计算机科学
COMPUTER SCIENCE

横向联邦学习后门的多方共治防范策略

许文韬, 王斌君, 朱莉欣, 王晗旭, 龚颖

引用本文

许文韬, 王斌君, 朱莉欣, 王晗旭, 龚颖. 横向联邦学习后门的多方共治防范策略[J]. 计算机科学, 2024, 51(11A): 240100176-9.

XU Wentao, WANG Binjun, ZHU Lixin, WANG Hanxu, GONG Ying. Multi-party Co-governance Prevention Strategy for Horizontal Federated Learning Backdoors [J]. Computer Science, 2024, 51(11A): 240100176-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100176

横向联邦学习后门的多方共治防范策略

许文韬[1] 王斌君[1] 朱莉欣[2] 王晗旭[1] 龚 颖[1]
1 中国人民公安大学信息网络安全学院 北京 100038
2 西安交通大学苏州信息安全法学所 江苏 苏州 215123
(1625592944@qq.com)

摘 要 联邦学习易受到基于模型替换的后门攻击。针对目前后门检测方法效果不佳的问题,提出横向联邦学习后门的多方共治防范策略,旨在建立联邦学习中心服务器与客户端共治机制,从而在不破坏数据隐私与主任务性能的前提下有效检测并防范模型中的后门。该策略涵盖浅层后门扫描、深层后门检测和模型修复等内容,均由客户端在中心服务器的协同下完成。其中,浅层后门扫描是一种轻量级的实时后门检测方案,其并不显著增加时间开销。该方案由客户端捕捉聚合后模型参数的异常变化,并向中心服务器报告。当报告数达到设定的阈值时,中心服务器启动深层后门检测,各客户端会暂停联邦学习进程,进行深度检测,以确定模型中的神经元是否受到后门攻击的影响而表现异常。若存在异常,各客户端采用良性模型与受攻击模型拼接的方法,将模型恢复至良性状态,并将深层后门检测的结果以及模型修复方案提交至中心服务器,由中心服务器决定最终修复方案,从而彻底清除后门。实验结果表明,该策略可以有效地检测并清除联邦学习模型中存在的后门,为横向联邦学习的安全运行保驾护航。

关键词:联邦学习;后门攻击;后门检测;多方共治
中图分类号 TP391

Multi-party Co-governance Prevention Strategy for Horizontal Federated Learning Backdoors

XU Wentao[1],WANG Binjun[1],ZHU Lixin[2],WANG Hanxu[1] and GONG Ying[1]
1 College of Information and Cyber Security,People's Public Security University of China,Beijing 100038,China
2 Suzhou Institute of Information Security Law,Xi'an Jiaotong University,Suzhou,Jiangsu 215123,China

Abstract Federated learning is susceptible to backdoor attacks based on model replacement. In response to the poor performance of current backdoor detection methods,multi-party co-governance prevention strategy is proposed. The aim is to establish a co-governance mechanism between the federated learning center server and the client,so as to effectively detect and prevent backdoors in the model without compromising data privacy and main task performance. This strategy covers shallow backdoor scanning,deep backdoor detection,and model repair,all of which are completed by the client in collaboration with the central server. Among them,shallow backdoor scanning is a lightweight real-time backdoor detection scheme that does not significantly increase time overhead. This scheme captures abnormal changes in the aggregated model parameters by the client and reports them to the central server. When the number of reports reaches the set threshold,the central server initiates deep backdoor detection,and each client pauses the federated learning process for deep detection to determine whether the neurons in the model are affected by backdoor attacks and exhibit abnormalities. If there are anomalies,each client adopts a method of concatenating a benign model and an attacked model to restore the model to a benign state,and submits the results of deep backdoor detection and model repair plans to the central server. It is up to the central server to decide the final repair plan,thereby thoroughly clearing the backdoor. Experimental results show that this strategy can effectively detect and remove backdoors in the federated learning model,ensuring the safe operation of horizontal federated learning.

Keywords Federated learning,Backdoor attack,Backdoor detection,Multi-party co-governance

1 引言

联邦学习[1-2]作为一种新兴的分布式机器学习技术,不仅能够有效地利用隐私数据,还能够防止数据泄露,成为解决数据孤岛、数据共享与隐私安全问题的新范式。其中,横向联邦学习[3]作为最早提出的联邦学习模式,是以数据特征空间为基础,拓展数据样本的联邦学习模式。由于参与方掌握的数据特征空间基本相同,横向联邦学习通过使各参与方的数据以独立样本的形式投入联邦训练,实现了数据样本的扩展和共享,使得联邦模型能够充分利用更多的数据进行训练。

后门攻击是神经网络训练过程中一种严重的潜在威胁。其攻击流程为:在神经网络模型训练前,攻击者在训练数据中

基金项目:国家社会科学基金重点项目(20AZD114)
This work was supported by the Key Program of National Social Science Foundation(20AZD114).
通信作者:王斌君(wangbinjun@ppsuc.edu.cn)

注入部分的后门样本，例如在图像的特定位置覆盖特定图案，并将后门样本的标签类别修改至目标类别。在训练过程中，神经网络不仅训练目标任务，还会学习后门样本的特征。在训练过程中遭受后门攻击的神经网络，当后门样本特征出现时，将输出攻击者所指定的目标类别，从而实现人为控制神经网络模型输出的结果，这导致模型可能在特定情况下做出错误的决策，造成伤害或财产损失。由于攻击者并未干扰神经网络模型在目标任务上的正常训练，因此模型仍然能够正常运作并完成预期的任务。因此，这种攻击方法具有高度的隐蔽性，不易被察觉，其危害性极大。在联邦学习中，攻击者为了提高后门攻击对全局模型的影响，通常采用基于模型替换的后门攻击[4-8]。后门攻击者在联邦学习的后期进行，此时，神经网络模型逐渐收敛，各良性客户端对全局模型的梯度贡献较小，甚至可以忽略不计。而后门攻击者通过放大与后门模型对应的参数梯度值，能够迅速替换全局模型。即便只有一名后门攻击者存在，在未部署防御与检测策略的联邦学习环境中，也能够成功实施对全局神经网络模型的有效后门攻击。

通常情况下，为了保护隐私数据安全，主流的联邦学习通常采用安全聚合算法[9]，这意味着参数梯度值对中心服务器是不可见的。然而，安全聚合算法虽然确保了客户端的隐私数据，却使得中心服务器失去了检测异常参数梯度值的机会，导致联邦学习成为后门攻击的天然靶场。

