计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200426
57(10): 2066-2085, 2020

机器学习的安全问题及隐私保护

魏立斐 陈聪聪 张 蕾 李梦思 陈玉娇 王 勤
(上海海洋大学信息学院 上海 201306)
(Lfwei@shou.edu.cn)

Security Issues and Privacy Preserving in Machine Learning

Wei Lifei, Chen Congcong, Zhang Lei, Li Mengsi, Chen Yujiao, and Wang Qin
(*College of Information Technology, Shanghai Ocean University, Shanghai* 201306)

Abstract In recent years, machine learning has developed rapidly, and it is widely used in the aspects of work and life, which brings not only convenience but also great security risks. The security and privacy issues have become a stumbling block in the development of machine learning. The training and inference of the machine learning model are based on a large amount of data, which always contains some sensitive information. With the frequent occurrence of data privacy leakage events and the aggravation of the leakage scale annually, how to make sure the security and privacy of data has attracted the attention of the researchers from academy and industry. In this paper we introduce some fundamental concepts such as the adversary model in the privacy preserving of machine learning and summarize the common security threats and privacy threats in the training and inference phase of machine learning, such as privacy leakage of training data, poisoning attack, adversarial attack, privacy attack, etc. Subsequently, we introduce the common security protecting and privacy preserving methods, especially focusing on homomorphic encryption, secure multi-party computation, differential privacy, etc. and compare the typical schemes and applicable scenarios of the three technologies. At the end, the future development trend and research direction of machine learning privacy preserving are prospected.

Key words machine learning; privacy preserving; security threat; secure multi-party computation; homomorphic encryption; differential privacy

摘 要 近年来,机器学习迅速地发展,给人们带来便利的同时,也带来极大的安全隐患.机器学习的安全与隐私问题已经成为其发展的绊脚石.机器学习模型的训练和预测均是基于大量的数据,而数据中可能包含敏感或隐私信息,随着数据安全与隐私泄露事件频发、泄露规模连年加剧,如何保证数据的安全与隐私引发科学界和工业界的广泛关注.首先,介绍了机器学习隐私保护中的敌手模型的概念;其次总结机器学习在训练和预测阶段常见的安全及隐私威胁,如训练数据的隐私泄露、投毒攻击、对抗攻击、隐私攻击等.随后介绍了常见的安全防御方法和隐私保护方法,重点介绍了同态加密技术、安全多方计算

收稿日期:2020-06-10;修回日期:2020-07-28
基金项目:国家自然科学基金项目(61972241,61802248,61672339);上海市自然科学基金项目(18ZR1417300);上海海洋大学骆肇荛科技创新基金项目(A1-2004-20-201312)
This work was supported by the National Natural Science Foundation of China (61972241, 61802248, 61672339), the Natural Science Foundation of Shanghai (18ZR1417300), and the Luo Zhaorao Science and Technology Innovation Fund of Shanghai Ocean University (A1-2004-20-201312).
通信作者:张蕾(Lzhang@shou.edu.cn)

技术、差分隐私技术等,并比较了典型的方案及3种技术的适用场景.最后,展望机器学习隐私保护的未来发展趋势和研究方向.

关键词 机器学习;隐私保护;安全威胁;安全多方计算;同态加密;差分隐私

中图法分类号 TP309

依托于云计算、物联网、大数据技术的发展,以数据挖掘和深度学习为代表的人工智能技术正在改变人类社会生活,并成为先进科技应用的代表和社会关注的热点.作为引领未来的战略性技术,人工智能技术被世界各国纷纷提升为发展国家竞争力、维护国家安全的重大战略.

机器学习是一种实现人工智能的方式,是近些年主要研究的领域.目前机器学习方案在很多领域都有着成熟的应用,如天气预报、能源勘探、环境监测等,通过收集相关数据进行分析学习,可以提高这些工作的准确性;还有如在垃圾邮件检测、个性化广告推荐、信用卡欺诈检测、自动驾驶、人脸识别、自然语言处理、语音识别、搜索引擎的优化等各个领域,机器学习都扮演着重要的角色.然而,蓬勃发展的机器学习技术使数据安全与隐私面临更加严峻的挑战,因为机器学习的更精准模型需要大量的训练数据为支撑.

自2013年斯诺登的"棱镜"事件以来,全球信息泄露规模连年加剧,引起社会的广泛关注.2016年9月Yahoo被曝出曾被黑客盗取了至少5亿个用户账号信息;2017年微软Skype软件服务遭受DDOS攻击,导致用户无法通过平台进行通信;2018年3月美国《纽约时报》和英国《卫报》均报道:剑桥分析(Cambridge Analytica)数据分析公司在未经用户许可的情况下,盗用了高达5千万个Facebook的用户个人资料[1].2019年美国网络安全公司UpGuard发现上亿条保存在亚马逊AWS云计算服务器上的Facebook用户信息记录,可被任何人轻易地获取;IBM在未经当事人许可的情况下,从网络图库Flickr上获得了接近100万张照片,借此训练人脸识别程序,并与外部研究人员分享[2].2020年4月《华盛顿邮报》报道视频会议软件Zoom存在的重大安全漏洞:数以万计的私人Zoom视频被上传至公开网页,任何人都可在线围观,很多视频都包含个人可识别信息,甚至是在家里进行的私密谈话[3].信息泄露的途径主要分为内部人员或第三方合作伙伴泄露、信息系统无法杜绝的漏洞、机构本身的防护机制不健全、对数据的重要程度不敏感,以及对安全配置的疏忽大意等.可见,数据隐私的泄露已不单单是满足某些外部人员好奇心所驱使,而是已成为一种重要的商业获利而被广泛关注,其中不乏内外勾结、合谋获取用户的隐私等行为.

由此可见,机器学习中的安全与隐私问题已经非常严重.目前,研究人员提出了许多解决机器学习中的隐私问题的方法.本文将全面介绍机器学习中的安全问题和隐私威胁,并重点介绍隐私保护方法.

1 机器学习及敌手模型

1.1 机器学习概念

机器学习(machine learning, ML)是人工智能的一个分支,主要是研究如何从经验学习中提升算法的性能[4],它是一种数据驱动预测的模型.它可以自动地利用样本数据(即训练数据)通过"学习"得到一个数学模型,并利用这个数学模型对未知的数据进行预测.目前机器学习被广泛应用于数据挖掘[5-6]、计算机视觉[7-8]、电子邮件过滤[9-10]、检测信用卡欺诈[11-13]和医学诊断[14-15]等领域.

机器学习主要可以分为监督学习、无监督学习、半监督学习和强化学习[16],总结如表1所示.监督学习就是给定一个包含1个或多个输入和输出(标签)的数据集,通过监督学习算法得到一个数学模型,该数学模型可以用来对给定的数据进行预测.常见的监督学习算法包括支持向量机、神经网络、回归分析和分类等,常用于对电子邮件进行过滤等应用.无监督学习是给定无人为标记标签的数据集,通过无监督学习算法可以识别出数据的共性,再根据数据的共性对每个数据是否存在此类共性做出反应.常见的无监督学习算法包括聚类,通常无监督学习用于聚类分析、关联规则和维度缩减等.监督学习和无监督学习的区别就是训练数据集的目标是否标记(含有标签)过.半监督学习介于监督学习和无监督学习之间,当未标记的数据与少量标记的数据结合使用时,可以大大提高模型的准确性.而强化学习强调如何基于环境而行动(与环境进行交互或者说是利用奖惩函数进行训练),以取得最大化的预期利益.常见强化学习算法如Q-Learning[17]等,该算法常应用于自动驾驶、游戏等领域.

Table 1 Machine Learning Classification

表 1 机器学习分类

Type	Dataset Features	Typical Model/Method	Application
Supervised Learning	One or more inputs and outputs (tags)	Linear Regression Logistic Regression *K*NN Classification Naive Bayes Decision Tree Random Forest Support Vector Machine Neural Networks	Trend prediction, text classification, spam filtering, sentiment analysis, etc.
Unsupervised Learning	No artificially tagged labels	Principal Component Analysis Factor Analysis *K*-means Clustering Spectral Clustering	Data reduction, market segmentation, consumer behavior division, design sampling plan, etc.
Semi-supervised Learning	Combination of the unlabeled data with a small amount of labeled data	Generative Method Graph-based Semi-supervised Learning Semi-supervised Support Vector Machine	Application of the corresponding model of supervised learning and unsupervised learning
Reinforcement Learning	No need of the label of input/output pairs	Q-Learning SARSA DQN Strategy Gradient	Games, robotics, autonomous driving, etc.

1.2 机器学习中的敌手模型

在机器学习安全中,常常利用敌手模型来刻画一个敌手的强弱.Barreno 等人[18]在 2010 年考虑了攻击者能力、攻击者目标的敌手模型.Biggio 等人[19] 2013 年在 Barreno 等人的基础上完善,提出了包含敌手目标、敌手知识、敌手能力和敌手策略的敌手模型,这也是目前普遍接受的敌手模型或攻击者模型.从这 4 个维度刻画敌手,能够比较系统地描述出敌手的威胁程度[19-21].

1.2.1 敌手目标

敌手期望达到的破坏程度和专一性称为敌手目标.破坏程度包括完整性、可用性和隐私性,专一性包括针对性和非针对性[19].破坏完整性目标就是未经过数据拥有者的同意对数据进行增删、修改或破坏,如对个人的医疗数据进行篡改,最后将训练得到的模型进行预测得到错误的疾病类型.破坏可用性目标就是使目标服务不可用,如在训练数据集中注入大量不良数据使训练出来的模型无用[22],从而达到服务不可用的目的.破坏隐私性目标可以理解为窃取隐私数据,如将训练数据集的信息窃取等.而专一性中的针对性目标和非针对性目标则是可以产生针对性的目标破坏或者非针对性的破坏,如对医疗数据可以产生针对性的窃取某个客户的隐私信息或者对数据的完整性产生非针对性的破坏[23].

