联邦学习安全与隐私保护研究综述

周　俊,方国英,吴　楠
(华东师范大学上海市高可信计算重点实验室, 上海 200062)

摘　要：数据孤岛以及模型训练和应用过程中的隐私泄露是当下阻碍人工智能技术发展的主要难题。联邦学习作为一种高效的隐私保护手段应运而生。联邦学习是一种分布式的机器学习方法,以在不直接获取数据源的基础上,通过参与方的本地训练与参数传递,训练出一个无损的学习模型。但联邦学习中也存在较多的安全隐患。本文着重分析了联邦学习中的投毒攻击、对抗攻击以及隐私泄露三种主要的安全威胁,针对性地总结了最新的防御措施,并提出了相应的解决思路。

关键词：联邦学习;投毒攻击;对抗攻击;隐私泄露

中图分类号：TP181;TP309　　**文献标志码**：A　　**文章编号**：1673-159X(2020)04－0009－09

doi:10.12198/j.issn.1673－159X.3607

Survey on Security and Privacy-preserving in Federated Learning

ZHOU Jun, FANG Guoying, WU Nan
(*Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai* 200062 *China*)

Abstract: The issue of data island has always been a difficult problem during the development of artificial intelligence. The risk of privacy disclosure in model training and application further impedes the development of artificial intelligence technology. Federated learning, emerging as an efficient means of privacy protection, is a distributed machine learning technique, which enables to train a lossless learning model through local training and parameter transfer of participants without directly obtaining data sources. However, study results show that there are still many security risks in federated learning. Aiming at the security problems in federated learning, this paper analyzes three main security threats, including poisoning attacks, adversarial attacks and privacy disclosure, and summarizes the latest defense measures. Finally, this paper discusses the security issues still existing in the current federated learning with related solutions.

Keywords: federated learning; poisoning attack; adversarial attack; privacy leakage

联邦学习(federated learning, FL)在 2016 年由谷歌最先提出[1],用于建立移动终端与服务器之间的共享模型,从而在大规模数据背景下有效地利用这些数据资源,并且保证用户的隐私安全。但这些分散的数据大多是异构且不均衡的,为此,Jakub 等[2]提出一个实用高效的优化算法来处理数据分布问

收稿日期:2020－02－15
基金项目:国家自然科学基金项目(61602180、U1636216);上海市自然科学基金项目(16ZR1409200)。
第一作者:周俊(1982—),男,副教授,主要研究方向为外包系统安全与隐私保护、安全多方计算、人工智能安全与区块链隐私保护等。
ORCID:0000－0003－3294－9774　　E-mail: jzhou@sei.ecnu.edu.cn

引用格式:周俊,方国英,吴楠. 联邦学习安全与隐私保护研究综述[J]. 西华大学学报(自然科学版),2020,39(4):9－17.
ZHOU Jun, FANG Guoying, WU Nan. Survey on Security and Privacy-preserving in Federated Learning[J]. Journal of Xihua University(Natural Science Edition), 2020, 39(4): 9－17.

题。之后,又有大量的研究来进一步优化联邦学习模型,如文献 [3] 提出了两种方法来减小通信消耗,从而实现更加高效的训练过程;文献 [4] 解决了之前联邦学习机制中共享模型可能会偏向于某些参与方的问题,保证了参与方间的公平性;文献 [5] 提出单样本/少样本探索式的学习方法来解决压缩式联邦学习中的通信问题。

联邦学习一经推出,就受到广泛的关注。各大科技金融龙头也开始进行开源项目的搭建,如 WeBank 开发的 FATE、Google 推出的 TensorFlow Federated(TFF)、Uber 开源的 Horovod等。联邦学习已经被广泛应用于无线通信与边缘计算[6]、智慧金融[7]、智慧医疗[8]、环境保护[9] 等领域,未来有望改变新时代的商业模式,深入影响到智能城市的建设。

然而,联邦学习中仍然存在巨大的安全隐患,比如参与方的安全等级较低,容易遭受恶意攻击,从而影响到整个模型的安全。本文针对联邦学习可能产生的安全问题进行分析,着重针对投毒攻击、对抗攻击以及隐私泄露这三个方面的安全威胁进行详细的说明,并有针对性地总结了防御措施,以期对减小联邦学习的安全性风险、促进其进一步发展与普及有一定帮助。

