文章编号：2095-2163(2021)01-0126-05 中图分类号：TN915.08 文献标志码：A

联邦学习与数据安全研究综述

王壮壮，陈宏松，杨丽敏，陈丽芳
（华北理工大学 理学院，河北 唐山 063210）

摘 要： 数据孤岛是制约人工智能技术发展和落地的主要障碍，随着国家与个人对隐私保护意识的增强，联邦学习在数据不共享的情况下，却能达到数据共享目的，受到广泛关注，联邦学习分为：横向联邦学习、纵向联邦学习和联邦迁移学习，具有数据隔离、质量保证、各参数方地位等同、独立性等优点，但联邦学习也存在很多的安全隐患，本文详细探讨了联邦学习的原理，提出了中央服务器、数据传输、单方数据污染、数据泄露以及对抗攻击等重要的数据安全问题，并汇总介绍了当前主要的防御措施。

关键词： 联邦学习；数据安全；对抗攻击；数据投毒

Review of federal learning and data security

WANG Zhuangzhuang, CHEN Hongsong, YANG Limin, CHEN Lifang
(College of Sciences, North China University of Science and Technology, Tangshan Hebei 063210, China)

【Abstract】 Data island is the main obstacle that restricts the development and implementation of artificial intelligence technology. With the enhancement of the awareness of privacy protection of the state and individuals, federal learning can achieve the purpose of data sharing without data sharing, which has been widely concerned. Federal learning is divided into horizontal federal learning, vertical federal learning and federal transfer learning. It has the advantages of data isolation, quality assurance, equal status and independence of various parameters, but federal learning also has many security risks. This paper introduces the principle of federal learning in detail, some important data security problems such as central server, data transmission, unilateral data pollution, data leakage and anti-attack are put forward. Meanwhile, the current main defense measures are summarized.

【Key words】 federal learning; data security; anti-attack; data poisoning

0 引 言

随着科技、信息化的迅速发展，通过信息共享来对数据进行资源整合是目前常用的一种手段，大数据[1]和人工智能时代[2]，AI 已经在方方面面得到体现，比如人脸识别、人工智能打败人类围棋手、无人驾驶等等，而经由研究可知，大规模的数据集能够提高 AI 的性能，数据对于 AI 的重要性就好比石油对于工业的重要性。然而，生活中大量的数据是私密的、不能共享的，大部分个人和企业所拥有的数据都存在质量较差、数据量有限的问题，只有为数不多的几家公司才能支撑 AI 技术的实现。个人隐私被窃取、公司数据泄露等安全问题日益凸显，想把分布在各个地方的数据集整合起来几乎是不可能的，或者需要耗费巨大成本。将各个用户、企业的数据共享，就会出现数据安全问题，当前无论国内还是国际都在进一步加强数据保护，陆续发布了相关法规，比如欧盟提出的新法案《通用数据保护条例》，中国起草的《数据安全管理办法（征求意见稿）》，颁布的《信息安全技术-个人信息安全规范》，在《网络安全法》中都提及了相关的数据保护[3-4]问题，同时人们对于自己数据隐私的重视程度也逐渐攀升。数据安全得以改善的同时，也造成了不同企业之间甚至同一企业不同部门之间的数据壁垒更加难以打破，从而形成了大量数据孤岛[5]，如果参与者之间无法实现数据交换，除了那些拥有巨体量用户具有绝对数据优势的企业之外，大多数企业或个人无法通过合法合规且低成本、高效率的方式跨越 AI 的数据“鸿沟”。联邦学习技术正是在数据不共享的情况下，达到数据共享目的，因其能有效解决数据孤岛，并已应用到很多领域，让数据在安全合法的前提下实现自由共享[6]。

基金项目： 河北省自然科学基金（F2014209086）。

作者简介： 王壮壮（1998-），男，硕士研究生，主要研究方向：联邦学习、网络空间安全；陈宏松（1998-），男，硕士研究生，主要研究方向：智能推荐、网络空间安全；杨丽敏（1997-），女，硕士研究生，主要研究方向：机器学习、网络空间安全；陈丽芳（1973-），女，博士，教授，主要研究方向：机器学习、数据挖掘、智能计算。