1.2.2 敌手知识

敌手知识是指敌手对目标模型或目标环境拥有的信息多少,包括模型的训练数据、模型结构及参数和通过模型得出的信息等.根据敌手拥有的信息量,可以将敌手拥有的知识称为有限知识和完全知识[21].而在机器学习的攻击中,可以根据敌手掌握的知识量将攻击方式划分为白盒攻击和黑盒攻击.白盒攻击是敌手掌握模型的一部分数据集或者完全数据集,了解模型结构、参数以及一些其他信息.而黑盒攻击则是敌手不了解模型的相关信息,但是敌手可以访问目标模型,因此敌手可以通过精心设计的输入来根据模型输出推断模型的信息[24].

1.2.3 敌手能力

敌手能力是指敌手对训练数据和测试数据的控制能力,可以将敌手对数据的影响定义为诱发性的(对数据集有影响)或者探索性的(对数据集无影响).或者将敌手能力定义为敌手是否可以干预模型训练、访问训练数据集、收集中间结果等.根据敌手对数据、模型的控制能力可进一步将敌手分为强敌手和弱敌手[20].强敌手是指敌手可以一定程度地干预模型训练、访问训练数据集和收集中间结果等;而弱敌手则只能通过攻击手段获取模型信息或者训练数据信息[25].

1.2.4 敌手策略

敌手策略是指敌手根据自身的目的、知识和能力为了对目标达到最优的结果所采取的攻击方式.敌手策略通常在机器学习的不同阶段采用不同的攻击

方式,如在训练阶段常采用的攻击方式为投毒攻击,在预测阶段常采取的攻击方式为对抗攻击、隐私攻击等.

2 机器学习中常见的安全及隐私威胁

机器学习目前广泛应用于数据挖掘[5-6]、计算机视觉[7-8]、电子邮件过滤[9-10]、检测信用卡欺诈[11-13]和医学诊断[14-15]、生物特征识别[26-28]、金融市场分析[29]等领域.机器学习在各个领域广泛使用的同时也带来了很多安全及隐私威胁,如在2019年Kyushu大学的Su等人[30]发现仅更改1个像素就可以欺骗深度学习的算法;研究人员还发现了可以干扰道路交通标志牌的方法,使自动驾驶的汽车进行错误标志牌划分[31-32];在2019年Heaven[33]发现对神经网络(neural network, NN)的对抗攻击可以使攻击者将自己的算法注入到目标AI系统中.

为了破坏机器学习模型,攻击者可以破坏其机密性(confidentiality)、完整性(integrity)和可用性(availability),这些性质构成了CIA安全模型.在该安全模型中[22],针对机密性的攻击目标是从ML系统中获取敏感数据,如攻击者想要知道某个特定的数据是否属于某个特定的训练数据集,如攻击者可以根据出院的信息(包含患者在各个医院呆过的时间、治疗方案等)[34]获取敏感数据.针对完整性的攻击目标比较多,如使目标分类错误(将"恶意类"分类为"良好类")、针对性的错误分类(如将停车标志分类为限速标志等)、误分类(将某个类分到另一个类里面)、置信度降低(降低模型的置信度)等.针对可用性的攻击目标是降低ML系统的可用性,如在训练数据集中注入大量不良数据,使训练出来的模型无用.如果能够保证CIA模型的这3个性质,那么这个系统或者协议将是安全的.

机器学习主要分为训练和预测2个阶段.一般在训练阶段前面有数据收集、清洗等工作,在训练阶段和预测阶段中间有测试阶段等,但是主要的工作还是在训练和预测阶段.常见的安全及隐私威胁也在训练和预测阶段出现,如图1所示:

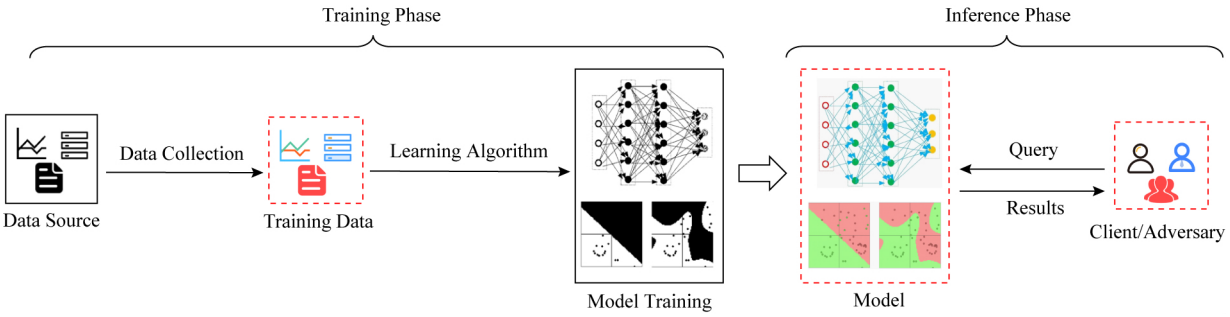


Fig. 1 Machine learning process with the security and privacy threats

图1 机器学习流程及其安全与隐私威胁

本文将从数据训练阶段和数据预测阶段分别论述机器学习中存在的安全及隐私威胁.其中,在训练阶段常见的安全及隐私威胁包含训练数据的隐私泄露和投毒攻击,在预测阶段常见的安全及隐私威胁包含对抗攻击、隐私攻击和预测数据的隐私泄露[22].

2.1 数据训练阶段存在的安全及隐私威胁

在数据训练阶段存在的隐私威胁主要为训练数据集的隐私泄露.在训练数据时,往往采用集中式学习(centralized learning)、分布式学习(distributed learning)或者是联邦学习(federated learning)的方式[20].其中,集中式学习[35]的方式就是将各方的训练数据集中到一台中央服务器进行学习;分布式学习[36]的方式就是将训练数据以及计算都分布到各个服务器节点进行学习,最后由中央服务器进行整合;联邦学习[37-39]的方式就是在保持训练数据集的分散情况下客户端与中央服务器联合训练一个模型.

联邦学习和分布式学习的方式类似,区别在于联邦学习没有数据收集阶段,分布式学习是首先进行数据收集,为了加快训练速度,将数据分发给多个服务器,然后由中央服务器协调进行迭代,从而训练出最终模型;而联邦学习是在各客户端本地进行学习得到子模型[40],然后交由中央服务器聚合得到最终模型.3种形式的学习如图2所示.

根据这3种学习的方式,不论从数据收集还是训练方式的角度出发,在模型训练阶段都不可避免会造成数据隐私的泄露.

在数据训练阶段存在的安全威胁主要为投毒攻击(poisoning attack)[41],投毒攻击的例子可以追溯

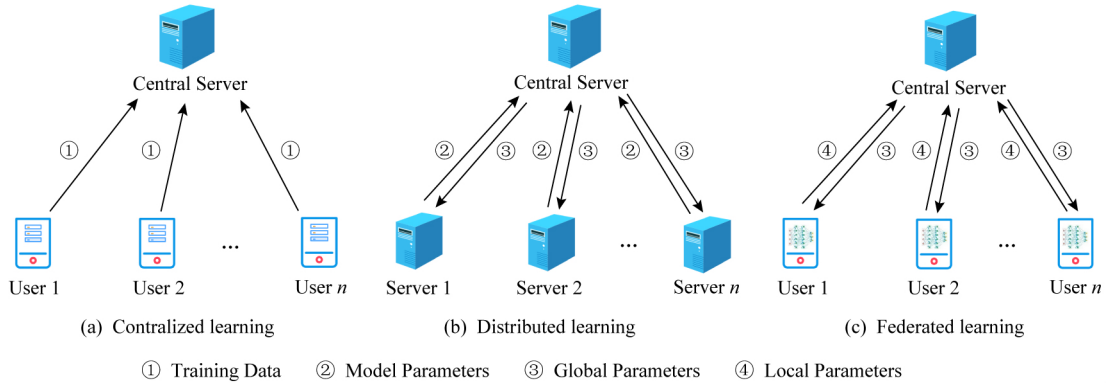


Fig. 2 The schematic diagram of three learning methods[20]

图 2 3 种学习方式示意图[20]

到 2004 年[42]和 2005 年[43]逃避垃圾邮件分类器的例子.投毒攻击就是敌手可以通过修改、删除或注入不良数据,改变原有数据集,使模型训练的结果产生偏差.最常见的投毒攻击就是使模型边界发生偏移,如图 3 所示:

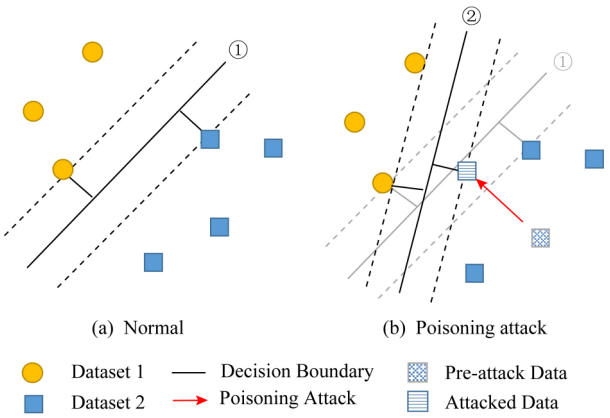


Fig. 3 The linear SVM classifier decision boundary[44]

图 3 线性 SVM 分类器决策边界[44]

2.1.1 训练数据的隐私泄露

训练数据的隐私泄露,就是在模型训练时可能发生数据泄露问题.目前大多数公司或者模型提供商都是使用集中式学习的方式训练模型,因此需要大规模的收集用户数据.但是对于收集用户数据时保护用户隐私没有一个统一的标准[21],所以在收集用户数据时可能会造成用户的数据隐私泄露的问题.在 2018 年图灵奖得主 Goldwasser[45]在密码学顶级会议 CRYPTO 2018 上指出了安全机器学习中密码学的 2 个主要发展方向:分布式模型训练和分布式预测,通过安全多方计算(secure multi-party computation, SMC)实现隐私保护机器学习.

