并过滤恶意更新, 同时最大限度地保留良性更新, 从而在有效抵御攻击的同时维持全局模型的泛化能力.

LP attack 等自适应攻击往往针对参数空间检测构造约束, 例如限制更新幅度或选择性投毒关键层, 使得基于参数距离统计的判别变得困难. FedProbe 第 2 阶段使用 SHAP 特征归因分布作为证据: 在多个类别的可信样本上, 良性模型的归因模式具有相对一致性. 攻击者若希望规避检测, 需要在保持攻击效果的同时, 还要使跨类别归因分

布与良性高度一致, 则必须弱化攻击特征与目标标签的关联, 这将导致其 ASR 断崖式下降. 这种归因一致性约束让自适应攻击完全规避该类归因证据具有较高难度.

5 讨 论

本节我们根据上述对 FedProbe 的解析和验证, 对其方法局限性和效率提升手段进行讨论.

5.1 方法局限性

尽管 FedProbe 在抵御无目标攻击、像素后门攻击及自适应后门攻击方面已通过实验验证了其鲁棒性, 但仍存在以下 3 个方面的局限性.

首先, 本方法的有效性依赖于一个核心先验假设: 即可解释性方法 (SHAP) 能够有效区分中毒模型与良性模型的行为差异. 虽然本文及先前的工作已通过大量实验证实了此假设的经验有效性, 但其背后仍缺乏坚实的理论保证. 因此, 为基于 SHAP 值的检测机制建立理论基础是未来一个重要的研究方向.

其次, FedProbe 的设计聚焦于在服务器端识别恶意模型更新, 因而未能有效抵御针对客户端的隐私攻击, 例如数据重建攻击或成员推理攻击. 作为潜在的解决方案, 在客户端与服务器端引入可信执行环境[42]等硬件加密技术, 能够在不根本改变本方法核心流程的前提下, 增强本地模型的隐私安全.

此外, 受 FLTrust 机制启发, 本文方案假设服务器端预置了一个小规模但分布可信的验证数据集, 旨在为恶意攻击检测提供行为度量基准与可解释性归因分析的支持. 在实际部署场景中, 该验证集的获取应遵循数据合规原则, 主要途径包括: 1) 任务发起方或服务提供方依法持有的少量业务样本; 2) 公开数据集或公共领域数据的代表性子集; 3) 系统初始化阶段, 在严格履行知情同意前提下由志愿者提供种子数据. 针对高隐私敏感场景, 建议将基于该验证集的检测逻辑部署于可信执行环境内部, 以最小化验证数据的隐私泄露风险.

值得注意的是, KL 散度与 SHAP 值归因的有效性高度依赖于验证数据集的分布代表性. 若验证集规模过小或存在严重的类分布不平衡, 将导致两方面的系统性偏差: 一方面, KL 行为距离的估计方差显著增大, 导致客户端的行为矩阵对采样噪声更敏感; 另一方面, SHAP 归因在不同采样批次下的稳定性降低. 因此, 本研究在实践中对大规模数据集建议采用分层抽样策略以保障特征覆盖率, 并在可疑分数计算阈值的设计中引入容错因子以抑制统计噪声. 关于验证集规模与分布对检测性能的敏感性分析, 将作为未来的研究工作进一步探讨.

最后, FedProbe 的有效性主要在图像分类任务上得到了显著验证, 但将其推广至其他领域, 例如目标检测乃至大语言模型场景, 仍面临挑战. 其根本原因在于, SHAP 分析的实现与具体任务紧密相关, 需要针对不同任务的特性调整其核心算法. 因此, 开发能够自适应不同任务场景的 SHAP 分析流程, 将是提升本方法泛用性的关键.

5.2 效率提升手段

FedProbe 在检测效率上仍有提升空间. 实验结果表明, 对单个本地模型进行 SHAP 值分析, 在 CNN 和 VGG-9 上分别需要近 7 s 和 10 s. 在拥有数百个客户端的大规模联邦学习场景中, 若无优化, 服务器每轮的检测时间将超过 10 min, 这会严重制约整体训练效率.

虽然本研究引入的代理模型引导机制已显著降低了检测开销, 但在客户端规模极为庞大且数据高度异构的场景下, 其有效性面临新的挑战: 如何精确地衡量数据异构性并确定最优的聚类数量, 成为一个关键难题.