1 联邦学习概述

联邦学习是一种分布式的机器学习方法,即参与方对本地数据进行训练后将更新的参数上传至服务器,再由服务器进行聚合得到总体参数的学习方法。与传统机器学习技术相比,联邦学习不仅可以提高学习效率,还能解决数据孤岛问题,保护本地数据隐私[10]。

1.1 联邦学习的定义

假设有 n 个参与方$U_1,U_2,\cdots U_n$,每个参与方U_i拥有各自的本地数据集D_i,现在需要在总的数据集$D=D_1\cup D_2\cup\cdots\cup D_n$中训练出模型$M_{Global}$。联邦学习指的是一种分布式的学习方式,即不直接把所有数据整合在一起统一进行训练得到模型M_{Sum},而是由各个参与方U_i根据服务器传过来的初始参数w_G,各自训练本地的数据,得到新的参数w_{Gi}',再将更新的参数值$\delta_i=w_{Gi}'-w_G$传到服务器端,服务器端采取一定的方式进行聚合,得到更新的总体参数

$$w_G'=w_G+f(\delta_1,\delta_2,\cdots,\delta_n)$$

由此经过多次迭代,最终得到总体训练模型M_{Fed}。此外,联邦学习需要能够保证模型M_{Fed}的效果V_{Fed}与模型M_{Sum}的效果V_{Sum}间的差距足够小[11],即

$$|V_{Fed}-V_{Sum}|<\varepsilon$$

其中:ε为任意小的正量值。

1.2 联邦学习的分类

联邦学习中各个参与方只需要维护本地的数据集D_i。但不同情况下,D_i之间用户和数据特征的差异也不尽相同。如表 1 所示,根据数据分布的不同情况,联邦学习大致分为 3 类:横向联邦学习、纵向联邦学习与联邦迁移学习。

表 1 三类联邦学习的对比

种类	用户重叠	数据特征重叠	训练方法
横向联邦学习	多	少	按用户维度切分
纵向联邦学习	少	多	按数据特征维度切分
联邦迁移学习	少	少	迁移学习

1.2.1 横向联邦学习

横向联邦学习指的是在不同数据集之间数据特征重叠较多而用户重叠较少的情况下,按照用户维度对数据集进行切分,并取出双方数据特征相同而用户不完全相同的那部分数据进行训练。

1.2.2 纵向联邦学习

纵向联邦学习指的是在不同数据集之间用户重叠较多而数据特征重叠较少的情况下,按照数据特征维度对数据集进行切分,并取出双方针对相同用户而数据特征不完全相同的那部分数据进行训练。

1.2.3 联邦迁移学习

联邦迁移学习指的是在多个数据集的用户与数据特征重叠都较少的情况下,不对数据进行切分,而是利用迁移学习[12] 来克服数据或标签不足的情况。

1.3 联邦学习的优势

与其他机器学习技术相比,联邦学习具有多重优势。

1)用户隐私保护。联邦学习数据只存储在本地,各参与方数据不共享,保证了用户数据的隐私,满足了《通用数据保护条例》[13] 的要求。

2)适应大规模数据的模型训练。大规模的训练数据可以提高训练模型的质量。采用联邦学习可以保证训练出的模型效果无损,同时可以减小对训练过程中的设备要求,提高模型训练速度。

3)增强了数据来源的灵活性。在联邦学习的技术支持下,一些原本因为特定因素无法参与训练的数据源,可以将数据存放在本地的同时参与总体模型的训练,更好地提升模型的泛化效果。

2　联邦学习中的安全问题

尽管联邦学习的优势明显,其出现和发展顺应时代的潮流,但在投入应用前应检测其安全性。近年来,大量研究成果表明,联邦学习机制中仍然存在安全问题,如投毒攻击,对抗样本攻击以及隐私泄露问题等。本节主要针对这三个安全问题进行详细说明。

2.1　投毒攻击

投毒攻击主要是指在训练或再训练过程中,恶意的参与者通过攻击训练数据集来操纵机器学习模型的预测[14]。联邦学习中,攻击者有两种方式进行投毒攻击:数据投毒和模型投毒,如图 1 所示。