Gu 等[25]首次提出了神经网络“BadNets”后门攻击，这种攻击方法通过在训练数据中注入特定触发模式，并修改标签类别，使模型在面对包含触发模式的输入样本时，能够按攻击者意愿输出结果。这一发现揭示了深度学习模式在训练过程中存在的后门隐患，后门攻击对联邦学习模型训练构成了安全威胁。然而，上述后门攻击方法在联邦学习聚合操作的限制下，对全局模型的影响被大幅削减。为此，Bagdasaryan 等[33]提出模型替换攻击方法，指出在联邦学习后期，全局模型趋于收敛，各参与者梯度更新信息较小，此时，攻击者可将后门攻击模型对应的参数梯度按照一定比例放大，抵消聚合策略的影响，实现在联邦学习一轮内将后门注入全局模型中。安全聚合策略保护用户隐私的同时，使中心服务器无法检查参与者提交的梯度信息，保护了放大梯度信息这一攻击行为。

为应对联邦学习中的后门隐患，研究者们针对性地设计了相关防御策略。梯度剪裁方法[10-11]等后门防御策略虽然能够抵御后门攻击，但通常以准确率降低、计算开销增加为代价。若联邦学习全程启用此类后门防御策略，势必会造成训练进程冗长、计算资源浪费等问题；若按满足后门攻击条件的联邦学习后期启用该防御策略，又难以把握防御策略的具体启用时间；另外，早期的联邦学习后门检测研究[12-14]多基于对明文参数梯度值的异常检测，难以适配当前主流联邦学习环境采用的安全聚合算法。Andreina 等[15]提出基于反馈的联邦学习后门检测方案，即每轮联邦学习结束时，均由客户端根据聚合后模型的性能，投票决定是否更新当轮参数梯度值。该方案提供了由客户端共同参与后门检测的新思路，但其无法有效检测基于触发器的后门攻击。Wang 等[16]提出逆向触发器工程，为集中式的神经网络模型提供了后门检测的新思路，但应用于对分布式的联邦学习环境时，存在漏检与误检的问题，效果不佳。

为应对此挑战，本文提出了横向联邦学习后门的多方共治防范策略（Multi-party co-governance Prevention Strategy，MPS），如图 1 所示。该策略涵盖浅层后门扫描、深层后门检测和模型修复 3 个关键环节。在训练开始后，各客户端在进行联邦学习模型训练的同时，采用轻量级的浅层后门扫描对聚合后的模型进行实时检测，一旦发现异常更新，即向中心服务器报告。随后，当中心服务器接收到的异常报告数超过预定阈值时，要求各客户端启动重量级的深层后门检测，即暂停联邦学习进程，并分析模型是否受到后门攻击。将检测结果向中心服务器报告，并提供模型修复方案。最后，中心服务器汇总并决策模型修复方案，部署后门防御策略，然后循环执行 MPS，直至联邦学习安全结束。

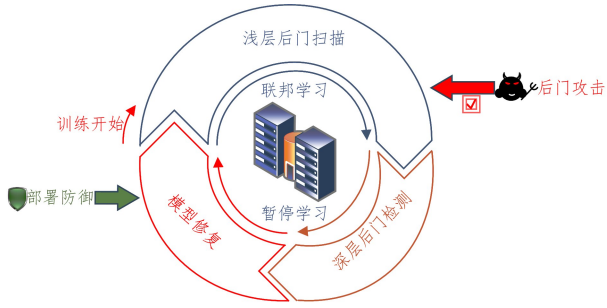


图 1 MPS 流程图

Fig. 1 MPS flowchart

这样，构成的先浅层后门扫描，再深层后门检测，最后进行模型修复的多方共治防范策略，并不要求中心服务器访问模型信息与用户数据信息，不影响模型主任务的准确性和性能，且保证了后门检测的精度。因此，该方案适合联邦学习生态环境。

本文的主要贡献可概括如下：

（1）提出横向联邦学习后门的多方共治防范策略，该策略可以适配现有基于安全聚合算法的联邦学习生态环境；

（2）该方案可以发现并清除神经网络中存在的后门，为后门的检测与修复提供了新思路；

（3）在 MNIST 和 Cifar-10 数据集上的实验验证了该方案的可行性。

2 相关技术

2.1 联邦学习

神经网络模型的训练需要大量数据，其中可能涉及隐私数据。为此，McMahan 等提出联邦学习[17]。联邦学习通常包含一个中心服务器与若干客户端，客户端始终将数据保留于本地，分布式训练神经网络模型。在这一过程中，通过与中心服务器交互神经网络的参数梯度值，替代了原始数据信息的流动，在数据共享和隐私保护的矛盾中寻得了一种良好的解决方案。与集中式神经网络的训练相比，数据隐私安全是联邦学习的首要前提。然而，研究表明，由于参数梯度值是基于原始数据训练得到的，诚实但好奇的中心服务器可以通过深度梯度泄露方法[18-19]，将参数梯度值还原出用户原始数据。因此，为了确保用户隐私不被窃取，当前主流联邦学习通常采用基于同态加密[20]的安全聚合策略。该策略通过在客户端对参数梯度值执行同态加密，然后将其上传至中心服务器，使真实的参数梯度值对中心服务器不可见，既保护了用户

隐私,又能满足中心服务器的相关聚合操作需求。

联邦学习协作过程可概括为:(1)中心服务器将模型及初始化参数发送至所有客户端;(2)中心服务器随机选择 m 个客户端,被选中的客户端在本地训练当前神经网络模型,并将参数梯度值经同态加密处理后上传至中心服务器;(3)中心服务器聚合全部的梯度信息,并将聚合结果发送至所有客户端,客户端对结果同态解密后,更新本地模型;(4)重复执行(2)和(3),直至模型收敛。

以全局平均聚合(FedAvg)策略为例,式(1)所示为第 $t+1$ 轮训练后全局神经网络模型的聚合结果。

$$G_{t+1}=G_t+\frac{\sigma}{m}\sum_{i=1}^{m}(L_{t+1}^{i}-G_t) \tag{1}$$

其中,L_{t+1}^{i} 为客户端 i 第 $t+1$ 轮训练后得到的本地模型,G_t 与 G_{t+1} 分别为第 t 轮和第 $t+1$ 轮的全局神经网络模型,σ 为全局学习率,m 表示第 $t+1$ 轮联邦学习的参与客户端数。

