第52卷第5期
2020年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 52 No. 5
Oct. 2020

DOI:10.16356/j.1005-2615.2020.05.001

联邦学习安全与隐私保护综述

陈　兵，成　翔，张佳乐，谢袁源

（南京航空航天大学计算机科学与技术学院/人工智能学院，南京，211106）

摘要：联邦学习是一种新型的分布式学习框架，它允许在多个参与者之间共享训练数据而不会泄露其数据隐私。但是这种新颖的学习机制仍然可能受到来自各种攻击者的前所未有的安全和隐私威胁。本文主要探讨联邦学习在安全和隐私方面面临的挑战。首先，本文介绍了联邦学习的基本概念和威胁模型，有助于理解其面临的攻击。其次，本文总结了由内部恶意实体发起的3种攻击类型，同时分析了联邦学习体系结构的安全漏洞和隐私漏洞。然后从差分隐私、同态密码系统和安全多方聚合等方面研究了目前最先进的防御方案。最后通过对这些解决方案的总结和比较，进一步讨论了该领域未来的发展方向。

关键词：计算机系统结构；联邦学习；模型安全；隐私保护

中图分类号：TP393　　**文献标志码：**A　　**文章编号：**1005-2615(2020)05-0675-10

Survey of Security and Privacy in Federated Learning

CHEN Bing, *CHENG Xiang*, *ZHANG Jiale*, *XIE Yuanyuan*

(College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics & Astronautics, Nanjing, 211106, China)

Abstract: Federated learning is a novel distributed learning framework which enables the sharing of training data across multiple participants without compromising their data privacy. However, such novel learning mechanism can still suffer from unprecedented security and privacy threats from various attackers. This article mainly explores the security and privacy challenges of federated learning by first introducing the preliminary knowledge and threat models to facilitate understanding of the potential attacks. Second, three types of attacks launched by the internal malicious entities are summarized and meanwhile the security and privacy vulnerabilities of federated learning architecture are analyzed. Third, the state-of-art protection solutions in aspects of differential privacy, homomorphic cryptosystem, and secure multi-party aggregation are surveyed. Finally, by summarizing and comparing these solutions, the promising directions are discussed.

Key words: computer system structure; federated learning; model security; privacy protection

随着机器学习方法在绝大多数识别相关的领域展现出了明显的优势[1]，使得在工业制造[2]、医疗卫生[3]、智能交通[4]和财务管理[5]等众多行业中涌现出了大量的智能应用。作为一种极具发展潜力的服务于大数据分析技术的解决方案，机器学习方法借助面向任务的诸多技术（例如：分类、回归和聚类和降维等）实现了自动识别与智能决策两个重要功能[6]。然而，随着数据驱动的智能应用快速发展，机器学习范式也面临着新的困境与挑战[7]。一方面，机器学习范式希望能为所有用户提供一种稳健

收稿日期：2020-05-30；**修订日期：**2020-08-10

作者简介：陈兵，男，教授，博士生导师。长期从事无线网络、网络安全及基于机器学习、数据挖掘的网络应用等领域教学和科研工作。承担各类科研项目30余项，发表论文40余篇，申请专利20余项。获得江苏省教学成果一等奖和二等奖各1项，国防科技进步三等奖2项。

通信作者：陈兵，E-mail：cb_china@nuaa.edu.cn。

引用格式：陈兵，成翔，张佳乐，等. 联邦学习安全与隐私保护综述[J]. 南京航空航天大学学报，2020，52(5)：675-684.
CHEN Bing, CHENG Xiang, ZHANG Jiale, et al. Survey of security and privacy in federated learning[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 675-684.

且高效的功能服务(例如谷歌的翻译系统)。而另一方面,数据作为学习算法的"营养"却很难做到充分的共享[8]。例如在工业应用场景中,很少有企业愿意共享其数据资源,这种情况主要受限于某些商业因素(例如:市场竞争和管理策略)[9]。即使对于个体参与用户,他们也依然担忧将本地数据集外包于服务提供商所带来的隐私泄露风险可能会远远超出这种便捷的在线服务所带来的收益[10]。

为了解决这一问题,联邦学习[11-14]作为一种极具潜力的解决方案应运而生,其主要创新性在于提供了一种具有隐私保护特性的分布式机器学习框架,并且能够以分布式的方式协同数以千计的参与者针对某个特定机器学习模型进行迭代训练。由于在联邦学习过程中训练数据仍然保存在参与者本地,这种机制既能实现对各参与者训练数据的共享,又能保证对每个参与者隐私的保护[15]。联邦学习的基本工作流程主要包括:(1)参与者从云服务器上下载已成初始化的全局模型,使用本地数据集对该模型进行训练,并生成最新的本地模型更新(即模型参数)。(2)云服务器通过模型平均算法收集各个本地更新参数并更新全局模型。由于联邦学习具有的独特优势——可在保护数据隐私的前提下实现由多个参与者的本地数据训练出统一的机器学习模型,因此在隐私敏感的场景中(包括金融业、工业和许多其他数据感知场景)联邦学习展现出了极好的应用前景[16-17]。