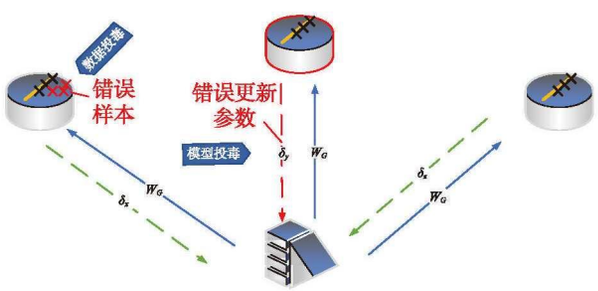

图 1　数据投毒与模型投毒

2.1.1　数据投毒

数据投毒是指攻击者通过对训练集中的样本进行污染,如添加错误的标签或有偏差的数据,降低数据的质量,从而影响最后训练出来的模型,破坏其可用性或完整性。文献 [15] 中提出了一种攻击方式,攻击者使学习模型的参数值接近他所期望的值,同时使模型输出对某些测试样本的错误预测。文献 [16] 采用混合辅助注入策略,通过注入少量有毒样本到训练集就获得了 90% 以上的攻击成功率。文献 [17] 中针对支持向量机算法(support vector machines , SVM)产生的优化梯度,预测其目标函数的变化方向,使用梯度上升策略显著提高了 SVM 分类器的错误率。为了提高攻击广度,文献 [18] 提出了一种基于反梯度优化思想的新型投毒算法,能够针对更广泛的学习算法中基于梯度的训练过程,包括神经网络(neural network, NN)和深度学习(deep learning, DL)体系结构。

2.1.2　模型投毒

模型投毒不同于数据投毒,攻击者不直接对训练数据进行操作,而是发送错误的参数或损坏的模型来破坏全局聚合期间的学习过程[19],比如控制某些参与方U_i传给服务器的更新参数δ_i,从而影响整个学习模型参数的变化方向,减慢模型的收敛速度,甚至破坏整体模型的正确性,严重影响模型的性能。文献 [20] 只假设了一个恶意代理(参与方),就实现了对整体模型的隐蔽性的攻击,使得目标模型无法对某类数据正确分类。

2.2　对抗攻击

对抗攻击是指恶意构造输入样本,导致模型以高置信度输出错误结果。这种通过在原始样本中添加扰动而产生的输入样本称为对抗样本[14]。

对抗攻击首先是由 Christian 等[21] 提出的,他们发现深度学习的输入与输出之间映射的不连续性,通过对图片进行一个细微的干扰,神经网络分类器就会完全改变对于这张图片的预测。文献 [22] 进一步研究发现,对抗攻击不仅能对复杂的深度学习网络起作用,而且在线性模型这样简单的网络中,对抗攻击也可以有很好的攻击效果。之后大量的研究探索出了多种对抗攻击的攻击方式,如 Least-Likely-Class Iterative Methods[23]、Jacobian-based Saliency Map Attack (JSMA)[24]、DeepFool[25] 等。

从攻击环境来说,对抗攻击可以分为黑盒攻击和白盒攻击。若知道机器学习模型中的参数与内部结构,攻击者可以把所需的干扰看作一个优化问题计算出来。这种情况下的对抗攻击属于白盒攻击。而另一种常见的情境下,攻击者不知道任何模型的信息,只能跟模型互动,给模型提供输入然后观察它的输出,这种情形下的对抗攻击属于黑盒攻

击。对抗攻击还可以根据攻击目的分为目标攻击和非目标攻击。根据干扰的强度大小分为无穷范数攻击、二范数攻击和零范数攻击等。

对抗攻击可以帮助恶意软件逃避检测,生成投毒样本,已经被攻击者广泛应用于图像分类、语义分割、机器识别以及图结构等多个领域,成为系统破坏者的一个有力攻击武器。

2.3 隐私泄露

联邦学习方式允许参与方在本地进行数据训练,各参与方之间是独立进行的,其他实体无法直接获取本地数据,可以保证一定的隐私安全,但这种安全并不是绝对安全,仍存在隐私泄露的风险。比如恶意的参与方可以从共享的参数中推理出其他参与方的敏感信息。

参与方的隐私安全通常易受到两种攻击:模型提取攻击和模型逆向攻击[14]。通过模型提取攻击,攻击者试图窃取模型的参数和超参数,破坏模型的保密性。比如恶意的参与方可以对共享模型进行预测查询,然后提取训练完成的模型。文献 [26] 针对 BigML 和 Amazon 机器学习在线服务进行了攻击,提取了一个几乎完全相同的模型,并且证明了同样的攻击适用于多种机器学习方法。通过模型逆向攻击,攻击者试图从训练完成的模型中获取训练数据集的统计信息,从而获取用户的隐私信息。文献 [27] 实现了一个攻击,可以推断模型构建过程中所使用的流量类型。模型逆向攻击推断出的训练集的信息,既可以是某个成员是否包含在训练集中,也可以是训练集的一些统计特性。根据这两种训练集信息,模型逆向攻击可以进一步分为成员推理攻击和属性推理攻击。这对联邦学习中的各参与方的隐私造成了严重的威胁。