在大数据驱动的云计算网络服务模式下,Jiang 等人在国际顶级安全会议 ACM CCS 2018 上提出了基于机器学习的密态数据计算模型[46],具体是由数据拥有者、模型提供商和云服务提供商组成.其中,数据拥有者对数据的拥有权和管理权是分离的;而云服务提供商也通常被假定为诚实且好奇(半可信)的,即在诚实运行设定协议的基础上,会最大程度发掘数据中的隐私信息,且该过程对数据拥有者是透明的.

使用机器学习对海量数据进行信息挖掘和学习,在安全模型增加模型提供商和云服务提供商后也变得更加复杂,这就需要考虑多方参与情形,构建隐私增强保护的数据计算模型和协议.

具体来说,在密态数据计算模型已经衍生出众多多方场景,例如:1)多方云服务提供商.数据拥有者通过秘密共享的方式将隐私数据信息分别分散到各个服务器(如亚马逊、微软、谷歌等多个服务提供商)上进行计算,各个服务器分别返回相应的计算结果,多个云服务提供商之间不进行主动合谋,最终由数据拥有者进行汇聚并得到结果.Shokri 等人[47]提出了与不同数据持有者合作的机器学习协议、分布式选择性随机梯度下降算法,以便在训练数据不共享的前提下展开联合机器学习模型的训练.在学习模型和学习目标协调的情况下,参与者可以训练自己的局部模型,并有选择地在每个局部随机梯度下降阶段异步交换其梯度和参数.2)多方数据拥有者.当前的企业组织多采用协作学习模型[48]或联邦学习模型[49].比如,分发相关疾病疫苗时,医疗组织希望基于大数据利用机器学习确定高爆发的地区,这就需要不同区域医疗组织的数据,但往往处于法律和隐私考量,数据无法完成及时共享.

2.1.2 投毒攻击

投毒攻击有 2 个目标:1)破坏完整性;2)破坏

可用性.破坏完整性目标就是敌手通过选择一个精心构造的恶意数据注入你的训练数据集中,通过这个精心构造的恶意数据,敌手可以得到一个“后门”,因此破坏完整性目标的投毒攻击也称为“后门”攻击.破坏可用性目标就是敌手通过注入很多不良数据到你的训练数据集,使得到的模型边界基本上变得无用[50].下面就2种不同的投毒攻击进行介绍.

2.1.2.1 破坏完整性目标的投毒攻击

对于破坏完整性目标的投毒攻击,也称“后门”攻击.对于此类攻击,模型拥有者可能无法发现模型已经被植入“后门”,但是敌手可以利用这个“后门”达到他的目的.例如一个敌手使分类器将含有某些特殊字符的文本划分为正常文本,那么敌手可以在自己编写的恶意软件中插入某些特殊字符完成任何攻击.这种攻击带来的后果非常严重,它甚至可以使ML系统完全被敌手控制[50].

Liu等人[51]提出了一种针对神经网络的攻击方法.该方法具有很强的隐蔽性,他们首先利用逆神经网络生成一个通用的触发器,然后用反向工程的训练数据重新训练模型,从而向模型注入恶意行为.恶意行为仅由带有触发器的输入激活.使用该攻击模型,可以造成严重的后果,如在自动驾驶领域可造成交通事故等[50].

2.1.2.2 破坏可用性目标的投毒攻击

破坏可用性目标的投毒攻击主要是使模型不可用.该类攻击主要是通过破坏原来训练数据的概率分布,使得训练出的模型决策边界偏离或者使得模型精度降低.在现实生活中,训练数据一般都是保密的,并且不会轻易被攻击者修改、删除等.但是因为数据的不断更新,模型也需要不断更新,这就使得攻击者可以注入大量的恶意数据,达到使模型不可用的目的.如生物特征识别、搜索引擎、电子邮件过滤、金融市场分析等领域的模型,需要进行定期更新,因此这类系统面临的破坏可用性的投毒攻击风险也更大[23].

Steinhardt等人[52]的研究表示,即使拥有强大的安全防御方案,在训练数据集中注入3%的中毒数据也可以使得模型的训练误差从12%提高到23%.目前已经存在针对情绪分析[53]、恶意软件聚类[54]、恶意软件检测[55]、蠕虫签名检测[56]、入侵检测[57-59]等投毒攻击的研究.

2.2 数据预测阶段存在的安全及隐私威胁

在模型训练完成后,通常会将训练好的模型用于预测特定的结果,以便人们做出高效的决策.因此,在预测阶段被敌手恶意攻击产生的后果往往会更加严重.预测阶段存在的安全及隐私威胁主要可以分为对抗攻击、隐私攻击和预测数据的泄露.其中,机器学习中的安全威胁为对抗攻击(adversarial attack),而隐私威胁则是隐私攻击(privacy attack).

2.2.1 对抗攻击

对抗攻击也称逃逸攻击(evasion attack)[19,60-61],是指敌手在模型原始输入上添加对抗扰动构建对抗样本(adversarial examples)从而使模型对预测结果或者分类结果产生偏差.例如,垃圾邮件发送者经常通过混淆垃圾邮件和恶意软件代码的内容来逃避检测,使得他们的垃圾邮件或者恶意软件代码是合法的.对抗攻击的过程中,选择和产生对抗扰动是非常关键的,对抗扰动一般是微小的并且有能力使模型产生错误输出.

2013年Szegedy等人[62]首次提出了对抗样本的概念;之后Goodfellow等人[63]最早提出了对抗攻击的防御方法,他们提出了一种快速梯度符号(fast gradient sign method, FGSM)的方法来生成对抗样本,FGSM用于扰动模型的输入;Moosavi-Dezfooli等人[64]基于迭代且线性近似的方案提出了一种计算对抗样本的方法DeepFool,他们利用DeepFool来更精确地生成对抗样本进行对抗训练可以有效提高分类器的鲁棒性;在2016年Papernot等人[65]提出了JSMA算法,JSMA构建一种针对性更强的对抗性样例;2017年Carlini等人[66]提出了一种更加高效的方法产生对抗样本;2020年Ru等人[67]将贝叶斯优化与贝叶斯模型选择结合优化对抗扰动和搜索空间的最佳降维程度,提高模型的鲁棒性;同年Zhou等人[68]提出了一种不需真实数据的替身模型训练的方法,他们表明该方法在黑盒攻击时具有较好的效果.

2.2.2 隐私攻击

隐私攻击是专注于隐私的一类攻击,包括模型的隐私和数据的隐私.其中模型隐私包括模型参数信息、模型结构、模型本身等关于模型的隐私信息,数据的隐私包括训练模型所用的数据集等.因此,总结常见的隐私攻击类型,可以将针对模型隐私的攻击称为模型提取攻击(model extraction attack),将针对数据隐私的攻击称为数据提取攻击(data extraction attack)和成员推理攻击(membership inference attack).

2.2.2.1 模型提取攻击

模型提取攻击的目的就是敌手通过对已经训练好的模型进行应用程序接口(application programming

interface, API)查询来非法窃取模型参数、模型结构,构建一个替代模型甚至非法获取模型本身的一种攻击方式[69],其流程图如图 4 所示.ML 系统中的模型是非常有价值的一部分.首先,对于一个组织、机构或者公司来说,模型的训练过程可能涉及到大量的数据,这些数据可能是组织、机构或者公司通过很大的代价获得并处理的,然后经过大量的时间和金钱来训练模型[70].其次,敌手可以通过模型提取攻击来获取模型参数、模型结构等信息,敌手获得这些信息之后可以更加方便地实施投毒攻击、对抗攻击等恶意攻击.因此,一旦模型泄露,可能对这些组织、机构或者公司带来巨大的损失.

Tramer 等人[69]证明了模型提取攻击对支持向量机、决策树、神经网络在内的大多数算法都有非常好的效果,并且表明对于 1 个 N 维模型,理论上只需要通过 $N+1$ 次查询就能够提取该模型.Shi 等人[71]通过深度学习高精度地提取了朴素贝叶斯和 SVM 分类器,并可以通过获取到的信息重构一个等价的模型.Wang 等人[72]提出一种针对超参数攻击的方法,通过实验表明,该方法使用了线性回归、逻辑回归、支持向量机和神经网络等算法,成功获取模型的超参数.