尽管联邦学习具有上述明显优点,由于以下3方面主要原因,其安全与隐私的问题依然存在[18]。(1)由于联邦学习框架中的云服务器没有访问参与者本地数据及其训练过程的权限,使得恶意参与者可以上传不正确的模型更新以达到并破坏全局模型的目的。例如,内部攻击者可以通过已经修改后的训练数据训练得到的投毒模型更新,达到有效影响全局模型准确性的目的。(2)由于将本地模型更新和全局模型参数相结合可得到训练数据中的隐含知识,使得用户的个人信息可能被泄露给不可信的服务器或者其他恶意用户。例如,即便是由自其他用户的训练数据产生的原型样本也有可能被恶意用户隐秘地窃取。(3)在不可信的云服务器和恶意参与者的合谋攻击下,每个人确切的隐私信息都是有可能被泄露的。

已有的研究结果表明,在联邦学习中学习模型的安全性和用户的隐私信息会受到一系列被动和主动攻击的威胁[19]。特别是那些由内部实体(如恶意用户和不受信任的服务器)发起的强大攻击具有更强的威胁性[20]。虽然在深度学习领域,特别是在神经网络学习场景中,保证安全性和隐私的工作已探索多年,但针对如何构建具有安全和隐私性的联邦学习系统的研究仍处于初级阶段[21-22]。因此,本文对联邦学习领域的安全和隐私威胁进行了综合的调研,并进一步整理和总结了该领域最新解决方案的发展趋势。与目前已有的联邦学习综述文章相比,本文主要关注联邦学习自身的架构安全性与隐私问题,从攻击模型、攻击方法、隐私威胁模型以及防御方案等几个方面,全面剖析联邦学习的脆弱性以及可应用性。

1 联邦学习

联邦学习是一种由谷歌[12]提出的非常有效的工具,其主要目的是借助多个移动设备生成的私有训练数据联合进行机器学习模型训练。涉及分布式移动用户的数据隐私可以通过只上传模型参数(例如梯度)代替上传原始数据进行保护。在某个特定设备上的训练样本通常遵循该移动用户的使用偏好,因此多个训练数据的分布特征是不平衡的(或者为非独立同分布)。例如,在一个单词预测任务中,潜在的训练数据可能包括该用户在此移动设备上输入的任何形式的信息(包括URL、密码和消息等形式)[23]。这些数据的特征分布很可能在多个移动用户之间具有很大差异,并且个体用户的数据特征分布与公共训练数据集(例如Wikipedia和其他数据集)特征分布的差异可能更为明显[24]。

由于联邦学习与传统的机器学习方法相比在模型训练阶段具有显著的优势,本文将重点分析和研究该阶段的主要特性。完整的联邦学习模型训练阶段主要包括以下步骤[25-26]。(1)初始化:所有的用户在他们的设备中都获取了一个预先分配好的神经网络模型,他们可以自愿加入联邦学习协议,并确定相同的机器学习及模型训练目标。(2)本地训练:在某轮特定的通信过程中,联邦学习参与者首先从中央服务器下载全局模型参数,然后又使用其私有训练样本对模型进行训练,生成本地模型更新(即模型参数),并将这些更新发送至中央服务器。(3)模型平均:通过聚合不同训练样本训练得到的所有模型更新并进行平均计算便可得到下一轮的全局模型。

在联邦学习过程中将迭代地执行上述步骤以达到优化当前全局模型的目的,整个迭代过程将在全局模型参数满足收敛条件时停止。在实际应用时,小批量随机梯度下降法(Mini-batch stochastic gradient descent,MSGD)非常适合建立联邦优化算法,其本地训练策略为

$$L_{t+1}=L_t-\eta\cdot\nabla_{L_t}\mathcal{L}(L_t,b) \tag{1}$$

式中：L_t代表第t轮通信轮次中服务器分发给用户的本地模型参数；b代表本地训练批量大小；η代表学习速率。

例如，在谷歌的FedSGD算法[12]中，每轮迭代都会选择一定比例的用户，通过最小化这些设备拥有的所有训练数据的损失函数来计算梯度。以这种方式，每个用户只需要在当前的全局模型上执行一轮随机梯度下降法计算，然后服务器将承担所有剩余的任务以得到当前模型权重的平均值，服务器端的全局模型更新机制为

$$G_{t+1}=G_t+\frac{1}{n_t}\sum\nolimits_{i\in[n_t]}\Delta L_{t+1}^{i} \tag{2}$$

式中：G_t代表第t轮通信轮次中服务器端的全局模型参数；n_t代表本轮次中选取的参与者数量；ΔL_{t+1}^{i}代表服务器接收到用户发出的本地模型更新参数。

2　安全和隐私威胁

在联邦学习场景中，攻击行为不但可以由不受信任的服务器发起也有可能由恶意用户发起[27]。一方面，由服务提供商部署的服务器被视为被动攻击者，其安全模型是诚实但好奇的。这意味着这些服务器通常会严格地按照既定的学习协议提供服务，但他们的同事也试图从本地模型更新中泄露用户的一些敏感信息。另一方面，参与者被视为主动攻击者，他们试图从由训练数据形成共享的全局模型参数中恢复出用户的敏感信息。