我们通常假设服务器是可信的,然而实际情况中并非如此,若服务器是恶意的(或者是诚实并好奇的),它可以识别更新的参数的来源,甚至进一步通过参与方多次反馈的参数推测参与方的数据集信息,这可能造成参与方的隐私泄露。

3 防御措施

针对联邦学习面临的多重安全威胁,本节讨论防御上述攻击的一些最新对策。

3.1 投毒攻击防御

联邦学习中的投毒防御主要从数据投毒防御和模型投毒防御两个方向考虑。

3.1.1 数据投毒防御

针对数据投毒,防御方法应从保护数据的角度出发。一方面,在训练模型之前应当保证数据来源的真实性与可靠性。另一方面,在使用不能保证安全性的数据之前,应当进行相应的检测以保证数据完整性不受篡改。

为保证数据源的真实与可靠,在与各参与方进行数据交互之前,可以使用健壮的身份验证机制,以防止欺骗攻击或将被攻占的节点中被污染的数据集加入训练集,从而降低数据的质量。

目前已经有多种防御机制来抵抗数据投毒的攻击。Nathalie 等[28] 使用起源和转换等上下文信息来检测训练集中的有毒样本点。该检测方法通过将整个训练集分为多部分,比较各部分数据训练出的效果,从而识别出哪一部分的数据表现最为异常,实验证明该方法能达到较高的检测率。文献 [29] 提出了一种防御机制来对抗回归中的投毒攻击,该技术集成了改进的鲁棒低秩矩阵逼近和鲁棒主成分回归,提供了强大的性能保证。

3.1.2 模型投毒防御

针对模型投毒,假定服务器是可信的,那么防御的重点在于对恶意参与方的识别以及对错误更新参数的检测。恶意参与方也可以用相关的身份管理技术进行防范。对于异常的更新参数,通常有两种检测方法[20]。一种是通过准确度检测。服务器利用参与方U_i返回的参数δ_i计算$w_{G1}' = w_G + f(\delta_i)$,利用其他参与方返回的参数计算$w_{G2}' = w_G + f(\Delta)$,其中$\Delta = \{\delta_j | j = 1,2,\cdots,n, j \neq i\}$。然后分别使用$w_{G1}'$和$w_{G2}'$作为模型的权重参数,比较两个模型在验证集上的准确度。若使用w_{G1}'的模型准确度明显小于使用w_{G2}'的模型,则推测δ_i异常。另一种方法是通过直接比较各个参与方提交的更新参数$\delta_1,\delta_2,\cdots,\delta_n$之间的数值统计差异,当某个参与方反馈的更新参数$\delta_i$与其他参与方的有很大的统计差异时,则推测$\delta_i$异常。

3.2 对抗攻击防御

在机器学习领域中,研究了大量对抗攻击防御

机制,这些机制也同样适用于联邦学习的对抗防御。

3.2.1　对抗训练

一个常用的防御手段是进行对抗训练,即将真实的样本和对抗样本一起作为训练集,来训练出最后的模型。对抗训练适用于多种监督问题[30],它可以使得模型在训练过程中就学习到对抗样本的特征,提高模型的健壮性。但这样的模型只能抵抗训练集中的对抗样本,不能很好地防范未知的攻击。

3.2.2　数据增强

数据增强是对抗攻击的一种扩充。在训练过程中不可能穷举所有对抗样本,但通过对原始数据集中的数据进行随机化处理可以增强模型的泛化能力。比如在图像处理中对训练集中的图片进行翻转、旋转、缩放比例、裁剪、移位以及颜色等处理,而且适度加入噪声也是一种常用的方法。文献 [31] 中对每个原始样本加入高斯噪声,生成了 10 个噪声样本,取得了较好的防御效果。

3.2.3　数据处理

数据处理采取与数据增强不同的方式,数据处理技术是指对样本进行降噪处理,以减小对抗样本的干扰。文献 [32] 中引入标量量化和平滑空间滤波两种经典的图像处理技术来降低噪声的影响。并且利用图像熵作为度量指标,实现了对不同类型图像的自适应降噪。通过比较给定样本的分类结果及其去噪后的版本,这种降噪处理方法可以有效地检测和剔除对抗样本,在 F1 度量标准下达到 96.39% 的准确度。