为进一步优化效率, 可从两个方面探索解决方案. 其一是在服务器端, 通过并行计算或利用 GPU 进行硬件加速, 可直接缩短 SHAP 分析的耗时. 其二是探索分布式计算方案, 即将部分检测逻辑 (如 SHAP 值分析) 迁移至客户端的可信执行环境中执行. 在该模式下, 繁重的计算任务在本地完成, 服务器仅需分析各个客户端返回的计算结果, 从而将计算压力从中心服务器有效分摊[43]. 然而, 后一种方案高度依赖客户端的硬件性能与安全保障, 使其在物联网或资源受限移动设备上的应用面临现实挑战.

6 结 论

本文提出了一种代理模型引导的可解释高效联邦学习投毒攻击防御框架 FedProbe. 首先, 该框架利用服务器

端的一个可信验证集, 通过计算各客户端模型更新间的 KL 散度进行聚类, 并为每个聚类构建代理模型. 此代理模型引导机制旨在克服传统 SHAP 值分析所面临的效率瓶颈, 同时提升在 non-IID 数据环境下的检测稳定性. 随后, 框架通过对代理模型进行 SHAP 值分析, 并结合一种跨类别可疑度分数计算方法来识别恶意模型. 这种基于可解释性的分析能够深入洞察模型的内部决策逻辑, 从而有效识别并抵御多样化的投毒攻击. 我们在 MNIST、Fashion-MNIST 和 CIFAR-10 这 3 个基准数据集上对所提框架进行了广泛评估. 实验结果表明, 与其他先进的防御策略相比, FedProbe 在多种攻击场景下均展现出更高的任务准确度稳定性与更低的攻击成功率. 最后, 本文亦探讨了 FedProbe 的潜在局限与未来效率优化的途径, 为该领域的研究工作指明了方向. 在未来的研究工作中, 本研究将对其目前局限性进行改进, 例如引入可信执行环境增加更高的隐私防护以及考虑针对拥有数万个边缘节点的超大规模联邦网络下对投毒攻击的高效防御, 进一步降低通信与计算延迟.

References

[1] Imteaj A, Thakker U, Wang SQ, Li J, Amini MH. A survey on federated learning for resource-constrained IoT devices. IEEE Internet of Things Journal, 2022, 9(1): 1–24. [doi: 10.1109/JIOT.2021.3095077]

[2] Chen JX, Yan H, Liu ZY, Zhang M, Xiong H, Yu S. When federated learning meets privacy-preserving computation. ACM Computing Surveys, 2024, 56(12): 319. [doi: 10.1145/3679013]

[3] Liu L, Wang Y, Liu GY, Peng K, Wang C. Membership inference attacks against machine learning models via prediction sensitivity. IEEE Trans. on Dependable and Secure Computing, 2023, 20(3): 2341–2347. [doi: 10.1109/TDSC.2022.3180828]

[4] Hu K, Gong S, Zhang Q, Seng C, Xia M, Jiang SS. An overview of implementing security and privacy in federated learning. Artificial Intelligence Review, 2024, 57(8): 204. [doi: 10.1007/s10462-024-10846-8]

[5] Bagdasaryan E, Veit A, Hua YQ, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 2938–2947.

[6] Luisier F, Blu T, Unser M. Image denoising in mixed poisson—Gaussian noise. IEEE Trans. on Image Processing, 2011, 20(3): 696–708. [doi: 10.1109/TIP.2010.2073477]

[7] Kumari K, Rieger P, Fereidooni H, Jadliwala M, Sadeghi AR. BayBFed: Bayesian backdoor defense for federated learning. In: Proc. of the 2023 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2023. 737–754. [doi: 10.1109/SP46215.2023.10179362]

[8] Li WJ, Fan K, Zhang JY, Li H, Lim WYB, Yang Q. Enhancing security and privacy in federated learning using low-dimensional update representation and proximity-based defense. arXiv:2405.18802, 2024.

[9] Li SH, Ngai ECH, Voigt T. An experimental study of Byzantine-robust aggregation schemes in federated learning. IEEE Trans. on Big Data, 2024, 10(6): 975–988. [doi: 10.1109/TBDATA.2023.3237397]

[10] Zhu CY, Roos S, Chen LY. LeadFL: Client self-defense against model poisoning in federated learning. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 1818.