通讯作者： 陈丽芳 Email：hblg_clf@163.com

收稿日期： 2020-10-30

1　联邦学习概述

1.1　联邦学习概念

谷歌 2017 年 4 月份在一篇文章中第一次提出联邦学习[7-9]的概念，联邦学习本质上是一种加密的分布式机器学习技术、一种机器学习框架，目的是保证数据隐私安全并以合法的方式使用数据。联邦学习有三大构成要素：数据源、中央服务器、各个客户端。其中，中央服务器先为参与方提供最初模型，各参与方训练自己所拥有的数据，将得到的本地模型上传给中央服务器，中央服务器整合得到新的初始模型之后，再分发给参与方们，迭代该流程直到模型收敛。每个参与者在联邦学习框架中地位相同，可以自由控制加入还是退出，联邦学习实现了在数据之上建模，达到与数据集中收集存储相同的建模效果，很大程度上解决了数据孤岛的问题。

1.2　联邦学习分类

1.2.1　横向联邦学习

2 个数据集的用户特征重叠较多而用户重叠较少，根据特征维度横向切分数据集，挑选双方相同用户特征而用户不完全相同的数据进行训练。通过联邦学习获得更多的符合某一特征样本，这种方法叫做横向联邦学习。

1.2.2　纵向联邦学习

2 个数据集的用户重叠较多而用户特征重叠较少，根据用户维度将数据集进行纵向切分，挑选双方相同用户而用户特征不完全相同的数据进行训练。通过联邦学习丰富样本的特征，更精准地刻画样本，比如通过纵向联邦学习描绘用户，更全方位地通过各种属性特征描绘一个人。

1.2.3　联邦迁移学习

联邦迁移学习是纵向联邦学习的一种特例，和纵向联邦学习相同点在于数据特征维度重叠部分较少，而联邦迁移学习面临的情况更加苛刻，因为用户特征维度重叠部分也很少，利用迁移学习来克服数据或标签不足的情况。比如说，人们小时候学习骑自行车后，此后再学习骑电车和摩托车就会更加得心应手，因为平衡感、方向选择、行车感觉基本都是一样的，而这就是一种迁移学习。

1.3　联邦学习优势

（1）数据隔离：数据映射为模型参数，不会离开存储的地方，保证参与方的数据安全和隐私保护。

（2）质量保证：虽然没有将数据集中存储，但能够保证训练模型的质量，不会比集中训练的模型质量差。

（3）地位相同：参数方地位等同，没有一方控制另一方的情况。

（4）独立性：拥有自己数据的绝对控制权，决定参与还是退出，保持独立性的情况下，加密交换各种信息。

2　联邦学习中数据安全问题

2.1　中央服务器的存在

在迭代过程中每个参与方训练模型历次更新的信息都需要发给中央服务器，中央服务器有机会对更新信息进行分析得出参与方的本地数据信息。

2.2　数据传输的问题

更新信息向中央服务器报告的时候，虽然梯度是原始信息的映射，但是攻击者可以通过与模型交互，对更新信息中的敏感部分，如梯度信息、参数特征等进行逆向推理，反推出参与方本地数据信息。

2.3　单方数据污染

联邦学习中，每个参与方都是独立的个体，中央服务器并没有具备检验参与方数据正常与否的能力，这就导致如果攻击者从联邦学习内部对数据投毒或模型投毒，如 Chen 等人[10]提出的攻击方式，仅仅利用少量有毒样本，就有九成以上的攻击成功率，还可以在生成的模型里埋下隐患，将模型参数训练值引导成拟欲得到的结果，使模型预测样本精确度降低，性能下降。

2.4　数据泄露

虽然联邦学习的数据训练是在本地进行，各参与方之间相互独立，可以在一定程度上保证数据的安全，但依然存在数据泄露的情况。比如存在恶意参与方，从中央服务器共享的参数中对其他参与方的部分数据进行推理，而有些数据只需要获得部分，就能知晓整体，从而达到窃取数据的目的。参与方一般受到模型提取攻击、模型逆向攻击和成员推理攻击。通常研究中的训练认定服务器是可信的，在实际情况中，这不是肯定的，如果是恶意服务器，就可通过与各参与方的交互，拥有更多泄露参与方隐私的可能，数据泄露的威胁增大。

这里将给出隐私攻击的 3 种技术方法，详述如下。

（1）模型提取攻击（model extraction attack）：向目标模型不断发送数据，通过观察得到的回应信息推测模型的参数和作用，从而生成精确模型或相似模型来实现模型的提取，精确模型是指攻击者构建

的一个在预测性能上相近的替代模型,如果窃取到精确模型,则模型拥有方的数据信息泄露,损失程度较大,而窃取相似模型可以生成对抗样本,对目标模型有极大威胁。

(2)模型逆向攻击(model inversion attack):攻击者从模型给出的预测结果中提取目标模型的数据信息,与 GAN 结合后效果更为显著。Hitaj 等人[11]的研究表明,联邦学习结构避免参与者数据集遭受 GAN 的攻击是很困难的,基于 GAN 的攻击者可以诱导受害者泄露更多隐私数据。