2.2 后门攻击

在神经网络模型训练过程中,后门攻击者向本地的部分训练数据中注入触发器,并修改对应数据的标签类别,训练一个既可以完成目标任务,又可以按攻击者意愿输出特定数据类别的神经网络模型,其训练目标如式(2)所示。

$$\arg\min_{(L_{t+1}^{\mathrm{att}})}\Big(\sum_{x_p\in D_p}Loss(y_p,f(x_p))+\sum_{x\in D_c}Loss(y_c,f(x))\Big) \tag{2}$$

其中,L_{t+1}^{att} 表示第 $t+1$ 轮联邦学习攻击者本地训练的攻击模型,$Loss$ 表示用于衡量分类误差的损失函数,D_p 表示注入触发器的后门数据集合,D_c 表示原始数据集合,y_p 表示后门目标标签,y_c 表示原始标签,$f(x)$ 表示将数据 x 置入神经网络后的输出值。采用安全聚合策略的联邦学习无法检测客户端提交的异常参数梯度值,联邦学习存在后门隐患。

然而,联邦学习中攻击者处于少数,平均聚合策略削减了少数后门攻击者对全局神经网络模型产生的影响。因此,联邦学习的后门攻击通常与模型替换攻击相结合,只需在整个联邦学习训练后期完成一次后门攻击,即可实现用本地模型替换全局神经网络模型[5]。由于联邦学习后期的全局神经网络模型趋于收敛,除攻击者外的 $m-1$ 个正常客户端训练得到的本地模型与上轮全局模型差异较小,对全局神经网络模型的贡献有限,故有式(3)。

$$\sum_{i=1}^{m-1}(L_{t+1}^{i}-G_t)\approx 0 \tag{3}$$

攻击者可按照式(4)将攻击模型 L_{t+1}^{att} 的梯度值($L_{t+1}^{\mathrm{att}}-G_t$)放大 $\frac{m}{\sigma}$(缩放因子)后得到 L_{t+1}^{en}。

$$L_{t+1}^{\mathrm{en}}=\frac{m}{\sigma}(L_{t+1}^{\mathrm{att}}-G_t)+G_t \tag{4}$$

将式(3)和式(4)代入式(1)可得式(5),从而实现在一轮内用本地包含后门的神经网络模型替换全局神经网络模型,完成对全局神经网络模型的有效后门攻击。

$$G_{t+1}=G_t+\frac{\sigma}{m}\Big[(L_{t+1}^{\mathrm{en}}-G_t)+\sum_{i=1}^{m-1}(L_{t+1}^{i}-G_t)\Big]=L_{t+1}^{\mathrm{att}} \tag{5}$$

因此,基于模型替换的后门攻击在当前联邦学习环境的实际应用中攻击性强、风险大。

2.3 后门检测

Wang 等[16]将后门触发器的一般注入形式定义为式(6)。

$$x_p=(1-mask)\cdot x+mask\cdot pattern \tag{6}$$

其中,x_p 为后门数据;x 为原始数据;$pattern$ 表示后门触发模式;$mask$ 表示触发掩码,用于表示后门注入具体像素位置的掩码矩阵,由若干 0~1 之间的数表示,当 $mask$ 中某一处值取 1 时,代表该像素点完全由后门触发模式所取代。

后门攻击者为神经网络模型添加了触发器特征,当数据中出现触发器时,与触发器特征相对应的神经元激活,并输出异常高值。当原始数据特征与触发器特征同时出现时,触发器特征将原始数据特征覆盖,使模型总向后门标签类别输出。即当神经网络模型中存在后门时,每个后门数据与后门标签间存在“便捷通道”,使被注入后门的原始数据总向后门标签归类。因此,Wang 等利用后门攻击的这一特性,通过检测是否存在“便捷通道”,判定模型中是否存在后门。具体而言,检测者可利用本地数据集,针对每一个标签类别,通过逆向工程优化 $mask$ 与 $pattern$。如式(7)所示,逆向工程包含两个方面的优化目标。将数据分类至相应标签类别以及获取 L1 范数最小的触发掩码 $mask$。

$$\min_{mask}(Loss(y_c,f(x))+\lambda\parallel mask\parallel_1) \tag{7}$$

其中,λ 表示 $mask$ 的 L1 范数在优化项中所占比重。

获取每类标签对应的 $mask$ 及其 L1 范数后,采用基于绝对中位差的异常值检测方法[21],计算各标签异常指数,筛选全部标签类别对应 $mask$ 的 L1 范数中的异常值,如式(8)所示。

$$\begin{aligned}&MAD=median(|M-median(M)|)\\&a\_index=\frac{|M-median(M)|}{MAD\times constant(1.4826)}\end{aligned} \tag{8}$$

首先,计算中值绝对偏差,其中 $median$ 表示中值计算函数,M 表示所有 $mask$ 的 L1 范数集合,即 $M=\{m_1,m_2,m_3\cdots\}$。其次,假设数据分布为正态分布,利用常数估计器($constant=1.4826$)对异常指数进行规范化解决,得到异常指数集合 a_index。任何异常指数大于 2 且表现为异常低值所对应的标签类别被认为存在后门。

逆向触发器工程的提出为后门检测工作带来了曙光。然而,该方法要求检测者掌握模型信息与训练数据。在联邦学习的环境下,中心服务器无法满足该要求;而由客户端进行的逆向触发器工程检测效果不佳,易造成后门漏检或误检。故本文基于现有逆向触发器工程方法的原理,提出联邦学习环境下的深层后门检测算法。

3 横向联邦学习多级共治防范策略

在主流的联邦学习环境中,为了维护用户隐私安全,通常采用安全聚合算法,这导致中心服务器无法获取各客户端的梯度更新信息和模型参数信息。因此,中心服务器面临着无法完成后门检测任务的挑战。为解决这一问题,本文提出 MPS,其涵盖了浅层后门扫描、深层后门检测和模型修复 3 个关键环节。在这一策略中,各客户端通过与中心服务器的协作完成这 3 个环节,实现了分布式训练与分布式后门检测的同步进行,形成了一种多方共治的后门检测格局。其中,浅层后门扫描提供了轻量级实时后门检测,客户端能够捕捉模型参数的异常变化,同时避免显著的检测计算开销;服务器捕捉到大量客户端模型异常后,则启动深层后门检测,暂停联邦学习进程,检测神经元的异常激活,通过高置信度的评判来判断模型是否受到后门攻击。MPS 巧妙地结合了轻、重量级的