3.2.4　数据压缩

数据压缩是一种特殊的数据处理方法,专门针对图像训练过程,即使用压缩后的图片进行训练。文献 [33] 中采用 PCA 降维压缩技术防御对抗样本攻击,在维度降至 50 时取得了最优的防御效果。但这样的处理方式在降低样本中噪声比例的同时,也会减小原始数据信息,所以压缩图像同时也会降低正常分类的准确率。

3.2.5　防御蒸馏

防御蒸馏的主要思想是先利用训练集得到一个模型,然后再通过模型提取,从原来的模型“蒸馏”提纯出另外一个模型,从而降低模型的复杂度。文献 [34] 对防御蒸馏技术的有效性进行了实证研究,发现防御蒸馏可使在 MNIST 数据集上的对抗攻击成功率从 95% 降低到 0.5% 以下,在 CIFAR10 数据集上也将攻击成功率降到了 5% 以下,而且没有对训练过程造成过多的干预,保证了模型训练的效率与质量。

3.2.6　梯度正则化

模型训练中常使用正则化来防止过拟合,即过度学习样本特征。若模型过拟合程度越高,其泛化能力越弱,越容易遭受到对抗样本的攻击。梯度正则化是指在训练模型的目标函数上对输入与输出的变化进行惩罚,从而限制了输入的扰动对于预测结果的影响。文献 [35] 使用梯度正则化来防御 FGSM 和 TGSM 生成的对抗样本,证明了梯度正则化技术能提高对抗攻击鲁棒性,且相比对抗攻击和防御蒸馏,梯度正则化的防御效果更好。

3.2.7　对抗样本检测

对抗样本检测也是一种常用的防御措施。若能区分出对抗样本与正常样本的不同之处,然后以较高精度检测出对抗样本,就能较好地防范对抗攻击。文献 [36] 中发现对抗样本的局部本征维数(local intrinsic dimensionality, LID)与正常样本差异较大,LID 根据样本到它的邻居样本间的距离分布,评估其周围区域的空间填充能力。文章利用 LID 对五种攻击策略进行防御,证明了该技术的检测率大大超出几种最先进的检测措施。

3.2.8　基于 GAN 的防御

生成式对抗网络(generative adversarial net, GAN)是一种机器学习模型,由两个模块组成。一个是生成模块 G,利用接收到的随机噪声生成虚假样本,另一个是判别模块 D,用以判断出某样本是否为 G 生成的虚假样本。文献 [37] 使用基于 APE-GAN 的生成式对抗网的有效框架来防御对抗攻击。其中 G 被训练成更改输入样本中的微小扰动,而 D 被用来分隔真实的样本与经过 G 处理的去除掉扰动的对抗样本。该技术在 MNIST、CIFAR10 和 ImageNet 三种数据集上的实验结果表明,APE-GAN 能够有效地抵抗对抗攻击。

3.3　隐私泄露防御

联邦学习中的隐私保护主要从两大主体——

参与方与服务器的角度进行保证。同时对于训练完成的模型也要防止模型提取攻击和模型逆向攻击。

3.3.1 差分隐私

考虑恶意参与方与诚实服务器的情形。由于任何一个参与方都可以从训练过程中获取总体参数,联邦学习方式易受到差分攻击[38]。通过分析共享模型,其他诚实的参与方的数据隐私会受到威胁。在这种情况下,常采用差分隐私保护技术。

设有随机算法M,R为所有可能的输出构成的集合,若对于任意两个邻近数据集D和D'以及R的任意子集S,都有$\Pr[M(D)\in S]\leqslant e^{\varepsilon}\times\Pr[M(D')\in S]$,则称算法$M$满足$\varepsilon$-差分隐私。

其中,当ε越小,算法提供更高等级的隐私保护,但在一定程度上会降低准确性。基于这个技术,文献 [38] 提出了一种针对参与方差分隐私保护的联邦优化算法——差分隐私随机梯度下降算法,其目的是在模型训练阶段隐藏参与方的更新参数,从隐私损失和模型性能之间找到平衡。该技术将数据样本随机分成小部分,在聚合的过程中加入高斯噪声,实现差分隐私保护,同时也维持了模型的高性能。文献 [39] 结合了联邦学习的具体情境,切实保护各个参与方的数据集,且通过差分隐私保护技术保证训练完成的模型不会泄露某一参与方是否参与了数据训练过程,即一定程度上可以抵抗成员推理攻击。实验表明,在参与方足够多的情况下,该技术能够以较小的模型性能成本维持客户级差异隐私。文献 [40] 为了提高过于严格的本地差分隐私保护的实用性,重新定义了保护机制,既保证了敏感信息安全,又放宽了对数据的限制,并且设计了新的局部最优差异隐私机制来解决所有隐私级别的统计学习问题,适用于大型分布式模型拟合和联邦学习系统。