[11] Cao XY, Fang MH, Liu J, Gong NZQ. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In: Proc. of the 28th Annual Network and Distributed System Security Symp. The Internet Society, 2021.

[12] Sandeepa C, Siniarski B, Wang S, Liyanage M. SHERPA: Explainable robust algorithms for privacy-preserved federated learning in future networks to defend against data poisoning attacks. In: Proc. of the 2024 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2024. 4772–4790. [doi: 10.1109/SP54263.2024.00271]

[13] Mishra P, Lehmkuhl R, Srinivasan A, Zheng WT, Popa RA. DELPHI: A cryptographic inference service for neural networks. In: Proc. of the 29th USENIX Security Symp. USENIX Association, 2020. 2505–2522.

[14] Ma ZR, Ma JF, Miao YB, Li YJ, Deng RH. ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning. IEEE Trans. on Information Forensics and Security, 2022, 17: 1639–1654. [doi: 10.1109/TIFS.2022.3169918]

[15] Nguyen TD, Rieger P, Chen HL, Yalame H, Möllering H, Fereidooni H, Marchal S, Miettinen M, Mirhoseini A, Zeitouni S, Koushanfar F, Sadeghi AR, Schneider T. FLAME: Taming backdoors in federated learning. In: Proc. of the 31st USENIX Security Symp. Boston: USENIX Association, 2022. 1415–1432.

[16] Nguyen TD, Rieger P, Yalame H, Möllering H, Fereidooni H, Marchal S, Miettinen M, Mirhoseini A, Sadeghi AR, Schneider T, Zeitouni S. FLGUARD: Secure and private federated learning. IACR Cryptology ePrint Archive, 2021, 2021: 25.

[17] Mosca E, Szigeti F, Tragianni S, Gallagher D, Groh G. SHAP-based explanation methods: A review for NLP interpretability. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: Int'l Committee on Computational Linguistics, 2022. 4593–4603.

[18] Gou JP, Yu BS, Maybank SJ, Tao DC. Knowledge distillation: A survey. Int'l Journal of Computer Vision, 2021, 129(6): 1789–1819.

[doi: 10.1007/s11263-021-01453-z]

[19] Data D, Diggavi SN. Byzantine-resilient high-dimensional federated learning. IEEE Trans. on Information Theory, 2023, 69(10): 6639–6670. [doi: 10.1109/TIT.2023.3284427]

[20] Xie C, Koyejo O, Gupta I. Fall of empires: Breaking byzantine-tolerant SGD by inner product manipulation. In: Proc. of the 35th Uncertainty in Artificial Intelligence Conf. PMLR, 2020. 261–270.

[21] Wang YK, Zhai DH, He YP, Xia YQ. An adaptive robust defending algorithm against backdoor attacks in federated learning. Future Generation Computer Systems, 2023, 143: 118–131. [doi: 10.1016/j.future.2023.01.026]

[22] Xu J, Huang SL, Song LQ, Lan T. Byzantine-robust federated learning through collaborative malicious gradient filtering. In: Proc. of the 42nd IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). Bologna: IEEE, 2022. 1223–1235. [doi: 10.1109/ICDCS54860.2022.00120]

[23] Li SH, Ngai ECH, Ye FH, Ju L, Zhang TR, Voigt T. Blades: A unified benchmark suite for byzantine attacks and defenses in federated learning. In: Proc. of the 9th IEEE/ACM Int'l Conf. on Internet-of-Things Design and Implementation. Hong Kong: IEEE, 2024. 158–169. [doi: 10.1109/IoTDI61053.2024.00018]

[24] Zhuang HM, Yu MX, Wang H, Hua Y, Li J, Yuan X. Backdoor federated learning by poisoning backdoor-critical layers. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[25] Yin D, Chen YD, Ramchandran K, Bartlett P. Byzantine-robust distributed learning: Towards optimal statistical rates. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: JMLR.org, 2018. 5636–5645.

[26] Siegel AF. Robust regression using repeated medians. Biometrika, 1982, 69(1): 242–244. [doi: 10.21236/ada092660]

[27] Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine learning with adversaries: Byzantine tolerant gradient descent. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 118–128.