2.1　威胁目标

无论在被动攻击还是主动攻击场景中，攻击者的主要目的都是破坏学习模型具有的基本性能，主要包括机密性、完整性和可用性[28-29]。首先，攻击者不仅可以窃取训练数据中嵌入的敏感信息，还可以通过暴露目标模型信息及其预测结果来破坏机密性，严重威胁用户的隐私。其次，对完整性和可用性的威胁主要集中在机器学习模型的输出上，将严重影响模型的正常使用。完整性威胁是指攻击者通过诱导模型行为，使得在预测过程中模型输出为指定的分类标签，第三，可用性威胁主要用于阻止用户获得模型的正确输出或者干预用户获取模型的某些特征，使得用户获取的模型不再具备可靠性。

2.2　攻击者能力

对于恶意用户，可以通过全局模型参数模拟其他用户的训练样本，例如另外部署一个生成式对抗网络（Generative adversarial net，GAN）便能实现这种策略。与此同时，恶意用户还可以完全控制本地训练过程，继而修改模型超级参数（例如批量大小、epoch数量和学习速率）或本地模型更新（例如本地模型训练结果）。对于不可信的服务器，他们可能会根据每个参与者上传的参数推断出一些非预期的信息（例如梯度变化和真值标签）。此外，不受信任的服务器还可以与一部分恶意用户串通，实现对其他用户的细粒度隐私信息的窃取（例如特定用户的梯度）[30-31]。

2.3　攻击方式

目前，联邦学习中的攻击主要来自参与联邦学习过程的内部攻击者和独特的模型训练策略。首先，由于本地模型参数中隐含了用户的相关信息，一旦这些参数被发送到服务器进行联邦平均后，用户的敏感信息就很可能被泄露给服务器。其次，由于全局模型参数会进行共享，那么用户的私有信息也可能被泄露给其他联邦学习参与者。最后，由于联邦学习中的服务器不可访问用户的本地数据，对于整个学习过程便无法判定各用户所上传的本地更新是否是通过正确的执行学习协议而生成，伪造的本地更新便很难被察觉。联邦学习中主要的攻击类型如表1所示。

表1　联邦学习中的攻击类型

Table 1　Categorization of attacks on federated learning

攻击类型	攻击方式	攻击目标	理论方法
投毒攻击	标签翻转[18]	使训练后的模型偏离原始的预测边界	标签翻转
	后门[19]	使训练后的模型偏离原始的预测边界	插入触发器
基于GAN的攻击	类级[21]	从全局模型参数中模拟类级训练样本	经典GAN
	用户级[22]	从用户的更新和身份中模拟用户级的训练样本	多任务GAN
推理攻击	服务器端[32]	根据每个用户的更新推理特征表示	梯度推理
	用户端[33]	根据全局模型参数推理特征	属性推理

3　安全挑战

3.1　投毒攻击

中毒攻击在集中式学习场景的训练阶段中已取得到了不错的研究成果[18]。联邦学习框架内部的攻击者可以试图同修改、删除训练数据或者向训练数据中嵌入恶意数据，以达到破坏原始数据的初始分布和改变学习算法逻辑的目的。在安全挑战部分将主要介绍两种常见的中毒攻击的示例：标签翻转攻击[34]和后门攻击[19]。

3.1.1 标签翻转攻击

标签翻转攻击是指恶意用户通过翻转样本标签，将预先准备好的攻击点嵌入到训练数据中，便可使训练后的模型偏离既定的预测边界[33]。在联邦学习系统中利用标签翻转方法进行投毒攻击的过程如图1所示。主动攻击者首先在第t轮通信时下载全局模型参数来更新其本地模型，攻击者接着用标签翻转训练数据对已更新的本地模型进行训练并将训练后生成的本地模型参数上传至服务器。当服务器基于最新上传的伪本地模型参数完成联邦平均之后，全局模型将在随后的通信过程中受到被攻击者的破坏[36]。

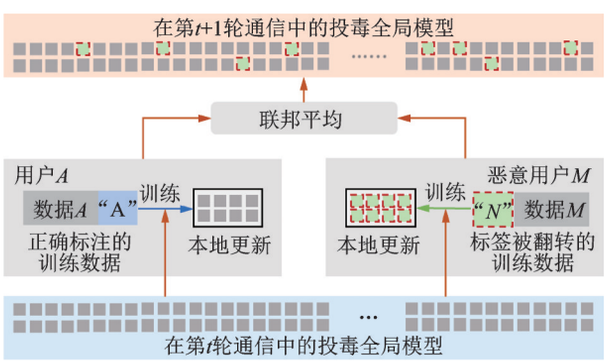

图1 在联邦学习中利用标签翻转进行投毒攻击

Fig.1 Poisoning attack with label-flipping in federated learning

由于攻击者是作为普通用户参与到整个联邦学习过程中的，他所进行的模型结构、全局参数等知识的获取都将被视为正常行为。与此同时，由于整个训练过程都是在本地执行的，服务器无法监督这些训练过程，因此很难检测检查攻击者的训练样本[37]。此外，该攻击者可以自适应地调整目标标签和本地超参数，实现最大限度地提高投毒攻击的效率。例如，可以通过引入本地更新的比例因子来放大攻击者的中毒参数，即有

$$\Delta L_{t+1}^{p}=S(L_{t+1}^{p}-\nabla_{L_t}\mathcal{L}(L_t,b)) \tag{3}$$

式中：L_{t+1}^{p}代表攻击者的中毒更新参数，S代表比例因子。

3.1.2 后门攻击

与标签翻转攻击不同，后门攻击需要攻击者在其精心设计的训练数据上使用一些特定的隐藏模式来训练目标深度神经网络(Deep neural network, DNN)模型。这些模式称为“后门触发器”，他们可以干预学习模型在预测阶段生成与真实情况大相径庭的结果[19]。例如图2所示的一个分类任务，攻击者的目标是标签“N”，后门触发模式存在于样本数据右下角的红色方块。在训练阶段结束后，后门模型将识别出那些携带触发器的样本，并将他们视为目标标签“N”，然而正常样本仍然可以被正确地分类。