3.3.2 秘密共享机制

考虑诚实参与方与恶意服务器(或者诚实并好奇服务器)的情形。服务器在联邦学习中扮演重要角色,它可以获取各个明确身份的参与方反馈的参数,并从中推测出参与方的敏感信息,这将对参与方隐私造成威胁,可以使用秘密共享机制来进行防范。

(n,t)秘密共享是指将一个秘密信息s分成n个碎片,交由n个不同的参与方保管,使得其中任意t个或t个以上的碎片可以重构出秘密s,而当碎片数量少于t时无法获得任何关于s的有用信息。

文献 [41] 基于 Shamir 秘密共享设计了一个实用的安全聚合方案,该方案可以在诚实并好奇的服务器背景下保证更新参数安全性,即保证各参与方数据的隐私,同时控制协议的复杂度,使之能在大规模数据集中保持较低的计算和通信开销,适用于联邦学习中的协同训练。但这个协议无法防止共谋攻击。

3.3.3 同态加密

考虑诚实参与方与恶意服务器(或者诚实并好奇服务器)的情形。采用加密的数据传输方式来保障隐私安全是有效防御措施。同态加密技术是一种常用的防御手段。

同态加密是一种有效的加密方式,它的特性在于不需要直接访问明文,对密文的操作结果解密后等于明文的操作结果。以加性同态加密为例,即有

$$\mathrm{Enc}_{\mathrm{pk}}(m_1)=c_1,\mathrm{Enc}_{\mathrm{pk}}(m_2)=c_2$$
$$\mathrm{Dec}_{\mathrm{sk}}(c_1\circ c_2)=m_1+m_2$$

其中,加密方案采用公钥加法同态加密,(pk,sk)是一对公私钥,$\circ$表示密文上某种特定的运算,如乘法或加法运算。

利用同态密码,服务器就对密文参数进行聚合而无法获取用户的隐私参数。比如文献 [42] 基于诚实并好奇的云服务器提出了一个新的深度学习系统,利用同态加密方案实现了梯度在诚实并好奇服务器上的聚合,并且保证了系统达到与所有参与方联合数据集上训练的相应深度学习系统相同的精度。文献 [43] 开发了 CryptoDL,用近似多项式代替原激活函数训练卷积神经网络,实验证明该技术在 MNIST 数据集的准确率高达到 99.52%,每小时可以做出接近 164 000 个预测,提供了一个高效准确的隐私保护方案。

3.3.4 混合防御机制

考虑恶意参与方与恶意服务器(或者诚实并好奇服务器)的情形。为了同时对参与方和服务器进行防范,可以将多种防御技术结合起来。文献 [44] 将差分隐私保护技术与同态密码相结合,参与方

利用初始参数计算出δ_i后,先加入噪声使之满足ε–差分隐私,然后再使用轻量级的同态加密方案进行加密,这样可以防范服务器与恶意参与者的勾结问题。

综上,对联邦学习中的三类安全威胁及其防御措施进行总结,见表 2。

表 2　联邦学习中三类安全威胁及其防御措施小结

类型	攻击方法	描述	防御措施
投毒	数据投毒[15－18] 模型投毒[19－20]	投毒攻击主要是指在训练或再训练过程中,恶意的参与者通过攻击训练数据集来操纵机器学习模型的预测	源信息检测[28] 鲁棒低秩矩阵逼近和鲁棒主成分回归[29] 参数检测[20]
对抗	对抗攻击[21－22] 对抗样本生成方法[23－25]	对抗攻击是指恶意构造输入样本,导致模型以高置信度输出错误结果	对抗训练[30] 数据增强[31] 数据处理[32] 数据压缩[33] 防御蒸馏[34] 梯度正则化[35] 对抗样本检测[36] 基于GAN的防御[37]
隐私	模型提取攻击[26] 模型逆向攻击[27]	模型提取攻击指攻击者试图窃取模型的参数和超参数,破坏模型的保密性; 模型逆向攻击指攻击者试图从训练完成的模型中获取训练数据集的信息,从而获取用户的隐私信息	差分隐私[38－40] 秘密共享机制[41] 同态加密[42－43] 混合防御机制[44]