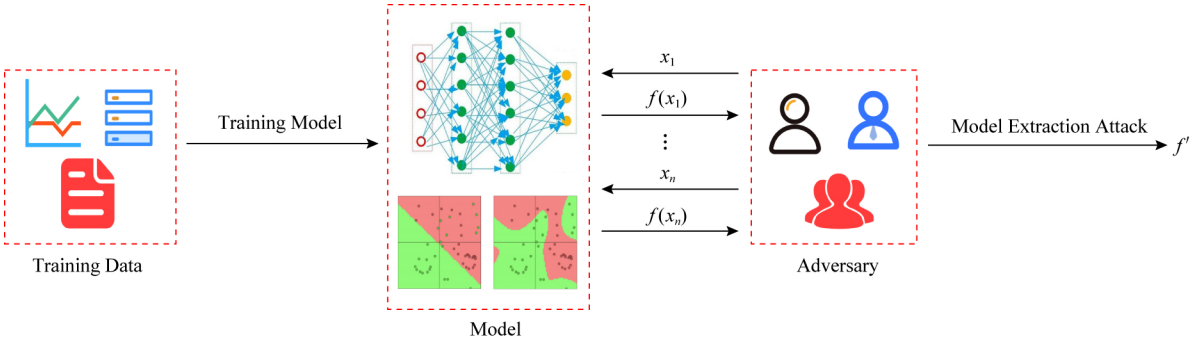


Fig. 4 The diagram of model extraction attack

图 4 模型提取攻击流程图

2.2.2.2 数据提取攻击

数据提取攻击也称模型逆向攻击(model inversion attack),由 Fredrikson 等人[73]首次提出,是指敌手通过访问模型 API,通过一系列的查询来获取模型训练数据里的隐私数据的一种攻击手段.数据提取攻击造成的隐私数据泄露可能会造成巨大的威胁,如针对训练的模型提取了病人的基因组信息[74],并且可以使药物错配,从而导致生命威胁.

Fredrikson 等人[73]通过训练好的模型,成功通过数据提取攻击重构了人脸图像;Ateniese 等人[75]构建了一种分类器使它可以攻击其他分类器并获取训练数据;Song 等人[76]证明了训练好的模型会"记忆"大量隐私信息,如果存在恶意 ML 算法的模型训练者,那么模型可能泄露训练数据集的信息;Carlini 等人[77]描述了一种可以提取隐私信息的算法,他们通过不断查询模型来获取如信用卡号码、ID 号码等隐私信息.

2.2.2.3 成员推理攻击

成员推理攻击指敌手通过访问模型 API 获取足够数据,然后构建一些"影子"模型来模仿目标模型,最后通过构建一个攻击模型来判断某些特定的数据是否在训练数据集中[34].成员推理攻击也可能造成个人敏感数据的泄露,如 Shokri 等人[34]通过成员推理攻击成功判断了特定病人是否已出院.

Pyrgelis 等人[78]对集合数据发起成员推理攻击,从而确定特定的用户是否在集合数据中;Yeom 等人[79]的研究表明,不管模型稳定的算法还是容易过度拟合的算法,都容易受到成员推理攻击;Song 等人[80]设计了一种可审计的自然语言文本深度学习模型,用于检测是否使用特定用户的文本数据来训练预言模型.Truex 等人[81]证明了 ML 模型何时以及为什么容易受到成员推理攻击;Hayes 等人[82]通过生成对抗模型(generative adversarial network, GAN)检测过拟合和识别出训练数据中的一部分,并利用鉴别者的能力了解分布的统计差异,在白盒攻击情况下,100%推断出哪些样本用于训练模型;在黑盒攻击情况下,成功率也达到了 80%.

2.2.3 预测数据的隐私泄露

预测数据的隐私问题与训练阶段的隐私问题类似,主要是数据的隐私泄露,但预测数据的泄露主要发生在人们利用模型进行预测时.对于恶意的攻击者,可能会获取到用户的隐私信息.目前有很多工作都考虑了预测数据的隐私问题,接下来的隐私保护

方法涉及到该类介绍,在此不再赘述.

2.3 小 结

本节从训练阶段和预测阶段2个角度出发,分别介绍了机器学习过程中常见的安全隐私威胁.如表2所示,总结了机器学习中常见的攻击手段所破坏的CIA安全模型的性质.在训练阶段中,训练数据的隐私泄露破坏了机器学习CIA安全模型的机密性;投毒攻击则破坏了其完整性和可用性,在一定程度下可能致使目标模型偏离甚至完全错误导致模型无用.在预测阶段中,对抗攻击则是通过添加对抗样本的方式破坏模型的完整性,导致模型可以分类错误的样本;隐私攻击和预测数据的隐私泄露破坏了CIA安全模型的机密性,其中隐私攻击可以在预测过程中泄露模型及训练数据的隐私.对于影响CIA性质中机密性的情况,可以归结为机器学习中的隐私威胁.

Table 2 The Nature of the CIA Affected by the Attack

表2 攻击手段所影响的CIA性质

Phase	Integrity	Availability	Confidentiality
Training Phase	Poisoning Attack	Poisoning Attack	Privacy Leakage of Training Data
Inference Phase	Adversarial Attack		Privacy Attack/Privacy Leakage of Inference Data

3 机器学习中常见的安全防御方法

3.1 针对投毒攻击的防御

投毒攻击常见的防御方法有异常检测、对模型进行准确性分析等[50].投毒攻击是在训练数据中注入异常数据,所以我们可以对数据进行分析,检测数据的分布情况,从而分离出异常数据[83].Baracaldo等人[84]于2017年提出了一种使用数据集中数据点的来源和转换的上下文信息来识别有毒数据的方法,从而使在线和定期再训练的机器学习应用程序能够在潜在的使用投毒攻击的敌对环境中使用数据集.

但是,攻击者可能生成与真实数据分布非常相似的异常数据,最后可以成功误导模型.Koh等人[83]提出了3种新的攻击方法,它们都可以绕过常见的投毒攻击防御手段,包括常用的基于最近邻的异常检测器、训练损失(training loss)和奇异值分解等.他们设计的攻击仅仅增加3%的有毒数据,就可以成功地将垃圾邮件检测数据集的错误率从3%增加到24%,将IMDB情感分类数据集的错误率从12%增加到29%.

因此,我们可以在训练新添加的样本数据时对模型进行准确性分析.如果收集的输入是有毒的,那么它最终的目的是破坏模型在测试集上的准确性.通过在将训练的新模型投入到生产环境之前对模型进行分析,我们可以避免投毒攻击给我们带来的影响.Suciu等人[85]提出的FAIL模型就是使用这种方法.

3.2 针对对抗攻击的防御

对抗攻击的防御方法有对抗训练[86-88]、梯度掩码[89-90]、去噪[91]、防御蒸馏[92-94]等.对抗训练(adversarial training)是通过在训练数据中引入对抗样本来提升模型的鲁棒性,是对抗攻击最有效的防御方式之一.梯度掩码(gradient masking)就是通过将模型的原始梯度隐藏起来达到抵御对抗攻击的目的.去噪(denoising)则是在输入模型进行预测之前,先对对抗样本去噪,尽可能地使对抗样本恢复成原始样本,从而提高模型鲁棒性.防御蒸馏(defensive distillation)首先根据原始样本训练一个初始的神经网络,得到一个概率分布,然后再根据这个概率分布构建一个新的概率分布,最后利用整个网络进行预测或分类,从而达到抵御对抗攻击的目的.

4 机器学习中常见的隐私保护方法

隐私保护机器学习(privacy-preserving machine learning, PPML)方法最早可追溯至2000年,Lindell等人[95]提出了允许两方在不泄露自己隐私的前提下,通过协作对联合数据集进行提取挖掘的方法,Agrawal等人[96]允许数据拥有者将数据外包给委托者进行数据挖掘任务,且该过程不会泄露数据拥有者的隐私信息.早期关于PPML的研究工作主要集中在决策树[95]、K-means聚类[97-98]、支持向量机分类[99]、线性回归[100-101]、逻辑回归[102]和岭回归[103]的传统机器学习算法层面.这些工作大多都使用Yao[104]的混淆电路(garbled circuit, GC)协议,将问题化简为线性系统的求解问题,但这不能推广到非线性模型,而且需要比较大的计算开销和通信开销,因此缺乏实施或评估案例.

在机器学习训练过程中,模型提供商会对一些

训练数据进行记录,而这些训练数据往往会涉及到用户个人的隐私等信息.在训练阶段,机器学习基于训练数据集展开模型训练,基于所学习数据的内在特征得到决策假设函数,而预测阶段目标模型的有效性则依赖于属于同一分布的训练数据集和预测数据集,但是攻击者仍可以通过修改训练数据的分布从而实施目标模型的攻击,在基于 ML 的医疗健康系统中,攻击者不仅可以窃取病人隐私还可以发起恶意数据注入或数据修改攻击,从而对病人的用药剂量产生影响.在机器学习预测过程中,攻击者无法修改训练数据集,但仍可以通过访问目标模型从而获得有效的参数信息发起攻击,使模型在预测阶段发生错误.

因此,采用隐私保护的手段保护数据和模型的安全是必不可少的,本节介绍密码学中保护机器学习中隐私的常见技术,主要包括同态加密技术、安全多方计算技术和差分隐私技术.相关工作的对比如表 3 所示:

Table 3 The Comparison of the Related Work

表 3 相关工作的对比

Reference	Scenarios			Privacy Preserving Phase		Privacy Preserving Objectives		Adversary Model			Technology			Type of Against Privacy Threats	
	Centralized Learning	Distributed Learning	Federated Learning	Training Phase	Inference Phase	Data	Model	Semi-Nonest Model	Malicious Model	Others	HE	SMC	DP	Privacy Leakage of Data	Privacy Attacks
Ref [105]		√			√	√	√	√			√	√		√	√
Ref [106]	√				√	√				√	√			√	
Ref [107]		√		√	√	√		√	√			√		√	
Ref [108]			√	√		√				√			√	√	√
Ref [109]	√			√	√	√	√			√			√	√	√
Ref [110]		√		√	√	√		√				√		√	
Ref [111]		√		√	√	√		√	√			√		√	
Ref [112]			√	√	√	√	√	√			√	√		√	√
Ref [113]	√				√	√				√	√			√	
Ref [114]	√				√	√				√	√			√	
Ref [115]		√		√	√	√		√			√	√		√	
Ref [116]	√			√	√	√				√	√			√	√
Ref [117]	√			√		√				√			√		√

4.1 同态加密技术

同态加密技术(homomorphic encryption, HE)允许直接在密文上做运算,运算之后解密的结果与明文下做运算的结果一样[118].同态加密技术常用于保护隐私的外包计算和存储中,主要是首先将数据加密,然后将加密的数据发给云进行存储或者计算,云直接在密态数据上进行操作,这样既不会泄露隐私又满足了需求.同态加密技术又可以分为全同态加密(fully homomorphic encryption, FHE)[119-125]、部分同态加密(partially homomorphic encryption, PHE)[126-127]、类同态加密(somewhat homomorphic encryption, SHE)[119-120,128-129]、层次型同态加密技术(leveled homomorphic encryption, LHE)[130-131]等.FHE 可以计算无限深度的任意电路;PHE 支持评估仅包含一种门类型的电路(如加法或者乘法);SHE 可以计算加法和乘法电路,但只支持有限次的乘法;LHE 支持对有界(预设)深度的任意电路进行计算.不同加密方案适用的场景如表 4 所示.