[28] Zhao LC, Jiang JL, Feng B, Wang Q, Shen C, Li Q. SEAR: Secure and efficient aggregation for Byzantine-robust federated learning. IEEE Trans. on Dependable and Secure Computing, 2022, 19(5): 3329–3342. [doi: 10.1109/TDSC.2021.3093711]

[29] Fereidooni H, Pegoraro A, Rieger P, Dmitrienko A, Sadeghi AR. FreqFed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning. In: Proc. of the 31st Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2024.

[30] Chu T, García-Recuero A, Iordanou C, Smaragdakis G, Laoutaris N. Securing federated sensitive topic classification against poisoning attacks. In: Proc. of the 30th Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2023.

[31] Zhu CC, Zhang JL, Sun XB, Chen B, Meng WZ. ADFL: Defending backdoor attacks in federated learning via adversarial distillation. Computers & Security, 2023, 132: 103366. [doi: 10.1016/j.cose.2023.103366]

[32] Fung C, Yoon CJM, Beschastnikh I. Mitigating sybils in federated learning poisoning. arXiv:1808.04866, 2018.

[33] Rieger P, Krauß T, Miettinen M, Dmitrienko A, Sadeghi AR. CrowdGuard: Federated backdoor detection in federated learning. In: Proc. of the 31st Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2024.

[34] Jiang WJ, Yang X, Li BX. Federated learning poisoning attack defense method based on interpretable contribution anomaly detection and dynamic pruning. Chinese Journal of Computers, 2025, 48(12): 2855–2874 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2025.02855]

[35] Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 721.

[36] Krizhevsky A. Learning multiple layers of features from tiny images [MS. Thesis]. University of Toronto, 2009.

[37] Deng L. The MNIST database of handwritten digit images for machine learning research [best of the Web]. IEEE Signal Processing Magazine, 2012, 29(6): 141–142. [doi: 10.1109/MSP.2012.2211477]

[38] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

[39] Alzubaidi L, Zhang JL, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 2021, 8(1): 53. [doi: 10.1186/s40537-021-00444-8]

[40] Wang N, Xiao Y, Chen YM, Hu Y, Lou WJ, Hou YT. FLARE: Defending federated learning against model poisoning attacks via latent space representations. In: Proc. of the 2022 ACM on Asia Conf. on Computer and Communications Security. Nagasaki: ACM, 2022. 946–958. [doi: 10.1145/3488932.3517395]

[41] Zhang KY, Tao GH, Xu QL, Cheng SY, An SW, Liu YQ, Feng SW, Shen GY, Chen PY, Ma SQ, Zhang XY. FLIP: A provable defense framework for backdoor mitigation in federated learning. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali:

OpenReview.net, 2023.

[42] Kuznetsov E, Chen YT, Zhao M. SecureFL: Privacy preserving federated learning with SGX and TrustZone. In: Proc. of the 2021 IEEE/ACM Symp. on Edge Computing (SEC). San Jose: IEEE, 2021. 55–67. [doi: 10.1145/3453142.3491287]

[43] Gu YH, Bai YB. Survey on security and privacy of federated learning models. Ruan Jian Xue Bao/Journal of Software, 2023, 34(6): 2833–2864 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6658.htm [doi: 10.13328/j.cnki.jos.006658]

附中文参考文献

[34] 蒋伟进, 杨璇, 李碧霞. 基于可解释贡献异常检测与动态裁剪的联邦学习投毒攻击防御方法. 计算机学报, 2025, 48(12): 2855–2874. [doi: 10.11897/SP.J.1016.2025.02855]

[43] 顾育豪, 白跃彬. 联邦学习模型安全与隐私研究进展. 软件学报, 2023, 34(6): 2833–2864. http://www.jos.org.cn/1000-9825/6658.htm [doi: 10.13328/j.cnki.jos.006658]

作者简介

曹益皓, 博士生, 主要研究领域为可信计算, 联邦学习, 信息安全.

张建标, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为可信计算, 网络安全, 区块链技术.

赵亚茹, 博士, 副教授, CCF 专业会员, 主要研究领域为网络安全, 联邦学习, 车联网.

韩宇飞, 博士生, 主要研究领域为可信计算, 信息安全, 云安全和隐私保护.

张承昱, 博士生, 主要研究领域为可信计算, 信息安全, 数据库安全.

叶涛, 博士, 教授, 主要研究领域为可信计算, 网络安全, 人工智能安全.