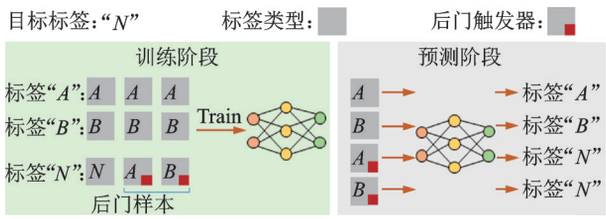

图2 后门攻击

Fig.2 An explanation of backdoor attack

对于联邦学习框架，攻击者可以利用后门数据对本地模型进行训练，并提交按比例缩小的训练结果以达到增强后门在全局模型中影响的目的[38]。经过模型平均后，全局模型对于后门样本的准确度将显著提高，而其他的主要任务(非后门样本)将不会受到影响。此外，后门攻击本质上不同于对抗性攻击[20]，因为对抗性攻击只能将攻击者选择的样本修改为错误分类的目标标签。因此，当这种修改被应用于其他样本时，这种攻击很难生效[39]。

基于以下原因，联邦学习框架面对这种投毒攻击时非常脆弱：(1)联邦学习系统中存在大量的参与者，很可能包含一个或多个恶意用户。(2)由于用户的本地数据和训练过程对服务器是不可见的，所以无法验证特定用户所上传更新的可靠性。(3)由于不同用户生成的本地更新可能会存在很大差异，给用户更新异常检测过程将带来巨大挑战。

3.2 用户端GAN攻击

Hitaj等[21]发现，联邦学习框架对于系统内部参与者发起的主动攻击是极为脆弱的。他们首次提出了一种由系统内恶意用户发起的基于GAN的重建攻击。在训练阶段，攻击者可以假扮成良性用户，训练一个GAN用于模拟其他用户训练数据产生的原型样本。通过不断注入伪造的训练样本，攻击能够逐步影响整个学习过程并且诱使受害者释放更多与攻击者目标类有关的敏感信息。

如图3所示，攻击者首先从服务器下载全局模型参数以更新其本地模型，紧接着通过创建一个新的本地模型的副本作为鉴别器(D)并在D上运行生成器(G)来实现对受害用户样本的模拟，GAN的工作原理为

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{4}$$

式中：$p_{\text{data}}(x)$代表原始数据分布；$p_z(z)$代表随机向量z的分布。生成的样本将被错误标记并输入

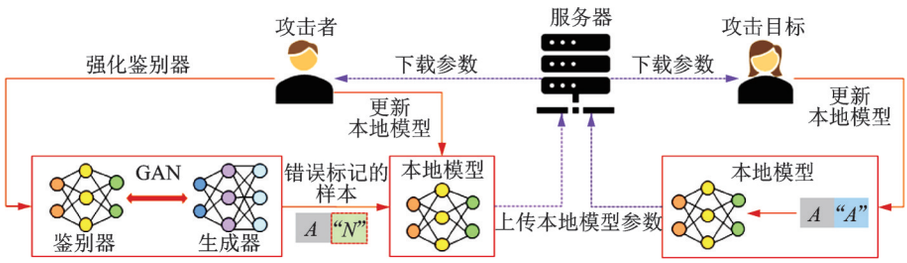

图 3　联邦学习中基于用户端的 GAN 攻击

Fig.3　User-side GAN-based attack in federated learning

本地模型达到更新全局模型参数的目的。通过这种方式，受害者将被强制对更多的样本进行本地训练，实现对正确与错误训练样本的区分，这对迭代式地改进鉴别器有很大的好处。用户端 GAN 攻击的强大之处在于，攻击者可以在不损害任何实体的前提下隐秘地完成所有的恶意行为，并可伪装成正常的参与者顺利地执行所建立的协议。

3.3　服务器端 GAN 攻击

服务器端 GAN 攻击依然暴露出了一些局限性。首先，错误标记的训练样本不仅会破坏全局模型，还会打破被攻击用户数据采样的平衡。其次，由于经过模型平均后错误标记样本的对抗性影响会大幅削弱，这种攻击在联邦学习中的效果将变得较差。第三，由于攻击者只能通过访问中央服务器来获取聚合和平均后的模型更新，所以这种攻击在类级样本重构阶段也同样受到了限制。

为了应对上述局限性，Wang 等引入了一种基于服务器端 GAN 攻击的方法用于推断用户级样本[22]。图 4 描述了联邦学习场景中基于 GAN 的服务器端攻击过程。恶意服务器首先假扮成正常的服务器为用户提供联邦学习服务，但其主要目的其实是为了重构特定攻击目标用户的训练样本。在获取到所有用户（其中包括攻击目标用户）的所有本地更新后，恶意服务器将攻击目标用户的本地更新添加到当前的全局模型以完成对识别器 D 的初始化过程，紧接着利用用户的本地更新来计算得到相应的数据样本。然后，恶意服务器通过将这些数据样本输入多任务鉴别器，进而识别出攻击目标用户的身份，并进逐步更新鉴别器 D 和生成器 G，最终实现服攻击目标用户所持有样本的再生重构。值得注意的是，用户身份的表征可以用来监督多任务 GAN 模型的训练阶段。