方法，最大程度地降低了后门检测的计算开销，同时有效保护了各客户端的隐私信息。随后，依赖模型修复环节将模型恢复至良性状态，由中心服务器部署后门防御策略，循环执行MPS，直至联邦学习安全结束。

3.1 浅层后门扫描

结合神经网络反向传播机制[22]可知，神经网络通过优化调整神经元之间的权重，构建输入数据与输出类别之间的正确关联。在神经网络中，末置隐藏层（即神经网络中最后一个隐藏层）与输出层相关性最强。因此，末置隐藏层的权重变化能够反映出神经网络的收敛状态。图2展示了由4个神经元的末置隐藏层与1个神经元的输出层构成的部分神经网络，其中紫色表示末置隐藏层中能表达数据特征的神经元（简称特征神经元，如图2中 x_2，x_3），与输出神经元（y）具有强关联。在训练过程中，通过反向传播机制，神经网络不断提高特征神经元与输出神经元间权重的绝对量，使特征神经元的激活状态向输出层神经元传导；同时，降低无关神经元（x_1，x_4）与输出神经元（y）间权重的绝对量，使其逐渐趋向于0，从而降低无关特征对输出神经元的影响。神经网络不断重复上述过程，以获取更高置信度的输出，使特征神经元与输出神经元间权重的L1范数（即 $\sum|W_i|$）不断提高，无关神经元与输出神经元间权重趋向于0，故总体表现为输出层各神经元权重L1范数（简称标签总权重值）随训练进程不断增加。

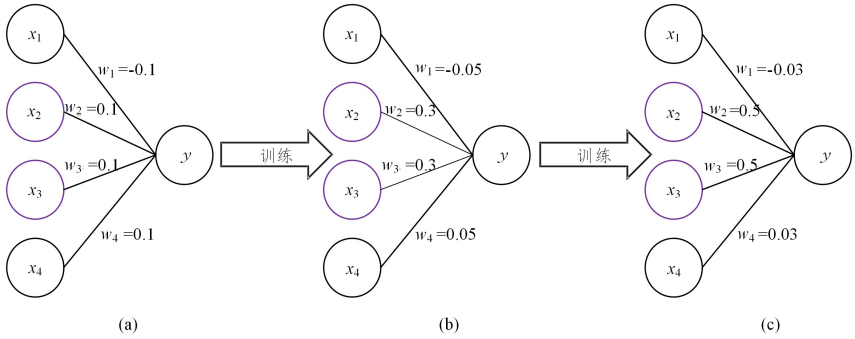


图2 末置隐藏层权重随训练变化情况

Fig. 2 Weights of the last hidden layer change with training

如2.2小节所述，联邦学习中有效的后门攻击往往在训练后期进行，此时模型主任务学习趋于收敛。攻击者为其本地部分数据添加触发器特征，并将其标签类别修改为目标标签。为使神经网络同时完成主任务与后门任务，模型利用反向传播机制重新调整目标标签相关权重的优化方向，破坏了目标标签的原有常规状态，无法维持目标标签总权重值的增长状态。而非后门目标标签正常完成主任务训练，所受影响较小。因此，在联邦学习受到后门攻击的训练轮次中，触发器的介入将使后门目标标签与非后门目标标签的总权重值变化量呈现出显著差异。

因此，浅层后门扫描通过检测每轮是否有标签类别总权重值相邻轮变化量与其他标签类别存在显著差异，从而发现后门目标标签。具体做法是：每轮均要求联邦学习客户端计算当前聚合后模型各标签总权重值矩阵与上一轮各标签总权重值矩阵的差（即差值矩阵），采用DBSCAN聚类算法[23]检测差值矩阵中的异常低值，该值所对应的标签类别被认为是潜在的后门标签类别，并向中心服务器报告扫描结果。当报告异常的客户端数达到阈值时，由中心服务器决定进行深层后门检测；反之，继续联邦学习进程。

浅层后门检测方案伴随联邦学习过程同时进行，计算量小，不显著增加计算开销。但并不意味着该异常必然由后门攻击造成。攻击者对联邦学习进行的可用性攻击[24-28]也可能导致模型参数的异常变化，因此，需要下一阶段的深层后门检测进行确认。

3.2 深层后门检测

后门攻击是攻击者通过毒化本地数据集，使神经网络模型学习触发器特征，而造成误分类。触发器特征经过神经网络的分析、提取，激活末置隐藏层中部分神经元（简称后门特征神经元），从而作用于模型输出。触发器具有特定模式和稀疏性[29]，触发器特征极易被神经网络分析、提取，使后门特征神经元获得相比于提取良性特征时更高的激活值。图3展示了由4个神经元的末置隐藏层与1个神经元的输出层构成的部分神经网络，当输入良性样本（x_{data}）时，特征神经元（x_3）被激活，模型正确输出 y 类；将该良性样本注入触发器形成后门样本（x_{poison}）后，后门特征神经元（x_1，x_2）被激活，以更高的激活值将原有特征覆盖，向后门目标标签（y_{att}）输出，造成误分类。

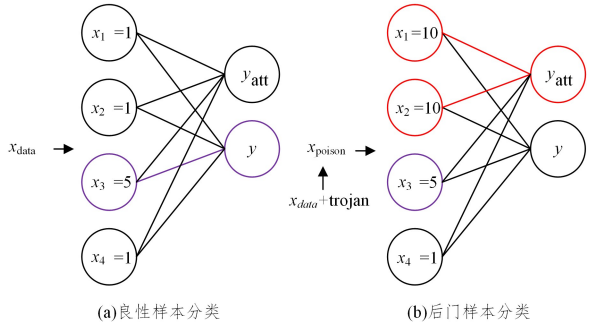


图3 末置隐藏层神经元随训练变化情况

Fig. 3 Neurons of the final hidden layer change with training

由此可见，触发器可导致末置隐藏层中更多的神经元异常激活，表现为末置隐藏层神经元激活值L1范数（简称总激活值，即图3中 $\sum|x_i|$）更高。

利用后门攻击的这一特性，基于逆向触发器工程[16]，深层后门检测要求各客户端本地逆向获取可以使样本误分类且总激活值异常高的最小触发器。具体做法是，中心服务器暂停联邦学习进程，要求所有客户端利用其本地数据检测后门；而各客户端运行如算法1所示的客户端深层后门检测算法，令逆向得到的触发器获得尽可能高的总激活值，并以总激活值是否异常作为评判模型中某一标签类别是否存在后门的