(3)成员推理攻击(membership inference attack):攻击者通过访问目标模型的 API 接口获得大量数据,以此模仿模型构建"影子"模型。攻击者不需要了解模型参数、结构、训练方法及数据,只需要得到预测分类的置信度,最终构建一个攻击模型,通过拥有的数据记录和模型的黑盒访问权限,将数据输入目标模型,把得出结果连同数据集的标签输入其中,就可以判断目标模型的数据集是否存在该记录。

2.5 对抗攻击

对抗攻击是由 Christian 等人提出的,通过生成针对性的对抗样本并放到目标模型,导致模型做出误判。张思思等人[12]介绍了什么是对抗样本、对抗样本的概念、出现的原因、攻击方式以及一些关键技术问题。根据攻击环境,对抗攻击可以分为 2 种。一种是白盒攻击,攻击者知道目标模型使用的算法与参数,借助优化问题计算所需干扰,攻击者在对抗数据生成的时候能够与目标模型交互。另一种是黑盒攻击,攻击者不知道目标模型所使用的算法与参数,只能通过为模型提供输入跟模型互动的时候,观察判断其输出,细微的数据修改也能为攻击者提供一种攻击手段。

3 防御机制

联邦学习在面临诸多风险时,主要采用防御措施。对此可做阐释论述如下。

3.1 投毒攻击防御

3.1.1 数据投毒防御

数据中毒的根本原因,是没有考虑到输入模型的数据可能有误,甚至遭遇人为破坏。所以这里提出的保护数据措施要在模型训练前排查数据的来源,在不能保证数据安全性前,要确保数据具有不被修改的完整性。在模型训练前,可以使用身份认证机制保证数据源的可靠性。Nathalie 等人[13]辨别有毒数据的方式是借助相关训练集中数据点的起源和上下文信息的转换。

3.1.2 模型投毒防御

当服务器是可信的时候,数据合理采样,不能过分重视那些混在样本中的恶意样本,可以通过限制每个参与方贡献数据的数量,或者根据数量使用衰减权重实现。对每轮更新的模型参数做异常点检测,当某个参与方提交的更新参数与其他参与方的参数有很大差异,则将该参数设定为异常点,在进行参数整合时不再考虑。

3.2 对抗攻击防御

很多机器学习适用的对抗攻击防御方法对于联邦学习也同样适用,如基于 GAN 的防御。研究内容详见如下。

3.2.1 基于 GAN 的防御

对抗攻击是需攻克的一大难关。防御中涉及的一个方向就是设计一种能够抵御更多对抗攻击算法、鲁棒性强且高效的防御模型,生成对抗网络(Generative adversarial networks, GAN)由生成器和鉴定器构成。其中,生成器就是利用噪声生成样本,鉴定器判断样本的真假,采用零和博弈的思想带动两者的更新与演变。孔锐等人[14]将攻击算法融入 GAN,提出一种基于 GAN 的对抗攻击防御模型(AC-DefGAN),其训练样本是通过对抗攻击算法生成的,为了模型的稳定,在模型训练期间加入条件约束,借助分类器对生成样本进行分类以完成 GAN 的训练,并以需要防御的攻击算法生成对抗样本训练判别器,最终得到的分类器可以抵御多种对抗攻击。实验表明该方法防御效果好、鲁棒性强。

3.2.2 数据泄露防御

防御成员推理攻击可以运用诸如 droupout、regularization 等一些防止过拟合的方法。黑盒攻击防御策略是:训练其他的分类器对抗输入,防止模型过度拟合,增强分类器的健壮性,可以让模型加速收敛,但增加一个分类器会极大降低分类效率。白盒攻击防御策略是:实行对抗训练程度。对于异常的更新参数,通常有 2 种检测方法。一种方法是检查准确度,比较 2 个模型在验证数据集上的准确度,若 2 个模型准确度差异较大则判定为异常。另一种方法是直接比较各个参与方提交的更新参数,当参数之间的差异较大时判定为异常。

3.3 差分隐私

2006 年,Dwork 等人[15]首次提出差分隐私来解决层出不穷的隐私攻击方式和现代隐私保护机制的

缺陷，当参与者与中央服务器交互时，直接发送信息可能会出现信息泄露，为避免这种情况，训练过程中传递的模型更新信息加入差分隐私，能够有效阻止攻击者反向推导出参与方的相关数据信息。差分隐私采用特定的随机算法对数据添加适量噪声，将数据模糊化，降低敏感数据信息泄露的风险，不会被数据量约束，这样即使攻击者得到交互的数据也不能对原始数据进行有效推理，在参与方与中央服务器共享数据前，对数据进行差分隐私，就不需要考虑中央服务器是否值得信任。差分隐私确保隐私的方式是引入随机性，通过牺牲一定的准确性得到更高的隐私安全，从而实现数据保护。当前，差分隐私的主要研究方向是确定隐私保护力度和在保证隐私的前提下能够最大程度不破坏原数据的有用信息。