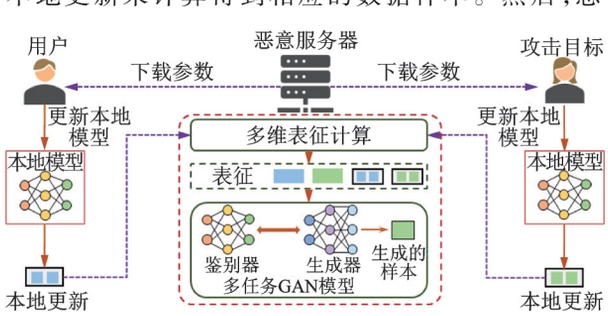

图 4　联邦学习中基于 GAN 的服务器端攻击

Fig.4　Server-side GAN-based attack in federated learning

3.4　推理攻击

正如上文所述，联邦学习机制要求所有参与者通过在本地数据集上训练全局模型用于上传梯度。在这种情况下，如果联邦学习系统中存在一个非可信且知识丰富的服务器，则无法保证用户的隐私数据信息。这种非可信服务器可以获取大量涉及每个参与者的局部训练模型的辅助知识（如模型结构、用户身份和梯度），并且有足够的能力进行用户隐私信息泄露。例如，Aono 等设计了一种服务器端推理攻击[32]，这种攻击通过使用周期性交换的模型参数来计算用户训练样本的隐私信息。然而，这种攻击仅限于单纯的训练设置，并要求共享模型是一个完全连接的网络，并且本地更新必须是通过单个样本训练生成的。

最近，Melis 等[33]指出用户的非预期信息很可能会被恶意用户推断得出。泄露的信息很可能被用于构建某些被动和主动的攻击，例如成员推理攻击。与服务器端推理攻击不同，恶意用户唯一拥有的知识是聚合生成的全局模型参数，因此如何在每一轮通信中获取其他用户的模型更新是重构数据样本的关键问题[40]。为此，攻击者首先在模型平均值之后获取全局模型参数的快照，并将这些快照保存在本地。然后通过计算后续快照之间的差异，并进一步去除新增的更新，实现从其他用户处获得聚合模型更新。通过这种方式，攻击者可以通过辅助数据集的协助来推断所有其他参与者共同合成的数据样本[41]。

首先，上文所提到的隐私威胁的影响是非常巨大的，因为攻击者只需要伪装成普通实体加入联邦学习系统，并秘密地执行恶意活动就可产生攻击效

果。其次,对于这两种基于GAN的攻击,很难区分生成的样本和来自同一类的训练输入。这是因为生成的样本只是在视觉上与目标训练数据相似,而无法用精确的数据体系。第三,攻击者在基于GAN的用户端攻击中,可以通过上传被覆盖的模型更新来强制攻击目标用户释放更多的敏感信息。然而,这种攻击方式只有在目标类的所有训练样本都属于攻击目标用户的情况下才能有效实施。

4 威胁对策

本节将从差分隐私(Differential privacy, DP)、同态密码系统(Homomorphic cryptosystem, HC)和安全多方聚合(Secure multi-party aggregation, SMA)3个角度,简要介绍几种具有可行性的策略,用于构建安全并可以保护用户隐私的联邦学习环境。本节所涉及到的解决方案如表2所示。

表2 联邦学习中的安全和隐私解决方案

Table 2 Security and privacy solutions for federated learning

涉及的问题	关键技术	可信模型
防御投毒攻击[42]	基于伪装K-means聚类方法识别恶意用户	恶意用户
保护训练数据[43]	构造差分隐私SGD	半信任服务器
防止信息泄露[44]	加法同态密码系统加密本地参数	半信任服务器
保护更新的参数[45]	基于双掩蔽和阈值秘密共享(Threshold secret sharing, TSS)方法进行安全多方聚合获取本地参数	半信任/恶意服务器和用户

4.1 安全防御方法

4.1.1 异常检测与对抗训练

Shen等提出了一种间接协作的深度学习框架[42],在此框架中用户不再向服务器传送模型梯度而是上传一种伪装特征,并设计了一种基于这些伪装特征的自动统计机制来抵御投毒攻击。在这种解决方案中,使用K-means算法对每轮通信过程中的本地模型更新(伪装特征)进行聚类,并识别出那些展现出异常分布的孤立点。这些检测到的孤立点可以用于进一步基于与攻击策略相关联的伪装特征对恶意用户进行识别[46]。

对抗性训练是一种主动防御技术,在这种防御技术的训练阶段模型就开始猜测对手攻击的所有排列,使机器学习模型对已知的对抗性攻击具有鲁棒性。文献[47]讨论了如何通过对抗训练的机器学习模型对攻击活动具有较强的鲁棒性。

4.1.2 知识蒸馏与数据清理

知识蒸馏是模型压缩技术的一种变体,在这种技术中,一个经过充分训练的神经网络根据需要将知识一步一步地传递到一个小型模型中。这种方式节省了训练模型所需的计算成本。共享知识的概念仅仅替代了在联邦学习中可利用的模型参数来提高客户端数据的安全性。