Gentry 基于理想格,构造了第一个理论上可行的 FHE 方案[119-120].该方案同时支持对密文上的加法和乘法运算,并可以构造执行任意计算的电路.这种构造首先设计了一个 SHE 方案,它的深度是有限的,因为每个密文在某种意义上都是有噪声的,而且这种噪声随着密文的增加而增加,直到最终噪声使得得到的密文无法辨认.Gentry 之后通过改进该方案提出自举技术,使得 SHE 可以转换为 FHE.之后也有很多人在 Gentry 的方案上进行改进,提出了更快、噪声更少的 SHE 方案和可以不使用自举技术就能评估有界(预设)深度电路的 LHE 方案[130-131].使用广泛的 RSA[126] 和 ElGamal[127] 等加密系统都是 PHE 方案.

接下来主要介绍全同态加密技术.全同态加密

技术一直被认为是进行隐私保护机器学习的一项重要技术.FHE主要分为层次型全同态加密技术(leveled fully homomorphic encryption, LFHE)和自举型同态加密技术(bootstrapping full homomorphic encryption, BFHE)[129].其中LFHE即前文提到的LHE.

Table 4 The Application Scenarios of Different Homomorphic Encryption Schemes

表4 不同同态加密方案对应的场景

Homomorphic Encryption	Application Scenarios
FHE	Evaluation of arbitrary circuits of infinite depth.
PHE	Evaluation of circuits composed of only one type of gate (i.e.addition or multiplication).
SHE	Evaluation of addition and multiplication, but the number of multiplication evaluations is limited.
LHE	Evaluation of arbitrary circuits with bounded (default) depth.

Note: For most homomorphic encryption schemes, the depth of the circuit is the main limitation for performing calculations on encrypted data.

2016年Gilad-Bachrach等人[113]提出了CryptoNets可以借助神经网络对加密数据进行相应推断,此后也有使用层次型同态加密方案对预先训练好的卷积神经网络(convolutional neural network, CNN)模型提供隐私保护性质,但是层次型同态加密技术会使得模型精度和效率严重下降.同时,模型中平方级的激活函数会被非多项式的激活函数和转换精度的权重代替,导致推导模型与训练模型得到的结果会有很大不同.此后,Hesamifard等人[114]采用低阶多项式近似逼近的方法对CryptoDL方案进行了改进,Chabanne等人[132]采用近似的激活函数和归一化操作改进加权值.Chillotti等人[133]、Bourse等人[134]则分别提出了相应的自举FHE方案,该方案均比层次型同态加密方案效率更高,值得注意的是自举FHE方案可以更加贴近实际地对单个实例进行预测.Jiang等人[46]则提出了一种基于矩阵同态加密的通用算术运算方法,提出在一些用户提供的密文数据上云服务提供商进行模型训练,该方法可以将加密模型应用到更新后的加密数据上.2019年Zheng等人[48]利用门限部分同态加密实现了Helen系统,该系统能够允许利用多个用户的数据同时训练模型,但不泄露数据.与之前的方案相比,Helen能够抵御m方中$m-1$方都为恶意的对手.但是该系统不能抵御投毒攻击、数据提取攻击和拒绝服务攻击.

尽管从理论层面认为FHE技术可以进行任意计算,但受当前相关实际方案约束,FHE普遍仅能支持整数类型的数据;同时,电路深度需要固定而不能进行无限次的加法和乘法运算;除此之外,FHE技术不支持比较运算操作.虽然,目前存在一些实数上计算并有优化的FHE方案,但数据规模大幅扩张、计算负载不断加剧、非线性激活函数的拟合计算误差等原因导致FHE的方案效率无法得到进一步提升.由此可见,FHE技术与机器学习算法的简单结合,将无法保证相关机器学习算法对密文数据进行操作.

4.2 安全多方计算技术

在基于安全多方计算技术的隐私保护机器学习[45]方面,SMC允许互不信任的各方能够在自身私有输入上共同计算一个函数,其过程中不会泄露除函数的输出以外的任何信息.但是,传统的SMC协议往往需要较为庞大的计算量和通信复杂度,导致其难以在实际机器学习中得以大规模部署.目前,常见的基于SMC的PPML解决策略有:1)基于混淆电路、不经意传输等技术的隐私保护机器学习协议,并执行两方SMC协议来完成激活函数等非线性操作计算;2)基于秘密共享技术允许多方参与方参与机器学习网络模型训练或预测,且该过程不会透露数据或模型信息.SMC主要方案的发展历程如图5所示.

4.2.1 两方隐私保护机器学习协议

典型的两方隐私保护机器学习协议包括TASTY[135], ABY[136], SecureML[115], MiniONN[105], DeepSecure[137]和GAZELLE[138]等.其中,ABY[136]利用高效乘法协议、快速的转换技术以及预处理的密码操作,构造用于两方计算环境的基于算术共享、布尔共享和Yao混淆电路的安全计算方案.值得注意的是,SecureML[115]是在两方计算环境中训练神经网络的隐私保护方法,该方案基于线性回归、逻辑回归和神经网络等模型提出一种高效的截断协议,相较于ABY提出的PPML算法更加高效.ABY和SecureML则均利用Beaver所提出三元组技术来实现部分乘法操作,但方案通信成本仍不理想.由于三元组会降低协议的效率,Chameleon[139]利用一个半诚实的第三方替代三元组从而减少了三元组的使用,另外,该方案基于ABY实现了加法秘密共享、GMW和GC的混合,大幅改进ABY的实用性和可

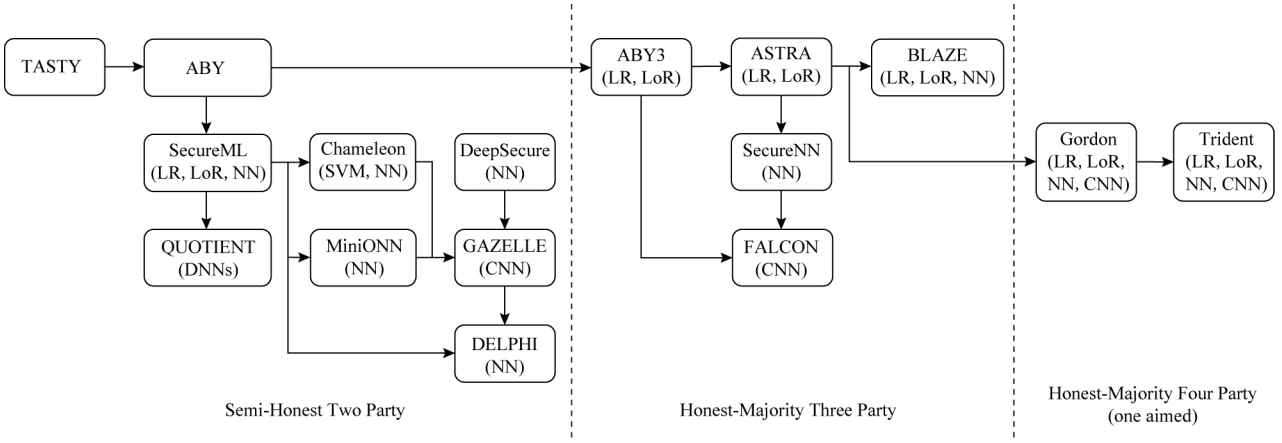


Fig. 5 Development of privacy preserving machine learning based on secure two-party/multi-party computation

图 5 基于两方/安全多方计算的隐私保护机器学习发展历程

扩展性.最近,QUOTIENT[140]和 DELPHI[141]则分别基于两方安全的神经网络完成对数据的训练和预测.这些两方的协议主要考虑在非共谋服务器情况下的安全保障需求:假设云服务器是诚实且好奇的被动攻击模型,在云服务器互不合谋情况下保证数据的安全性与机器学习方案的可用性.

为免受半可信参数服务器的攻击,MiniONN[105]基于简化的同态加密技术并将其应用于交换权重和梯度问题,把原始神经网络转换为遗忘神经网络再进行网络模型训练,并将混淆电路作为近似非线性激活函数.DeepSecure[137]基于 Yao 混淆电路[104]对深度学习模型上的加密数据进行计算和推理,并完成相应的安全证明.Juvekar 等人[138]指出 MiniONN, DeepScure 等工作表明同态加密在矩阵向量乘法具有明显的优势,但在线性运算方面并不明显.需要注意的是,虽然这些混合协议可以提高识别率,但带宽和网络延迟方面的效率并不理想.技术上来说,这些协议均是采用同态加密方法对标量乘法进行计算,采用安全多方计算对激活函数进行计算.