3.4　同态加密

同态加密（homomorphic encryption，HE）是一种可以直接对密文进行运算的加密方式，运算的结果与直接在明文上做运算的结果相同。HE 技术在密码学领域是公认的圣杯之一，根据发展阶段、支持密文运算的种类和次数，HE 分为全同态加密、部分同态加密、类同态加密、层次同态加密等。鲍海燕等人[16]在传统非对称密钥（RSA）的基础上，提出了提升加密速度的改进 RSA 算法，实验表明该方法加密过程耗时少，抵御攻击成功率高。赵秀凤等人[17]基于噪声淹没技术，提出循环安全的公钥同态加密方案，降低了系统参数的复杂度，缩减了公钥和密文大小，在实用性上突破了其研发瓶颈，提升了 HE 的性能。仝秦玮等人[18]提出一种基于 DGHV 适应智能合约的同态加密方法，可以提升保护数据、对布尔型或整型数据交易的效率。

3.5　安全多方计算

著名图灵奖获得者、华裔科学家姚期智教授[19]提出了姚氏百万富翁问题，即 2 个不暴露自身财富的富豪，怎么才能判断谁更加富有，这个问题演变成了当下的安全多方计算（Secure Multi－Party Computation，SMC）。即在不分享原始数据的情况下，获得想要的结论。多方安全计算技术[20]每次随机加密，不能重复使用加密的数据，直接在加密数据上运算，原始数据不被还原，每次计算之前先确定参与方，需要所有参与方共同协调，能够不泄露原始数据而得到数据中的价值。使用多方安全计算进行模型梯度更新的整合，可以降低信息泄露的可能性。

4　结束语

联邦学习可以有效解决跨设备、跨机构间的数据融合问题，从目前数据产业可以看出，联邦学习能够扩充数据总量，解决数据孤岛问题。站在企业的角度[21-22]，联邦学习能够帮助各企业以低成本的方式合法取得更利于自身的数据信息，相互之间可以加强合作，利用整个社会资源，做到数据共享的同时避免信息泄露。目前联邦学习最流行的算法之一是联邦平均算法（Federated averaging），在联邦学习中，各参与方对自己的设备和数据拥有绝对的控制权，受多因素影响，可能随时加入或退出联邦学习。本文对联邦学习中的几种典型的数据安全问题和防御措施进行了梳理总结，对联邦学习的工作原理和优缺点进行阐述，随着联邦学习的发展与应用，对数据隐私攻击的方式会更加多样化，后续则需要进一步探索针对投毒攻击、对抗攻击、数据泄露等攻击手段的防御技术。

参考文献

[1] 冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014，37(1)：246-258.

[2] 张夏明，张艳. 人工智能应用中数据隐私保护策略研究[J]. 人工智能，2020(4)：76-84.

[3] 杜雁芸. 大数据时代国家数据主权问题研究[J]. 国际观察，2016(3)：1-14.

[4] 赵朋. 大数据背景下的数据安全[J]. 计算机与网络，2020，46(14)：51.

[5] 杜小勇. 消除信息孤岛，实现“数据福利”[J]. 国家治理，2020(30)：20-23.

[6] 王利明. 数据共享与个人信息保护[J]. 现代法学，2019，41(1)：45-57.

[7] 王佳，苗璐. 联邦学习浅析[J]. 现代计算机，2020(25)：27-31，36.

[8] NELSON P. Federated learning improves how AI data is managed，thwarts data leakage[J]. Network World (Online)，2020.

[9] 杨强. AI 与数据隐私保护：联邦学习的破解之道[J]. 信息安全研究，2019，5(11)：961-965.

[10] CHEN Xinyun，LIU Chang，LI Bo，et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv：1712.05526，2017.

[11] HITAJ B，ATENIESE G，PEREZ-CRUZ F. Deep models under the GAN：Information leakage from collaborative deep learning[C]//Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. New York：ACM，2017：603-618.

[12] 张思思，左信，刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报，2019，42(8)：1886-1904.

[13] NATHALIE B，CHEN B，LUDWIG H，et al. Mitigating poisoning attacks on achine learning models：A data provenance based approach[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York，USA：ACM，2017：103-110.

（下转第 133 页）