数据清理这是一种可以防御多种投毒攻击的主动防御技术。文献[48]介绍了如何实现数据清理防御技术来驱逐中毒攻击者。笔者在文献[49]的数据清理防御实验中指出,较强的数据投毒攻击有可能破坏数据清理防御的效果。

4.2 隐私保护技术

4.2.1 差分隐私保护

DP是工业界和学术界广泛使用的隐私保护技术。DP保护隐私的主要概念是给私有敏感属性添加相应的噪声。因此,每个用户的隐私都能受到保护。与此同时,与增加的隐私保护能力相比,为每个用户的附加噪声所造成的统计数据质量损失微不足道。在联邦学习中,为了避免逆向数据检索,引入DP向参与者上传的参数中添加噪声。

DP机制已广泛应用于数据发布系统,主要是通过在数据集中加入随机噪声(如拉普拉斯噪声或高斯噪声),将数据查询操作的实际结果隐藏起来。在深度学习的背景下,DP可以作为本地隐私解决方案来保护用户梯度的私密性。Abadi等[43]提出了一种将DP机制与SGD算法相结合的隐私保护深度学习方法,该方法主要是通过在小批量步骤后利用噪声干扰本地梯度实现隐私保护。值得关注的是,隐私保护预算开销和联邦学习效率之间的平衡是很难抉择的,这是由于较高的隐私保护开销预算可能对一些大规模攻击活动(如基于GAN的攻击)并没有很大作用[50],然而较低的隐私保护开销又会阻碍本地模型的收敛。

4.2.2 安全多方计算

为了保证多个参与者在共同计算一个模型或函数时的输入数据的具有安全性,安全多方计算(Secure multi-party computation, SMC)的概念应运而生。安全多方计算中多个参与者间的通信是具有安全性的,并且通过加密方法加以保护。最近,安全多方计算也被用来保护客户端所上传的

更新数据的安全。与传统的安全多方计算不同，联邦学习算法只需对参数进行加密，而不需要大量的输入，大大提高了计算效率。这种性能特点使得安全多方计算成为联邦学习环境下的首选技术。

Aono 等[44]指出，在协同学习中，用户即使上传很少的本地参数，其私有数据信息也很有可能被非可信服务器秘密的窃取。为了解决这一问题，笔者提出了一种新颖的隐私保护深度学习系统来抵御非可信服务器发起的梯度推理攻击。在中央服务器挖掘客户端上传的更新信息被泄漏的可能性，将加密与异步随机梯度下降相结合，有效地防止了客户端数据在中央服务器泄漏。该系统利用基于 HC 的预协商公钥对本地训练过程生成的局部梯度进行加密。服务器收到密文后，利用加法同态属性计算统计值的总和，实现了更新聚合和模型更新[51]。在完成模型平均之后，参与者使用各自持有的私钥对全局模型参数进行解密。通过对客户端更新进行加密可以确保没有隐私信息被泄漏。但是加密技术在更大规模环境中的使用成本是很高的，并且可能会影响机器学习模型的效率。

Bonawitz 等提出了一种实用的安全聚合协议[45]，该协议允许非可信服务器计算各方参与者的高维数据矢量求和的结果，这种方法适用于联邦学习参数设置时聚合多个用户的私有梯度向量的过程。在这种方案中，每对用户共享一个由安全伪随机生成器生成的种子来计算他们之间的伪装梯度向量，这策略意味着一个用户在他的私有向量上添加掩码，另一个用户可以使用相同的种子去剔除这些掩码。通过这种方式，服务器可以通过对接收到的梯度向量执行求和操作来消除所有掩码。此外，该协议还考虑了用户退出场景，设计了一个 TSS 方法来解密平均更新[52-53]。

总体而言，可以从以下两个方面控制安全与隐私的威胁：结合安全方法或者改变学习策略。例如，将 DP，HC，SMA 方法与联邦学习相结合，可以保证用户训练数据的安全性和私密性。然而，必须考虑安全机制在联邦学习过程产生的负面影响，例如 DP 的隐私开销、加密系统的计算复杂度、多方聚合的通信开销等因素。在今后的研究过程中，可以设计一个隐藏的机制来阻止攻击者对其他参与者的学习结果进行估计，这是由于隐私泄露取决于对目标类较高的模型准确性。此外，应深入研究身份验证和本地训练完整性机制，实现对每个参与者可信度的验证。