指标。其中,*labels* 为标签类别总数,*e* 为优化轮次,λ_1 和 λ_2 表示对应值在优化项中所占比重,*a_threshold* 为异常指数阈值。

算法 1 客户端深层后门检测算法算法

输入:(model)

输出:(backdoor,mask,pattern)

```
1. f=model[:-2];
   /* 创建空数组分别用于存储各类别总激活值、异常指标值、触发掩码、触发模式 */
2. all_active_value,a_index,mask,pattern=[]
3. for label←0 to labels do:
/* 初始化 mask 与 pattern */
4.     mask[label],pattern[label]=0;
5.     for round←1 to e do:
6.         x_p=(1-mask[label])·x+mask[label]·pattern[label]
7.         loss=loss_f(label,model(x_p))+λ1 L1_norm(mask)-λ2 L1_norm(f(x_p))
8.         mask[label]=mask[label]- ∂loss/∂mask[label]
9.         pattern[label]=pattern[label]-∂loss/∂pattern[label]
10.    end for
       /* 存储当前类别总激活值 */
11.    all_active_value[label]=L1_norm(f(x_p))
12. end for
    /* 按式(8)计算各类别异常指数 */
13. a_index=MAD (all_active_value,constant)
    /* 将异常指数大于阈值的位置置 1,其余置 0 */
14. backdoor=discriminator(a_index,a_threshold)
15. return backdoor,mask,pattern
```

各客户端针对每个标签类别,初始化触发掩码和触发模式,并在 *e* 轮逆向优化过程中,按当前触发掩码为本地训练集中的数据注入当前触发模式,通过设定损失函数(*loss*),在最小化触发掩码的基础上,使后门数据 x_p 被误分类至当前标签类别,同时,使神经网络获取尽可能高的总激活值。优化结束后,获取所有标签类别的总激活值。使用如 2.3 小节所述的基于绝对中位差的异常值检测方法计算所有标签类别总激活值的异常指数,当异常指数大于异常指数阈值(*a_threshold*)且总激活值表现为异常高值时,意味着该标签类别拥有触发器特征,其对应标签类别被认为是后门目标标签。

3.3 模型修复

通过浅层后门扫描及深层后门检测确定神经网络模型中存在后门后,需将其恢复至良性状态。采用直接将模型回退至后门攻击前状态的方法并不理想。若攻击者在联邦学习后期进行后门攻击,其本地保存一份攻击所使用的异常梯度值,当后门被发现时,模型回退至后门攻击前的状态,攻击者仍可以直接使用该异常梯度值重新完成对模型的后门攻击,大大降低了攻击成本,提高了模型再次面对后门攻击的风险。因此,本文采用部分层回退的模型修复策略。

触发器特征经神经网络正向传播,被分析、提取,激活神经网络隐藏层中的部分神经元(简称触发器神经元),并将激活状态向后置神经网络层传播。因此,全局神经网络无需回退至后门攻击前的状态,保留触发器特征未被完全分析、提取的前若干层(即触发器神经元前若干层)神经网络,仅将其后的神经网络层回退至攻击前状态,也可消除触发器神经元的影响,从而修复后门。同时,拼接后的模型使攻击者无法使用相同的异常梯度值再次后门攻击。

为寻找到具体的拼接位置(即寻找触发器神经元),客户端可通过遍历全部的拼接方案,利用深层后门检测算法中优化所得后门标签的触发模式及触发掩码($mask_b$,$pattern_b$),检查不同拼接方案的总激活值是否恢复至良性水平,并结合模型主任务性能,即可确定拼接的具体位置。由于最终检视拼接方案的总激活值是否恢复至良性水平,因此拼接模型不需要包含输出层。

部分层回退的模型修复如算法 2 所示。各客户端定义一个新神经网络为当前模型(*model*)的 0 至 *n* 层与后门攻击前良性模型($model_{pre}$)的 *n* 至(*layers*−2)层的拼接,从后至前(即从 *n*=*layers*−2 始,至 *n*=0 止)遍历 *n* 的全部可取值,将算法 1 得到的后门标签对应触发掩码与触发模式注入本地数据集,得到不同拼接方案的总激活值,检查该值是否恢复至良性水平(即低于修复阈值,修复阈值可定义为良性标签总激活值的最大值)。其中,$mask_b$ 和 $pattern_b$ 为算法 1 所得后门触发器,*r_threshold* 为修复阈值,*layers* 为神经网络层数。

算法 2 客户端部分层回退算法

输入:(model,$model_{pre}$,$mask_b$,$pattern_b$)

输出:(n)

```
1. /* model 为待修复模型,model_pre为后门攻击前的良性模型 */
2. for n←layers-2 to 0 do:
3. /* 为原始数据注入由算法 1 逆向得到的触发器 */
4.     x_p'=(1-mask_b)·x+mask_b·pattern_b;
5. /* 拼接模型的前半部分 */
6.     f=model[0:n];
7. /* 拼接模型的后半部分 */
8. f_pre=model_pre[n:-2];
9. model_joint=f⊕f_pre;
10.    /* 计算当前拼接方案总激活值 */
11.    activite_value=L1_norm(model_joint(x_p'));
12.    /* 判断当前总激活值是否小于修复阈值 */
13.    if activite_value<r_threshold do:
14.        /* 结束循环,输出拼接位置 */
15.        returnn;
16. end for
```

客户端在执行算法 1、算法 2 后,将后门标签(backdoor)与修复位置(*n*)一并提交至中心服务器,后者根据各客户端所提供的线索,最终确定修复位置,并下发至各客户端,完成对后门的修复。随后,可部署可行的后门防御策略,继续完成后续联邦学习进程。

4 实验分析

本节将 MPS 部署于存在后门攻击者的联邦学习环境中,评估浅层后门扫描、深层后门检测、模型修复方法的有效性。MPS 中各项内容均由客户端投票参与决策。在实际应用中,为防止攻击者伪造良性客户端并提交大量虚假投票信息,可采用数字签名等技术保障投票的真实性。