4.2.2 多方隐私保护机器学习协议

利用传统分布式机器学习算法[142-143],典型的多方参与的 PPML 方案有 ABY3[107], SecureNN[110], ASTRA[144], FLASH[145], Trident[146], BLAZE[147], FALCON[111],如图 6 所示.

ABY3[107]提出一种在半诚实环境和恶意环境下的新方案,可以完成三方之间的算术共享、布尔共享和 Yao 共享,它对 SecureML 中的近似定点数乘法进行改进,使其得以在三方环境下可以使用,并相应设计了一种计算分段多项式函数的协议.最终在两方和三方的情况下分别将效率提升 93 倍和 8 倍,但是 SecureNN[110]仅考虑了神经网络训练方案的设计.ASTRA[144]基于 ABY3 考虑了半诚实环境

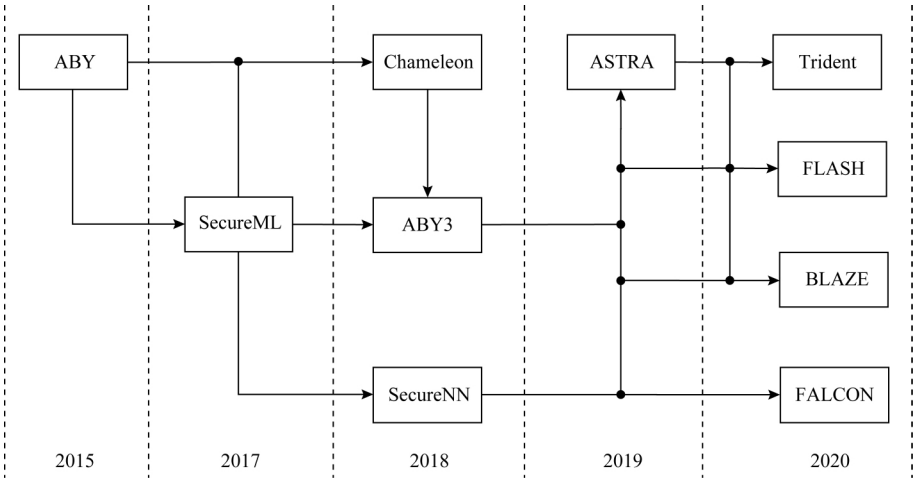


Fig. 6 Development of privacy preserving machine learning schemes involving multi-party

图 6 多方参与的隐私保护机器学习相关方案的发展历程

和恶意攻击环境,提出一种新的更加安全的三方PPML协议,它构建了一个公平的重构协议,解决了发送者在无广播频道情况下通过私人频道向其他参与方发送不同消息所可能导致的混乱,保证了当且仅当诚实方接收到输出时,恶意攻击方才会接收到输出.对于分类任务,ASTRA[144]充分运用秘密共享方案中的不对称性性质,放弃SecureML和ABY3中一些消耗较高的混淆电路或并行前缀加法器,最终提出一种新的安全比较协议.在半诚实的环境中,ASTRA则是将33%的整体通信开销转移到离线阶段,而在恶意环境基于高效的点积协议进一步提升在线通信效率.然而,ABY3[107],SecureNN[110],ASTRA[144]等方案均是专注于半诚实环境下的PPML框架,进行相关点积计算是与向量大小成线性关系,在恶意环境下向量点积、最高标记位的提取和截断的效率都会不同程度地降低.

虽然,ABY3和SecureNN提出使用Abort的安全构建组件,且ASTRA[144]将Abort的安全性提升到公平性,FLASH[145]则进一步实现了保证输出传递的最强安全概念(无论对手的行为如何,各方都获得输出)的四方PPML框架,且健壮性仅需要对称密码原语来实现.与文献[148]方案不同,FLASH不需要使用数字签名、广播等密码原语,它在协议中引入一个新的诚实参与方大大提升协议效率.Trident[146]提出了最多可以容忍一方腐败的四方PPML协议,其中第四方在除输入共享和输出重构的阶段外,所有的在线阶段均为非活动状态.在使用新的秘密共享方案后,Trident方案中25%的在线通信阶段部分转移到了离线阶段部分,在线效率得以充分提升.BLAZE[147]则是效率较高的一个PPML协议,在3个服务器的情形下可以容忍1个恶意腐败方的存在,并得到了更强的公平性保证(所有诚实方和恶意攻击方都获得相同的输出),基于所提出的点积协议、截断和位操作方法,方案效率比ABY3和ASTRA等方案更加高效[149].因此,仅有少数PPML方案考虑到了恶意攻击环境,当恶意攻击者存在情况下,还需考虑协议的公平性与一致性.

2020年微软、普林斯顿大学、以色列理工学院和Algorand基金会的研究人员基于SecureNN和ABY3推出了一个名为FALCON[111]的框架.该框架支持批归一化(batch normalization, BN)并且保护隐私的安全框架,它可以支持训练像AlexNet和VGG16这样的大容量网络.FALCON只使用算术秘密共享而避免使用转换协议(在算术、布尔和乱码电路之间)来实现针对非线性运算的恶意安全协议.

这些方案的详细比较如表5所示:

Table 5 Comparison of Privacy Preserving Machine Learning Schemes Based on Multi-party Computation

表5 多方参与的隐私保护机器学习相关方案的具体比较

Schemes	Number of Participants	Model	Features
ABY[136]	2	Semi-honest	Propose an efficient conversion ABY framework.
SecureML[115]	2	Semi-honest	Propose a new approximate fixed-point multiplication protocol.
Chameleon[139]	3	Semi-honest	Utilize a linear secret sharing model.
ABY3[107]	3	Semi-honest, Malicious	Expand the fixed-point number multiplication protocol for tripartite computation; Propose a protocol for calculating the piecewise function.
SecureNN[110]	3 or 4	Honest(majority)	Implement a PPML scheme based on neural network algorithm.
ASTRA[144]	3	Semi-honest, Malicious	Propose an efficient dot product protocol.
FLASH[145]	4	Malicious (at most one)	Propose a PPML framework with high robustness (symmetric key primitive); Propose a dot product protocol that has nothing to do with vector size, a new efficient truncation protocol; No need for digital signatures and expensive primitives (such as broadcasting and public key settings).
Trident[146]	4	Malicious (at most one)	No need for expensive multiplication triples; Dot product operation can be performed at a cost independent of the size of the two vectors; Propose an efficient truncation protocol.
BLAZE[147]	3	Malicious (at most one)	Propose a dot product protocol independent ofthe vector size, an efficient new truncation protocol and an efficient bit operation protocol.
FALCON[111]	3	Semi-honest, Malicious	First secure framework that supports operations on the batch normalization layer in privacy machine learning; Support the training of large-capacity networks like AlexNet and VGG16 on the Tiny ImageNet dataset.

4.3 差分隐私技术

差分隐私技术(differential privacy, DP)是通过添加噪声来保护隐私的一种密码学技术[150],因为加入少量噪声就可以取得较好的隐私保护效果,因此从它被 Dwork 等人[150]提出来就被广泛接受和使用.相比于前面 2 种密码学技术,差分隐私技术在实际场景中更易部署和应用[151].在机器学习中一般用来保护训练数据集和模型参数的隐私.DP 技术主要分为中心化差分隐私技术和本地化差分隐私技术,中心化差分隐私技术主要采用拉普拉斯机制[150](Laplace mechanism)、指数机制[152](exponential mechanism)等方法,而本地化差分隐私技术则采用随机响应[153-154](randomized response)方法.差分隐私技术保护用户隐私查询过程如图 7 所示:

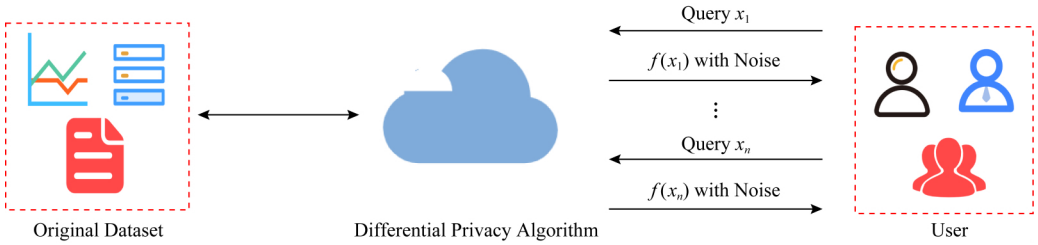


Fig. 7 Query process of schematic diagram of the differential privacy technology

图 7 差分隐私技术查询过程

Dwork 等人[150]提出的差分隐私技术在数据发布[155-156]、数据分析[157]、数据查询[158-159]、数据挖掘[150,152]等领域都受到了广泛的应用.Bindschaedler 等人[160]提出了一种合理可否认性(plausible deniability)的标准,保证了敏感数据的隐私.Huang 等人[117]提出了一种基于差分隐私的 ADMM 分布式学习算法 DP-ADMM,该算法在迭代过程中将近似增强拉格朗日函数与时变高斯噪声相结合,保证了良好的收敛性和模型准确性.Wang 等人[108]提出一种新的本地差分隐私机制来收集数值属性,并将其扩展到可以同时包含数字和类别属性的多维数据.他们实验证明,该算法可支持许多重要的机器学习任务,并比目前方案更加有效.

4.4 小 结

本节介绍了机器学习中常见的 3 种隐私保护技术,即同态加密技术、安全多方计算技术、差分隐私技术.并且根据各个技术的发展介绍了一些典型的隐私保护机器学习算法.随着深度学习的兴起,人工智能也迎来了发展契机,但是随着人工智能的广泛应用,其安全与隐私问题也越来越引起人们的关注,安全与隐私问题已经成为阻碍人工智能发展的绊脚石.