在联邦学习安全和隐私问题的探索以及未来研究方向上，本文针对投毒攻击类型，提出了基于对抗生成网络的数据生成方法，进一步实现用户端的主动式投毒[33,35]。该方法能够最大限度地降低传统投毒攻击方法对攻击者的限制条件。其次，在用户训练数据的隐私保护方面，本文提出了基于差分隐私和同态加密的强隐私保护联邦学习算法[49-51]。其中，基于差分隐私的方法可以支持在边缘计算环境下的计算外包，进而实现轻量级模型训练。在目前的研究基础上，本文进一步给出以下 3 点未来研究意见：(1)可信与可追溯性。联邦学习面临的其中一个主要挑战是在底层机器学习过程的整个生命周期中跟踪全局机器学习模型。例如，如果一个预测值在全局机器学习模型中发生了变化，需要具备后向跟踪能力用于识别哪些客户端的聚合值导致了这一变化[54-55]。如果机器学习模型行为下隐含的逻辑无法被掌握，将失去对逻辑现实的控制，从而盲目依赖于人工智能。(2)隐私保护与开销。目前的研究工作表明了如何在牺牲效率或准确性的前提下强化联邦学习中的隐私保护的能力[56]。然而，目前还没有研究成果表明应当为安全多方计算增添何种级别的加密算法和何种规模的噪声数据是相对合适的。如果加密级别或噪声规模不足，联邦学习参与者仍然面临着隐私泄露的风险。相反，如果加密级别太高或参数中添加了太多的噪声，则联邦学习模型的精度会急剧下降。(3)构建高效隐私保护联邦学习框架。目前已有一些联邦学习框架可以用来实现基于联邦学习的系统，比如 TensorFlow federed 和 FATE[57-58]。但是，目前还没有能够集成库或工具箱的联邦学习框架用于执行安全多方计算或差分隐私的应用。因此，开发联邦学习隐私保护增强框架是目前亟待解决的研究方向，倘若能实现这种框架不但有利于学术研究，而且也有利于联邦学习在工业界的广泛应用。

5　结　　论

本文重点讨论了联邦学习中的安全和隐私挑战，揭示了在参与者和服务器之间共享模型参数的独特特性可能会带来前所未有的安全和隐私挑战。本文还介绍了由内部攻击者发起的 3 种不同攻击类型，并指出了这些攻击能够成功构建的原因，还总结了近年来在这一领域已有的防御对策，为构建一个安全、隐私的联邦学习系统提供了新的研究方向。

参考文献：

[1] RIBEIRO M, GROLINGER K, CAPRETZ M A

M. MLaaS: Machine learning as a service[C]//Proceedings of 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). [S. l.]: IEEE, 2015: 896-902.

[2] GE Z, SONG Z, DING S X, et al. Data mining and analytics in the process industry: The role of machine learning[J]. IEEE Access, 2017, 5: 20590-20616.

[3] WARING J, LINDVALL C, UMETON R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare[J]. Artificial Intelligence in Medicine, 2020, 104: 101822.

[4] LOPEZ K L, GAGNE C, GARDNER M A. Demand-side management using deep learning for smart charging of electric vehicles[J]. IEEE Transactions on Smart Grid, 2018, 10(3): 2683-2691.

[5] LIN W Y, HU Y H, TSAI C F. Machine learning in financial crisis prediction: A survey[J]. IEEE Transactions on Systems Man and Cybernetics, 2012, 42(4): 421-436.

[6] ZHOU L, PAN S, WANG J, et al. Machine learning on big data: Opportunities and challenges[J]. Neurocomputing, 2017, 237(10): 350-361.

[7] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning[J]. arXiv, 2016, 16: 11-19.

[8] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems, 2019, 10(2): 12.1-12.19.

[9] BARRENO M, NELSON B, JOSEPH A D, et al. The security of machine learning[J]. Machine Learning, 2010, 81(2): 121-148.

[10] HUNT T, SONG C, SHOKRI R, et al. Privacy-preserving machine learning as a service[J]. Proceedings on Privacy Enhancing Technologies, 2018(3): 123-142.

[11] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver, CO, USA: ACM, 2015: 1310-1321.

[12] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of AISTATS. Fort Lauderdale, USA: JMLR, 2017: 1-10.

[13] KONEN J, MCMAHAN H B, YU F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv, 2016, 16: 1-10.

[14] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge[C]//Proceedings of ICC. Shanghai, China: IEEE, 2019: 1-7.

[15] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[16] WANG S, TUOR T, SALONIDIS T, et al. When edge meets learning: Adaptive control for resource-constrained distributed machine learning[C]//Proceedings of INFOCOM. Paris, France: IEEE, 2019: 63-71.

[17] TRAN N H, BAO W, ZOMAYA A, et al. Federated learning over wireless networks: Optimization model design and analysis[C]//Proceedings of INFOCOM. Paris, France: IEEE, 2019: 1387-1395.

[18] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning[C]//Proceedings of 2018 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2018: 19-35.

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

[20] YUAN X, HE P, ZHU Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[21] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: Information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, TX, USA: ACM, 2017: 603-618.

[22] WANG Z, SONG M, ZHANG Z, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning[C]//Proceedings of IEEE INFOCOM Conference on Computer Communications. Paris, France: IEEE, 2019: 2512-2520.

[23] YANG K, JIANG T, SHI Y, et al. Federated learning via over-the-air computation[J]. IEEE Transactions on Wireless Communications, 2020, 19(3): 2022-2035.

[24] WANG X, HAN Y, WANG C, et al. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning[J]. IEEE Network, 2019, 33(5): 156-165.

[25] KIM H, PARK J, BENNIS M, et al. Blockchained on-device federated learning[J]. IEEE Communications Letters, 2019, 24(6): 1279-1283.

[26] SAMARAKOON S, BENNIS M, SAADY W, et al. Distributed federated learning for ultra-reliable low-latency vehicular communications[J]. IEEE Transactions on Communications, 2019, 68(2): 1146-1159.