4.1 实验设置

4.1.1 数据集与训练模型

本文使用了两个数据集 MNIST 和 Cifar-10[31] 进行评估,其基本信息如表 1 所列。

MNIST 数据集来自美国国家标准与技术研究所

(NIST)。训练集由来自 250 个不同人手写的数字构成,用于识别 10 种不同的手写数字灰度图像,共包含 7 万张 28×28 灰度图像,总计 10 个分类,其中 6 万张训练图像和 1 万张测试图像。Cifar-10 是由 Alex 等收集的一个用于普适物体识别的计算机视觉数据集,它包含 60000 张 32×32 的 RGB 彩色图片,总计 10 个分类,其中包括 5 万张用于训练集,1 万张用于测试集。

表 1 数据集基本信息

Table 1 Dataset basic information

数据集	训练样本数	测试样本数	数据尺寸	模型结构
MNIST	60000	10000	1 * 28 * 28	7 个卷积层与 3 个全连接层
Cifar-10	50000	10000	3 * 32 * 32	

本文采用的联邦学习训练模型结构信息如表 2 所列,共包含 7 个卷积层与 3 个全连接层。其中,全连接层神经元节点个数根据不同数据集进行调整。

表 2 训练模型结构信息

Table 2 Training model structure information

层序	卷积层(输入通道,输出通道,滤波器尺寸)/全连接层	激活函数	池化层(池化窗口尺寸,移动步幅)
1	Conv2d(1 or 3,64,3)	Relu	
2	Conv2d(64,64,3)	Relu	Maxpool(2,2)
3	Conv2d(64,128,3)	Relu	—
4	Conv2d(128,128,3)	Relu	Maxpool(2,2)
5	Conv2d(128,256,3)	Relu	—
6	Conv2d(256,256,3)	Relu	—
7	Conv2d(256,256,3)	Relu	Maxpool(2,2)
8	Linear	Relu	—
9	Linear	Relu	—
10	Linear	Softmax	—

4.1.2 参数配置

联邦学习实验参数信息如表 3 所列。联邦学习共包含 10 个客户端,随机平均分配训练样本,其中包含 1 名后门攻击者,于联邦学习第 20 轮进行后门攻击(模型已趋于收敛),将触发器(位于图片右下角的 4×4 固定位置像素值)注入后门数据在本地数据集的 5%,并修改其标签为 4,对全局神经网络模型采用基于模型替换的后门攻击,缩放因子 $\gamma=5$。

表 3 联邦学习参数信息

Table 3 Federated learning parameter information

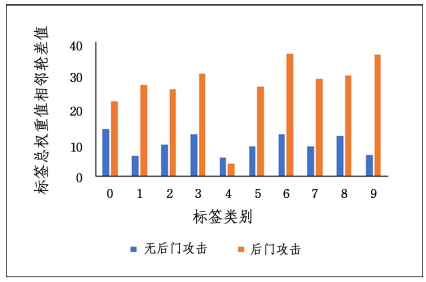
客户端总数 m	采样率/r	全局学习率/σ	缩放因子 $\gamma=m\cdot r/\sigma$
10	0.5	1	5

良性客户端在每轮联邦学习结束时均对聚合后模型执行浅层后门扫描,采用基于 DBSCAN 聚类的异常值检测算法,向中心服务器报告异常标签类别。经实验测试,DBSCAN 聚类算法领域半径的设置与末置隐藏层神经元数量成正相关。当末置隐藏层神经元较多时,各标签总权重值的相邻轮变化差异越大,可赋予更大的邻域半径,以避免误检。设 $\beta=4096$,当末置隐藏层神经元数量为 $1/2\beta,\beta,2\beta,4\beta$ 时,对应邻域半径 r 可取 5,6,8,10,可有效避免误检。本文实验环境中,MNIST 数据集末置隐藏层神经元数量为 $1/2\beta$,设定邻域半径为 5;cifar-10 数据集末置隐藏层神经元数量为 β,设定邻域半径为 6。

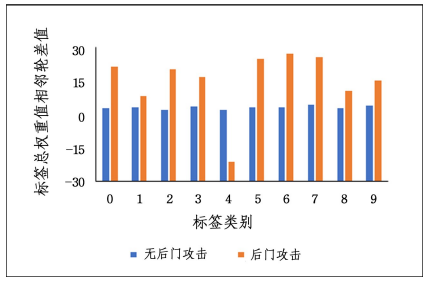
中心服务器得到超过半数的扫描异常报告时,要求客户端执行深层后门检测。客户端采用基于绝对中位差的异常值检测算法检测各标签总激活值中的异常值,设定异常指数阈值 $a\_threshold=3.5$(经实验测试,该阈值可有效检出后门并避免误检)。最后,各客户端执行模型修复。

4.2 浅层后门扫描与深层后门检测的有效性测试

图 4 分别显示了神经网络在常规状态、第 4 类标签受后门攻击的异常状态下,各标签总权重值相邻轮(第 20 轮与第 19 轮)差值变化的情况。其中,横坐标轴表示标签类别,纵坐标轴表示标签总权重值相邻轮的差值。由图 2 可知,常规良性状态下,各标签总权重值与上轮相比,各类别均小幅增加;而当神经网络受到后门攻击时,后门数据加入训练,后门目标标签相关权重进行了重构,导致其与良性标签总权重值的相邻轮差值存在显著差异,该差异大于所设定的邻域半径。后门扫描通过捕捉这一差异,可发现潜在的后门标签类别。



(a)MNIST 浅层后门扫描



(b)Cifar-10 浅层后门扫描

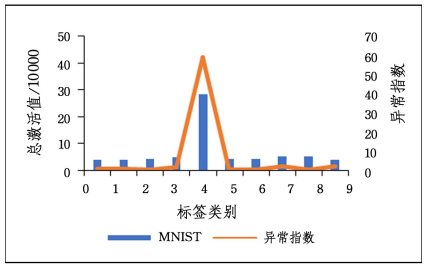
图 4 浅层后门扫描

Fig. 4 Shallow backdoor scanning

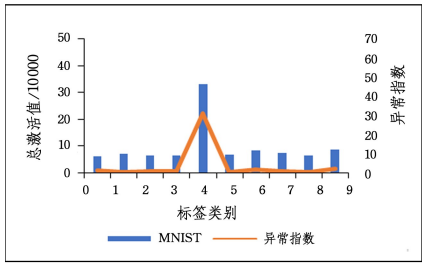
获取浅层后门扫描结果后,各客户端对预更新神经网络模型执行深层后门检测。图 5 显示了联邦学习第 20 轮受后门攻击时神经网络深层后门检测的结果。其中,横坐标轴表示标签类别,左纵坐标轴表示总激活值,右纵坐标轴为异常指数 a_index。由图可知,后门标签对应的触发模式,可使神经网络模型获得异常高的总激活值,其异常指数显著高于阈值。为形成对照,图 6 则显示了当其他条件相同时,联邦学习第 20 轮未受后门攻击时客户端运行深层后门检测的结果。由图可知,当不存在后门标签时,各标签总激活值整体无显著差异。由此证明,深层后门检测可用于检测后门标签。