4　总结与展望

随着人工智能技术的发展与普及,人们感受技术带来的便利的同时,也逐渐提高了对隐私保护的需求,尤其近期欧盟颁布的《通用数据保护条例》,更加凸显出联邦学习的优势,促进联邦学习的进一步发展。

但目前联邦学习中仍存在较多的安全问题,本文主要针对投毒攻击、对抗攻击及隐私泄露这三类安全问题,总结了针对性的安全与隐私保护防御措施。然而这不是一项简单的任务,现有的防御方法只能在一定的条件下,在一定的范围内提高模型的鲁棒性。在联邦学习的安全性问题中,还有一些问题仍待解决。

1)数据质量问题。由于数据集存储在本地,服务器无法接触到数据源,难以保证数据的标签是否正确,数据是否发生了混淆等问题。而且各参与方之间数据的异构程度也无从得知,若数据规模不够大,很容易因为罕见样本过多而导致对抗攻击频繁,对抗防御难度增大。可以考虑使用零知识证明和承诺协议来实现对恶意用户数据的可验证,从而保证数据质量。

2)通信效率问题。当前的联邦学习大多都是同步的,一次迭代中,服务器要与众多的参与方进行数据交互。如果要采用多种防御手段保证模型与敏感信息的安全,势必会加重服务器的通信负担,甚至会造成拒绝服务攻击或单点失败。若考虑多个服务器,则服务器之间的交互安全也是一个值得深入探索的课题。因此,如何实现高效的隐私保护,在不得不使用公钥密码来保护用户隐私的条件下,减少其使用的次数[45－46]。

3)模型可解释性问题。联邦学习方式进一步加大了模型的复杂度,缺乏可解释性可能会导致联邦学习应用过程中的潜在威胁。可解释性是指向人类解释或以呈现可理解的术语的能力[47],提高联邦学习模型的可解释性和透明性有利于消除内在的安全隐患,进一步提高模型的可靠性和安全性。由于联邦学习的内在性质,未来可能需要着重研究事后可解释性方法。

联邦学习是一个非常有前景的研究领域,已经吸引了众多学者进行相关领域的研究,也取得了一系列重要研究成果。但联邦学习技术的发展还处于初级阶段,仍然存在许多问题尚待解决。在未来工作中,要继续研究联邦学习领域的安全问题,加快研究和发展相关安全与隐私保护技术,促进联邦学习的进一步发展。

参 考 文 献

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging[J]. arXiv preprint, arXiv: 1602.05629,2016.

[2] KONEČNÝ J,MCMAHAN H B,RAMAGE D,et al. Federated optimization:distributed machine learning for on-device intelligence[J]. arXiv preprint, arXiv: 1610.02527, 2016.

[3] KONEČNÝ J, MCMAHAN H B,YU F X,et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint, arXiv: 1610.05492,2016.

[4] MOHRI M,SIVEK G,SURESH A T.Agnostic federated learning[J]. arXiv preprint, arXiv: 1902.00146,2019.

[5] YUROCHKIN M, AGARWAL M, GHOSH S, et al. Bayesian nonparametric federated learning of neural networks[J]. arXiv preprint, arXiv: 1905.12022, 2019.

[6] NIKNAM S, DHILLON H S, REED J H. Federated learning for wireless communications:Motivation, opportunities and challenges[J]. arXiv preprint, arXiv: 1908.06847, 2019.

[7] SHELLER M J, REINA G A, EDWARDS B,et al. Multi-institutional deep learning modeling without sharing patient data:A feasibility study on brain tumor segmentation[C]// International MICCAI Brainlesion Workshop. Springer, Cham, 2018:92 − 104.

[8] CHEN Y, WANG J, YU C, et al. FedHealth: A federated transfer learning framework for wearable healthcare[J]. arXiv preprint, arXiv: 1907.09173, 2019.

[9] 胡彬轩. 基于联邦学习的空气质量监测系统设计与实现 [D]. 北京: 北京邮电大学, 2019.

[10] CUSTERS B, SEARS A,DECHESNE F, et al.EU Personal Data Protection in Policy and Practice[M].Springer, 2019.

[11] YANG Q, LIU Y, CHEN T, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1 − 19.