同态加密技术、安全多方计算技术和差分隐私技术是目前使用的比较广泛的技术,从通信、算力、隐私等多个角度考虑,这 3 种隐私保护技术各有本身适用的计算场景.同态加密技术具有计算开销大、效率低、可用性差、使用场景广的特点,它适合集中式学习、外包计算等场景;安全多方计算技术具有可用性高、通信开销大、效率低的特点,它适合分布式学习、联邦学习等场景;差分隐私技术具有计算开销小、效率高、可用性差的特点,它适合训练数据的收集、模型参数保护等场景.具体如表 6 所示:

Table 6 Applications Scenarios of the Three Privacy Preserving Technologies

表 6 3 种隐私保护技术适合应用的场景

Cryptography Technology	Advantages and Disadvantages	Application Scenarios
HE	High computation overhead, low efficiency, low availability, and wide usage scenarios	Centralized learning, outsourced computation, etc.
SMC	High availability, high communication overhead, and low efficiency	Distributed learning, federated learning, etc.
DP	Low computation overhead, high efficiency, and low availability	Training data collection, model parameters protection, etc.

5 总 结

综上所述,随着机器学习算法和人工智能应用领域的研究逐步深入,机器学习算法的特殊性给用户数据和网络模型的隐私保护带来巨大挑战,迫切需要进一步考虑更高的安全及隐私威胁,特别是可以实施恶意攻击的攻击者.

针对安全威胁,我们还需要进一步探索针对投毒攻击、对抗攻击等攻击手段的防御技术,提高模型的鲁棒性,研究更强攻击的防御手段.针对隐私威胁,全同态加密一直被认为是隐私保护机器学习的首选技术,但由于其具有数据扩张、计算负载、激活函数拟合误差等不利因素,使得基于安全多方计算的隐私保护机器学习得到了迅速发展.然而,现有隐私保护机器学习方案往往假设云服务器是诚实且好奇的被动攻击模型,考虑在云服务器互不合谋情况下数据的安全性与机器学习的可用性;并且,在更高的安全等级推广到多方场景下,还需考虑存在恶意攻击者情形下的公平性与一致性.因此,针对目前所提的方案,还需要再提高精度、效率,降低误差,并考虑更强的威胁场景,如恶意场景等.因此,未来的研究方向建议如下:

1) 建立完善统一的安全评估标准.对于大部分隐私数据,可在源头上控制好这些数据的使用范围和收集过程.但是由于目前缺乏合理完善的安全评估标准,各类机构对于隐私数据的使用和收集都没有统一的标准,因此不可避免地会造成隐私的泄露[21].建立完善统一的安全评估标准势在必行、刻不容缓.

2) 研究具有更强鲁棒性的隐私保护的机器学习模型.随着如对抗攻击、投毒攻击等攻击手段的发展,普通模型已经不能够满足隐私需求,模型的泄露给组织、机构带来的损失不可估量.研究能够抵抗更强攻击手段的高鲁棒性机器学习模型是未来的工作.

3) 考虑更强威胁场景的机器学习方案.分布式、联邦学习的场景是目前的趋势,但是对于大多数分布式的方案来说,它们都是考虑半诚实的威胁模型.因此需要将目前的半诚实模型推广到恶意模型,使机器学习方案在恶意攻击者存在的情况下依然能够保持公平性和一致性.

4) 提高现有方案的精度、效率.目前的大部分隐私保护方案都是基于同态加密、安全多方计算和差分隐私的,目前这几类技术的通信、计算等开销比较大,这大大降低了算法的效率,带来了不必要的资源浪费.并且这些方案存在精度丢失的问题,因此研究更加高效、精度更高的隐私保护方案是一个重要的研究方向.

参 考 文 献

[1] Julia C W. The Cambridge analytica scandal changed the world-but it didn't change Facebook[EB/OL]. The Guardian, 2019. (2019-03-18)[2020-07-20]. https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook

[2] Olivia S.Facial recognition's "dirty little secret": Millions of online photos scraped without consent[EB/OL]. NBC News, 2019. (2019-03-12)[2020-07-20]. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[3] Drew H. Thousands of zoom video calls left exposed on open Web[EB/OL]. The Washington Post, 2020-04-04.[2020-07-20]. https://www.washingtonpost.com/technology/2020/04/03/thousands-zoom-video-calls-left-exposed-open-web/

[4] Mitchell T M. Machine Learning[M]. New York: McGraw-Hill, 2003

[5] Liao Guohui, Liu Jiayong. Amalicious code detection method based on data mining and machine learning[J]. Journal of Information Securyity Research, 2016, 2(1): 74-79 (in Chinese)
(廖国辉, 刘嘉勇. 基于数据挖掘和机器学习的恶意代码检测方法[J]. 信息安全研究, 2016, 2(1): 74-79)

[6] Han Ying, Li Shanshan, Chen Fuming. The seismic anomaly data mining model based on machine learning[J]. Computer Simulation, 2014, 31(11): 319-322 (in Chinese)
(韩莹, 李姗姗, 陈福明. 基于机器学习的地震异常数据挖掘模型[J]. 计算机仿真, 2014, 31(11): 319-322)

[7] Chen Xueyun, Xiang Shiming, Liu Chenglin, et al. Vehicle-detection in satellite images by hybrid deep convolutional neural networks[J]. IEEE Geoence and Remote Sensing Letters, 2014, 11(10): 1797-1801

[8] Chen Sizhe, Wang Haipeng, Xu Feng, et al. Target classification using the deep convolutional networks for SAR images[J]. IEEE Transactions on Geoence and Remote Sensing, 2016, 54(8): 4806-4817

[9] Wittel G L, Wu S F. On attacking statistical spam filters[C/OL]//Proc of Conf on Email & Anti-spam. Mountain View: CEAS, 2004[2020-08-28]. http://www.dextremes.com/research/wittel-wu-ceas-2004.pdf

[10] Launchbury J, Archer D, DuBuisson T, et al. Application-scale secure multiparty computation[C]//Proc of European Symp on Programming Languages and Systems. Berlin: Springer, 2014: 8-26

[11] Ling Chentian. Evolutionaryneural network for credit card fraud detection[J]. Microelectronics & Computer, 2011, 28(10): 14-17 (in Chinese)
(凌晨添. 进化神经网络在信用卡欺诈检测中的应用[J]. 微电子学与计算机, 2011, 28(10): 14-17)

[12] Fu Kang, Cheng Dawei, Tu Yi, et al. Credit card fraud detection using convolutional neural networks[C]//Proc of Int Conf on Neural Information Processing. Berlin: Springer, 2016: 483-490

[13] Roy A, Sun J, Mahoney R, et al. Deep learning detecting fraud in credit card transactions[C]//Proc of 2018 Systems and Information Engineering Design Symp (SIEDS). Piscataway, NJ: IEEE, 2018: 129-134

[14] Acharya U R, Oh S L, Hagiwara Y, et al. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals [J]. Computers in Biology and Medicine, 2018, 100: 270-278

[15] Arabasadi Z, Alizadehsani R, Roshanzamir M, et al. Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm [J]. Computer Methods and Programs in Biomedicine, 2017, 141: 19-26

[16] Alpaydin E. Introduction to Machine Learning [M]. Cambridge, MA: MIT Press, 2010

[17] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning [C] //Proc of the 3rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2094-2100

[18] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning [J]. Machine Learning, 2010, 81(2): 121-148

[19] Biggio B, Fumera G, Roli F. Security evaluation of pattern classifiers under attack [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(4): 984-996

[20] Tan Zuowen, Zhang Lianfu. Survey on privacy preserving techniques for machine learning [J]. Journal of Software, 2020, 31(7): 2127-2156 (in Chinese)
(**谭作文, 张连福. 机器学习隐私保护研究综述**[J]. **软件学报**, 2020, 31(7): 2127-2156)

[21] Song Lei, Ma Chunguang, Duan Guanghan. Machine learning security and privacy: A survey [J]. Chinese Journal of Network and Information Security, 2018, 4(8): 1-11 (in Chinese)
(**宋蕾, 马春光, 段广晗. 机器学习安全及隐私保护研究进展**[J]. **网络与信息安全学报**, 2018, 4(8): 1-11)

[22] Moisejevs I. Willmy machine learning system be attacked? [EB/OL]. Towards Data Science. (2019-07-15) [2020-06-01]. https://towardsdatascience.com/will-my-machine-learning-be-attacked-6295707625d8

[23] Li Pan, Zhao Wentao, Liu Qiang, et al. Security issues and their countermeasuring techniques of machine learning: A survey [J]. Journal of Frontiers of Computer Science & Technology, 2018, 12(2): 171-184 (in Chinese)
(**李盼, 赵文涛, 刘强, 等. 机器学习安全性问题及其防御技术研究综述**[J]. **计算机科学与探索**, 2018, 12(2): 171-184)

[24] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning [C] //Proc of the 2017 ACM on Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519

[25] Laskov P. Practical evasion of a learning-based classifier: A case study [C] //Proc of 2014 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2014: 197-211

[26] Biggio B, Didaci L, Fumera G, et al. Poisoning attacks to compromise face templates [C] //Proc of 2013 Int Conf on Biometrics (ICB). Piscataway, NJ: IEEE, 2013: 1-7

[27] Biggio B, Fumera G, Roli F, et al. Poisoning adaptive biometric systems [C] //Proc of Joint IAPR Int Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Berlin: Springer, 2012: 417-425

[28] Biggio B, Fumera G, Roli F. Pattern recognition systems under attack: Design issues and research challenges [J/OL]. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28(7): [2020-08-29]. https://doi.org/10.1142/s0218001414600027

[29] Bogdanov D, Talviste R, Willemson J. Deploying secure multi-party computation for financial data analysis [C] //Proc of Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2012: 57-64