本文在 MNIST 与 Cifar-10 数据集上进行了大量实验,分别测试了浅层后门扫描与深层后门检测的性能,并与客户端采用逆向触发器工程方法检测后门的实验结果进行了对比。在两数据集上包含后门攻击者的环境下各运行 10 次,每次均选定不同的后门标签类别,客户端在第 1—20 轮结束时均对神经网络模型进行浅层后门扫描,得出浅层后门扫描总体真阳率(TPR)与假阳率(FPR)。各客户端第 20 轮结束时对神经网络模型执行深层后门检测与逆向触发器工程,得出两方法的真阳率。以此对比浅层后门扫描、深层后门检测与逆向触发器工程的检测性能。此外,在未包含后门攻击者的联邦学习环境下同样运行 10 次,各客户端于第 20 轮各执行一

次深层后门检测和逆向触发器工程,得出两方法的假阳率。表 4 列出了浅层后门扫描、深层后门检测与逆向触发器工程的相关检测指标。



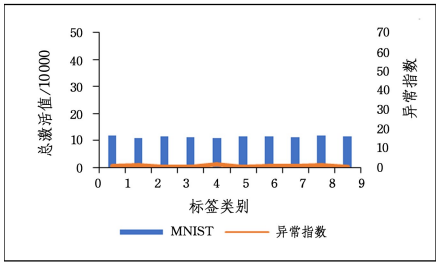
(a)MNIST 深层后门检测



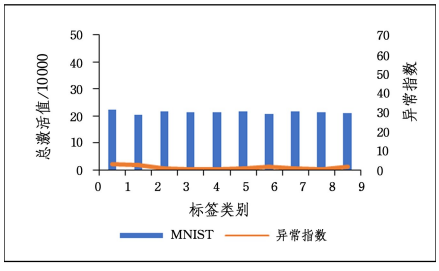
(b)Cifar-10 深层后门检测

图 5　深层后门检测(含攻击者)

Fig. 5　Deep backdoor detection(including attackers)



(a)MNIST 深层后门检测



(b)Cifar-10 深层后门检测

图 6　深层后门检测(未含攻击者)

Fig. 6　Deep backdoor detection(no attacker included)

表 4　不同后门检测方法的性能

Table 4　Performance of different backdoor detection methods

(%)

数据集	浅层后门扫描		深层后门检测		逆向触发器工程	
	TPR	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*
MNIST	100	0.52	95	0	83	0
Cifar-10	100	1.57	99	1	90	5

由表可知,浅层后门扫描可以检出全部的后门攻击行为,且误检较少,说明后门攻击会导致模型参数异常变化,被浅层后门扫描捕捉。但当攻击者采取其他攻击行为时,也可能导致参数的异常变化,被浅层后门扫描捕捉。因此,仍有必要进一步采取深层后门检测以确定参数异常的原因。由于各参与方仅掌握其本地数据,且每轮结束时,模型参数经过聚合运算,因此将集中式训练背景下的逆向触发器工程方法用于联邦学习后门检测中存在局限性,表现为在 MNIST 数据集、Cifar-10 数据集上的真阳率分别为 83%和 90%,存在一定漏检情况,且在 Cifar-10 数据集上假阳率为 5%,误检较多。而深层后门检测从后门标签拥有触发器特征的本质出发,利用总激活值检测后门标签,更适用于联邦学习环境下的后门检测工作,表现为更高的真阳率与更低的假阳率。

综上所述,在联邦学习环境下,相比于逆向触发器工程,浅层后门扫描与深层后门检测拥有更好的检测性能;且应用过程中,浅层后门扫描与深层后门检测均可以定位潜在的后门标签类别。浅层后门扫描作为轻量级检测方案,于深层后门检测前运行,定位潜在的后门标签类别,可以提高深层后门检测的效率。检测者可以仅针对浅层后门扫描异常的标签类别及其他几个良性标签类别进行深层后门检测,从而使检测者可以更有针对性地进行深层后门检测,降低检测时间开销。

4.3　模型修复

通在浅层后门扫描与深层后门检测确认模型中后门的存在后,执行模型修复方案。各客户端寻求第 19 轮(受攻击前良性模型)与第 20 轮(受攻击模型)全局神经网络模型的良性拼接方案,利用深层后门检测中逆向获取的后门触发模式 *pattern* 与触发掩码 *mask*,将其注入本地数据集后,迭代尝试的拼接位置 n,得到不同拼接位置的总激活值与良性数据测试集准确率,如图 7、图 8 所示。为验证模型修复的有效性,同时绘制了不同位置拼接的后门数据准确率。经实验测试,各客户端可得到相似结果。其中,横坐标为神经网络拼接位置 n,左纵坐标轴表示总激活值,右纵坐标表示准确率。本实验神经网络中,共 10 个神经网络层,当 $n=9$ 时,拼接模型为当前的受攻击模型;当 $n=0$ 时,模型为后门攻击前模型;取深层后门检测所得的良性标签总激活值的最大值为修复阈值。根据图 5,将良性标签总激活值最大值设定为修复阈值,得到 MNIST 数据集与 Cifar-10 数据集的修复阈值分别为 50 118 和 84 315。