[12] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345 − 1359.

[13] CUSTERS B, SEARS A, DECHESNE F, et al. EU personal data protection in policy and practice[M]. TMC Asser Press, 2019.

[14] 何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049 − 2070.

[15] JIANG W, LI H, LIU S, et al. A flexible poisoning attack against machine learning[C]//ICC 2019-2019 IEEE International Conference on Communications (ICC). Shanghai: China, IEEE, 2019: 1 − 6.10.1109/ICC. 2019. 8761422.

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

[17] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[J]. arXiv preprint, arXiv: 1206.6389, 2012.

[18] MUÑOZ-GONZÁLEZ L, BIGGIO B, DEMONTIS A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017: 27 − 38.

[19] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey[J]. arXiv preprint, arXiv: 1909.11875, 2019.

[20] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[J]. arXiv preprint, arXiv: 1811.12470, 2018.

[21] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2013.

[22] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014.

[23] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv preprint, arXiv: 1607.02533, 2016.

[24] PAPERNOT N, MCDANIEL P, JHA S, ET AL. The limitations of deep learning in adversarial settings [C]// 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Saarbrucken, Germany: IEEE, 2016: 372 − 387.

[25] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE,2016: 2574-2582.

[26] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[C]//25th {USENIX} Security Symposium ({USENIX} Security 16). New York, NY, USA:[s.n.],2016: 601 − 618.

[27] ATENIESE G, FELICI G, MANCINI L V, et al.

Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers[J]. arXiv preprint, arXiv: 1306.4447, 2013.

[28] BARACALDO N, CHEN B, LUDWIG H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, Texas, USA: ACM, 2017: 103-110.

[29] LIU C, LI B,VOROBEYCHIK Y,et al.Robust linear regression against training data poisoning [C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.Dallas, Texas, USA:ACM, 2017: 91 − 102.

[30] MIYATO T, MAEDA S, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1979 − 1993.

[31] LIANG B, LI H, SU M, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2018: 1 − 14.

[32] ZANTEDESCHI V, NICOLAE M I, RAWAT A. Efficient defenses against adversarial attacks[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.Dallas,Texas, USA:ACM, 2017: 39 − 49.

[33] 吴嫚,刘笑嶂. 基于 PCA 的对抗样本攻击防御研究[J]. 海南大学学报:自然科学版, 2019(2): 134 − 139.

[34] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE, 2016: 582 − 597.

[35] ROSS A S, DOSHI-VELEZ F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]//Thirty-second AAAI conference on artificial intelligence.Lousiana, USA:[s.n.], 2018.

[36] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv preprint, arXiv: 1801.02613, 2018.

[37] SHEN S, JIN G, GAO K, et al. Ape-gan: Adversarial perturbation elimination with gan[J]. arXiv preprint, arXiv: 1707.05474, 2017.

[38] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: A client level perspective[J]. arXiv preprint, arXiv: 1712.07557, 2017.

[39] 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.

[40] BHOWMICK A, DUCHI J, FREUDIGER J, et al. Protection against reconstruction and its applications in private federated learning[J]. arXiv preprint, arXiv: 1812.00984, 2018.

[41] 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 Texas, USA: ACM, 2017: 1175 − 1191.

[42] 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.

[43] HESAMIFARD E, TAKABI H, GHASEMI M. Cryptodl: Deep neural networks over encrypted data[J]. arXiv preprint, arXiv: 1711.05189, 2017.

[44] HAO M, LI H, XU G, et al. Towards efficient and privacy-preserving federated deep learning[C]//ICC 2019-2019 IEEE International Conference on Communications (ICC).Shanghai, China: IEEE, 2019: 1 − 6.

[45] 周俊,董晓蕾,曹珍富. 推荐系统的隐私保护研究进展[J]. 计算机研究与发展, 2019, 56(10): 2033 − 2048.

[46] 曹珍富,董晓蕾,周俊,等. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展, 2016, 53(10): 2137 − 2151.

[47] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071 − 2096.

作者介绍

周俊(1982—)，博士，副教授，主要研究方向为外包系统安全与隐私保护、安全多方计算、AI 安全与区块链隐私保护等。近年来以第一作者或通信作者身份在国际密码与安全领域权威期刊或会议上发表文章 20 余篇，其中 CCF A 类或中科院一区论文 14 篇，ESI 高被引论文 3 篇，ESI 热点论文 1 篇。

(编校:侯雪婷)