[27] TRUEX S, BARACALDO N, ANWAR A, et al. A hybrid approach to privacy-preserving federated learning[C]//Proceeding of the 12th ACM Workshop on Artificial Intelligence and Security. London, UK: ACM, 2019: 1-11.

[28] ROBIN C G, TASSILO K, MOIN N. Differentially private federated learning: A client level perspective[C]//Proceedings of NIPS. Long Beach, CA, USA: MIT Press, 2017.

[29] XU G, LI H, LIU S, et al. VerifyNet: Secure and verifiable federated learning[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 911-926.

[30] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031-2063.

[31] LU Y, HUANG X, DAI Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186.

[32] LE T P, AONO Y, HAYASHI T. Privacy-preserving deep learning: Revisited and enhanced[C]//Proceedings of International Conference on Applications and Techniques in Information Security. Singapore: [s.n.], 2017: 100-110.

[33] MELIS L, SONG C, de CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2019: 691-706.

[34] XIAO H, BIGGIO B, NELSON B, et al. Support vector machines under adversarial label contamination[J]. Neurocomputing, 2015, 160: 53-62.

[35] ZHANG J, CHEN J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets[C]//Proceedings of IEEE Trustcom. Rotorua, New Zealand: IEEE, 2019: 374-380.

[36] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]//Proceedings of ICML. Long Beach California, USA: ACM, 2019: 634-643.

[37] ZHAO Y, CHEN J, ZHANG J, et al. PDGAN: A novel poisoning defense method in federated learning using generative adversarial network[C]//Proceedings of ICA3PP. Melbourne, VIC, Australia: IEEE, 2019: 595-609.

[38] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]//Proceedings of AISTATS. Palermo, Sicily, Italy: JMLR, 2020: 2938-2948.

[39] XIE C, HUANG K, CHEN PY, LI B. DBA: Distributed backdoor attacks against federated learning[C]//Proceedings of ICLR. Addis Ababa: IEEE, 2020.

[40] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning[C]//Proceeding of IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2019: 739-753.

[41] TRUEX S, LIU L, GURSOY M E, et al. Demystifying membership inference attacks in machine learning as a service[J]. IEEE Transactions on Services Computing, 2019, 1: 1.

[42] SHEN S, TOPLE S, SAXENA P. Auror: Defending against poisoning attacks in collaborative deep learning systems[C]//Proceedings of the 32nd Annual Conference on Computer Security Applications. Los Angeles, CA, USA: IEEE, 2016: 508-519.

[43] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM, 2016: 308-318.

[44] AONO Y, HAYASHI T, WANG L, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345.

[45] 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. Dallas, TX, USA: ACM, 2017: 1175-1191.

[46] CAO D, CHANG S, LIN Z, et al. Understanding distributed poisoning attack in federated learning[C]//Proceeding of the 25th International Conference on Parallel and Distributed Systems (ICPADS). Tianjing, China: IEEE, 2019: 233-239.

[47] FLORIAN T, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and Defenses[EB/OL].(2017-05-07)[2020-03-10]. https://arxiv.org/abs/1705.07204.

[48] CRETU G F, STAVROU A, LOCASTO M E, et al. Casting out demons: Sanitizing training data for anomaly sensors[C]//Proceedings of the 2008 IEEE Symposium on Security and Privacy. Oakland Califor-

nia, USA: IEEE, 2008: 81-95.

[49] ZHANG J, ZHAO Y, WU J, et al. LVPDA: A lightweight and verifiable privacy-preserving data aggregation scheme for edge-enabled IoT[J]. IEEE Internet of Things Journal, 2020, 7(5): 4016-4027.

[50] ZHANG J, ZHAO Y, WANG J, et al. FedMEC: Improving efficiency of differentially private federated learning via mobile edge computing[J]. Mobile Networks and Applications, 2020, 1: 13.

[51] ZHANG J, CHEN B, YU S, et al. PEFL: A privacy-enhanced federated learning scheme for big data analytics[C]//Proceedings of 2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa, HI, USA: IEEE, 2019: 1-6.

[52] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for federated learning on user-held data[C]//Proceedings of NIPS. Barcelona, Spain: MIT Press, 2016.

[53] SATTLER F, WIEDEMANN S, MÜLLE K R, et al. Robust and communication-efficient federated learning from non-iid data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(9): 3400-3413.

[54] MOWLA N I, TRAN N H, DOH I, et al. Federated learning-based cognitive detection of jamming attack in flying AD-HOC network[J]. IEEE Access, 2020, 8: 4338-4350.

[55] HUANG X, DING Y, JIANG Z L, et al. DP-FL: A novel differentially private federated learning framework for the unbalanced data[J]. World Wide Web, 2020, 23: 2529-2545.

[56] ZHAO R, YIN Y, SHI Y, et al. Intelligent intrusion detection based on federated learning aided long short-term memory[J]. Physical Communication, 2020, 42: 101157.

[57] KAISSIS G A, MAKOWSKI M R, DANIEL R, et al. Secure, privacy-preserving and federated machine learning in medical imaging[J]. Nature Machine Intelligence, 2020, 2: 305-311.

[58] CHEN H, LI H, XU G, et al. Achieving privacy-preserving federated learning with irrelevant updates over e-health applications[C]//Proceedings of ICC 2020 - 2020 IEEE International Conference on Communications (ICC). Dublin, Ireland: IEEE, 2020.

(编辑:刘彦东)