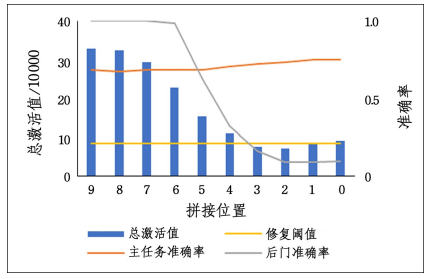


图 7　Cifar-10 模型修复

Fig. 7　Cifar-10 model repair

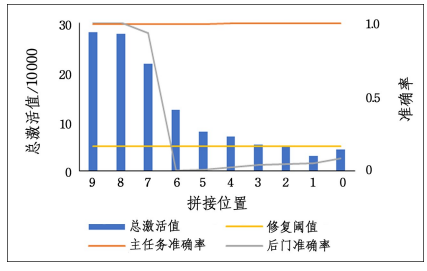


图 8　MNIST 模型修复

Fig. 8　MNIST model repair

可以看出，后门数据准确率随拼接位置 n 的降低而显著降低；相比于当前受攻击模型，拼接模型总激活值均有所降低。同时，测试集准确率相比于当前受攻击模型有所提高。在 Cifar-10 数据集上，当 $n=3$ 时，总激活值首次低于修复阈值，恢复至处于良性水平，后门攻击率仅为 15.7%，模型面对后门数据时表现为随机性输出；在 MNIST 数据集上，当 $n=2$ 时，总激活值首次低于修复阈值，此时后门攻击率仅为4.6%。因此，模型修复可以有效清除模型中存在的后门。

然而，即便知晓攻击者的存在，联邦学习的隐私要求与安全聚合策略仍使中心服务器无法定位后门攻击者的具体位置并访问客户端本地数据。因此，建议后续训练过程中继续采用 MPS，同时部署后门防御策略，如基于梯度剪裁的后门防御策略[10]、基于随机断层与梯度剪裁的防御策略[11]，保护全局神经网络模型免受后门攻击的影响，保障联邦学习安全结束。

结束语 在采用安全聚合策略的联邦学习环境中，中心服务器无法根据客户端提交的梯度更新值来检测恶意客户端。全程采用后门防御策略的解决方案通常会导致准确率降低及计算开销增加等问题。若选择在满足后门攻击条件的联邦学习后期启用防御策略，又面临确定具体启用时间的难题。MPS 成功解决了这一难题。其中，浅层后门扫描从模型参数入手进行检测，可捕捉神经网络模型训练时参数的异常变化；深层后门检测则从神经元激活值入手，检查是否存在标签类别拥有触发器特征，并由模型修复策略进行修复。在无法确定联邦学习中是否存在后门攻击者之前，可采用轻量级的浅层后门扫描方案，以降低检测与防御工作的计算开销。在中心服务器收到一定数量的报警后，才启用深层后门检测与模型修复，并有针对性地进行防御部署。综上所述，MPS 的提出为联邦学习中的后门检测与模型修复带来了全新的方法和思路。

参 考 文 献

[1] CHEN H L, FU C, ZHAO J S, et al. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks[C]// IJCAI. 2019: 4658-4664.

[2] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[3] 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.

[4] GU T Y, DOLAN-GAVITT B, GARG S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv: 1708.06733, 2017.

[5] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]// International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[6] WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: Yes, you really can backdoor federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084.

[7] GU T, LIU K, DOLAN-GAVITT B, et al. Badnets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.

[8] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv: 1712.05526, 2017.

[9] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[10] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning? [J]. arXiv: 1911.07963, 2019.

[11] XU W T, WANG B J. Backdoor Defense of Horizontal Federated Learning Based on Random Cutting and Gradient Clipping [J]. Computer Science, 2023, 50(11): 356-363.

[12] YIN D, CHEN Y, KANNAN R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]// International Conference on Machine Learning. PMLR, 2018: 5650-5659.

[13] MI Y, GUAN J, ZHOU S. Ariba: Towards accurate and robust identification of backdoor attacks in federated learning[J]. arXiv: 2202.04311, 2022.

[14] FUNG C, YOON C J M, BESCHASTNIKH I. The limitations of federated learning in sybil settings[C]// 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020). 2020: 301-316.

[15] ANDREINA S, MARSON G A, MÖLLERING H, et al. Baffle: Backdoor detection via feedback-based federated learning[C]// 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). IEEE, 2021: 852-863.

[16] WANG B, YAO Y, SHAN S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019: 707-723.

[17] GEIPING J, BAUERMEISTER H, DRÖGE H, et al. Inverting gradients-how easy is it to break privacy in federated learning? [J]. Advances in Neural Information Processing Systems, 2020, 33: 16937-16947.

[18] ZHU L, HAN S. Deep leakage from gradients[C]// Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2019: 14747-14756.

[19] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Artificial Intelligence and Statistics. Florida: PMLR, 2017: 1273-1282.

[20] FANG H, QIAN Q. Privacy preserving machine learning with homomorphic encryption and federated learning[J]. Future Internet, 2021, 13(4): 94.

[21] LEYS C, LEY C, KLEIN O, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median[J]. Journal of Experimental Social psychology, 2013, 49(4): 764-766.

[22] ZOPH B, LE Q V. Neural architecture search with reinforce-

ment learning[J]. arXiv:1611.01578,2016.

[23] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algo-rithm for discovering clusters in large spatial databases with noise[C]// KDD. 1996:226-231.

[24] GARBER L. Denial-of-service attacks rip the Internet[J]. Computer,2000,33(4):12-17.

[25] FANG M,CAO X,JIA J,et al. Local model poisoning attacks to {Byzantine-Robust} federated learning[C]// 29th USENIX security symposium(USENIX Security 20). 2020:1605-1622.

[26] XIE C,KOYEJO O,GUPTA I. Fall of empires:Breaking byzantine-tolerant sgd by inner product manipulation[C]// Uncertainty in Artificial Intelligence. PMLR,2020:261-270.

[27] BLANCHARD P,EL MHAMDI E M,GUERRAOUI R,et al. Machine learning with adversaries:Byzantine tolerant gradient descent[J]. Advances in Neural Information Processing Systems,2017,30.

[28] FUNG C,YOON C J M,BESCHASTNIKH I. The limitations of federated learning in sybil settings[C]// 23rd International Symposium on Research in Attacks[C]// Intrusions and Defenses (RAID 2020). 2020:301-316.

[29] GU T,DOLAN-GAVITT B,GARG S. Badnets:Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv:1708.06733,2017.

[30] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoim-mune Diseases,2009,1(4).

XU Wentao, born in 1999, postgraduate. His main research interests include federated learning and backdoor attack.

WANG Binjun, born in 1962, Ph.D, professor, Ph.D supervisor, is a member of CCF(No. E200014787M). His main research interests include network security and law enforcement.