[30] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841

[31] Lim H S M, Taeihagh A. Algorithmic decision-making in AVs: Understanding ethical and technical concerns for smart cities [J].Sustainability, 2019, 11(20): 1-28

[32] Ackerman E. Slight street sign modifications can completely fool machine learning algorithms [J/OL]. IEEE Spectrum, 2017 [2020-06-03]. https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms

[33] Heaven D. Why deep-learning AIs are so easy to fool [J]. Nature, 2019, 574(7777): 163-166

[34] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models [C] //Proc of 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3-18

[35] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response [C] //Proc of the 2014 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 1054-1067

[36] Low Y, Gonzalez J, Kyrola A, et al. Distributed graphlab: A framework for machine learning in the cloud [J]. Proceedings of the VLDB Endowment, 2012, 5(8): 716-727

[37] Konečný J, McMahan H B, Ramage D, et al. Federated optimization: Distributed machine learning for on-device intelligence [OL]. arXiv preprint, 2016 [2020-07-20]. https://arxiv.org/abs/1610.02527

[38] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency [OL]. arXiv preprint, 2016 [2020-07-20]. https://arxiv.org/abs/1610.05492

[39] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging [OL]. arXiv preprint, 2016 [2020-07-20]. https://arxiv.org/abs/1602.05629

[40] Jiang Han, Liu Yiran, Song Xiangfu, et al. Cryptographic approaches for privacy-preserving machine learning [J]. Journal of Electronics and Information Technology, 2020, 42(5): 1068-1078 (in Chinese)

(蒋瀚,刘怡然,宋祥福,等.隐私保护机器学习的密码学方法[J].电子与信息学报,2020,42(5):1068-1078)

[41] Barreno M, Nelson B, Sears R, et al. Can machine learning be secure? [C] //Proc of the 2006 ACM Symp on Information, Computer and Communications Security. New York: ACM, 2006: 16-25

[42] Dalvi N, Domingos P, Sanghai S, et al. Adversarial classification [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 99-108

[43] Lowd D, Meek C. Good word attacks on statistical spam filters [C] //Proc of Conf on Email & Anti-spam. Mountain View: CEAS, 2005: 125-132

[44] Miller D J, Xiang Z, Kesidis G. Adversarial learning in statistical classification: A comprehensive review of defenses against attacks [OL]. arXiv preprint, arXiv: 1904.06292, 2019 [2020-07-20]. https://arxiv.org/abs/1904.06292

[45] Goldwasser S. Fromidea to impact, the crypto story: What's next? Cryptography & machine learning: What else? [C/OL] //IACR Distinguished Lecture of CRYPTO 2018. 2018 [2020-06-01]. https://www.iacr.org/cryptodb/data/paper.php?pubkey=29941

[46] Jiang Xiaoqian, Kim M, Lauter K, et al. Secure outsourced matrix computation and application to neural networks [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 1209-1222

[47] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310-1321

[48] Zheng Wenting, Popa R A, Gonzalez J E, et al. Helen: Maliciously secure coopetitive learning for linear models [C] //Proc of 2019 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 724-738

[49] Stoica I, Song D, Popa R A, et al. A Berkeley view of systems challenges for AI [OL]. arXiv preprint, arXiv: 1712.05855, 2017 [2020-07-20]. https://arxiv.org/abs/1712.05855

[50] Moisejevs I. Poisoning attacks on machine learning [EB/OL]. Towards Data Science. (2019-07-15) [2020-06-01]. https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db

[51] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks [C] //Proc of Network and Distributed System Security Symp. Reston, VA: Internet Society, 2018: 1-5

[52] Steinhardt J, Koh P W W, Liang P S. Certified defenses for data poisoning attacks [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 3517-3529

[53] Newell A, Potharaju R, Xiang Luojie, et al. On the practicality of integrity attacks on document-level sentiment analysis [C] //Proc of the 2014 Workshop on Artificial Intelligent and Security Workshop. New York: ACM, 2014: 83-93

[54] Biggio B, Rieck K, Ariu D, et al. Poisoning behavioral malware clustering [C] //Proc of the 2014 Workshop on Artificial Intelligent and Security Workshop. New York: ACM, 2014: 27-36

[55] Xiao Huang, Biggio B, Brown G, et al. Is feature selection secure against training data poisoning? [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2015: 1689-1698

[56] Newsome J, Karp B, Song D. Polygraph: Automatically generating signatures for polymorphic worms [C] //Proc of 2005 IEEE Symp on Security and Privacy (S&P'05). Piscataway, NJ: IEEE, 2005: 226-241

[57] Newsome J, Karp B, Song D. Paragraph: Thwarting signature learning by training maliciously [C] //Proc of Int Workshop on Recent Advances in Intrusion Detection. Berlin: Springer, 2006: 81-105

[58] Chung S P, Mok A K. Advanced allergy attacks: Does a corpus really help? [C] //Proc of Int Workshop on Recent Advances in Intrusion Detection. Berlin: Springer, 2007: 236-255

[59] Chung S P, Mok A K. Allergy attack against automatic signature generation [C] //Proc of Int Workshop on Recent Advances in Intrusion Detection.Berlin: Springer, 2006: 61-80

[60] Biggio B, Corona I, Nelson B, et al. Security Evaluation of Support Vector Machines in Adversarial Environments [M]. Berlin: Springer, 2014: 105-153

[61] Nelson B, Rubinstein B I P, Huang L, et al. Query strategies for evading convex-inducing classifiers [J]. Journal of Machine Learning Research, 2012, 13(5): 1293-1332

[62] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [OL]. arXiv preprint, 2013 [2020-07-20]. https://arxiv.org/abs/1312.6199

[63] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [OL]. arXiv preprint, 2014 [2020-07-20]. https://arxiv.org/abs/1412.6572

[64] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks [C] //Proc of the IEEE Conf onComputer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2574-2582

[65] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of 2016 IEEE European Symp onSecurity and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2016: 372-387

[66] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of 2017 IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2017: 39-57

[67] Ru B, Cobb A, Blaas A, et al. Bayesopt adversarial attack [C/OL] //Proc of Int Conf on Learning Representations. Addis Ababa (Virtual Conference): ICLR, 2020 [2020-08-29]. https://openreview.net/forum?id=Hkem-lrtvH

[68] Zhou Mingyi, Wu Jing, Liu Yipeng, et al. DaST: Data-freesubstitute training for adversarial attacks [OL]. arXiv preprint, 2020 [2020-07-20]. https://arxiv.org/abs/2003.12703

[69] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIS [C] //Proc of the 25th USENIX Security Symp. Berkeley, CA: USENIX Association, 2016: 601-618

[70] He Yingzhe, Hu Xingbo, He Jinwen, et al. Privacy and security issues in machine learning systems: A survey [J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070 (in Chinese)
(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)

[71] Shi Yi, Sagduyu Y, Grushin A. How to steal a machine learning classifier with deep learning [C] //Proc of 2017 IEEE Int Symp on Technologies for Homeland Security (HST). Piscataway, NJ: IEEE, 2017: 1-5

[72] Wang Binghui, Gong N Z. Stealing hyperparameters in machine learning [C] //Proc of 2018 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2018: 36-52

[73] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[74] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing [C] //Proc of the 23rd USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2014: 17-32

[75] Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers [J]. International Journal of Security and Networks, 2015, 10(3): 137-150

[76] Song Congzheng, Ristenpart T, Shmatikov V. Machine learning models that remember too much [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 587-601

[77] Carlini N, Liu Chang, Erlingsson Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks [C] //Proc of the 28th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2019: 267-284

[78] Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? Membership inference on aggregate location data [OL]. arXiv preprint, 2017 [2020-07-20]. https://arxiv.org/abs/1708.06145

[79] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting [C] //Proc of the 31st Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2018: 268-282

[80] Song Congzheng, Shmatikov V. Auditing data provenance in text-generation models [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2019: 196-206

[81] Truex S, Liu Ling, Gursoy M E, et al. Demystifying membership inference attacks in machine learning as a service [J]. IEEE Transactions on Services Computing, 2019: 1-1

[82] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership inference attacks against generative models [J]. Proceeding on Privacy Enhancing Technologies, 2019, 2019(1): 133-152

[83] Koh P W, Steinhardt J, Liang P. Stronger data poisoning attacks break data sanitization defenses [OL]. arXiv preprint, arXiv: 1811.00741, 2018 [2020-07-20]. https://arxiv.org/abs/1811.00741

[84] Baracaldo N, Chen B, Ludwig H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security. NewYork: ACM, 2017: 103-110

[85] Suciu O, Marginean R, Kaya Y, et al. When does machine learning FAIL? generalized transferability for evasionand poisoning attacks [C] //Proc of the 27th USENIX Security Symp (USENIX Security 18). Berkeley, CA: USENIX Associate, 2018: 1299-1316

[86] Huang Ruitong, Xu Bing, Schuurmans D, et al. Learning with a strong adversary [OL]. arXiv preprint, arXiv: 1511.03034, 2015 [2020-07-20]. https://arxiv.org/abs/1511.03034

[87] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [OL]. arXiv preprint, 2017 [2020-07-20]. https://arxiv.org/abs/1705.07204

[88] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale [OL]. arXiv preprint, arXiv: 1611.01236, 2016 [2020-07-20]. https://arxiv.org/abs/1611.01236

[89] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [OL]. arXiv preprint, 2017 [2020-07-20]. https://arxiv.org/abs/1706.06083

[90] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1660-1669

[91] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] //Proc of the 2016 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2016: 582-597

[92] Zantedeschi V, Nicolae M I, Rawat A. Efficient defenses against adversarial attacks [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security. NewYork: ACM, 2017: 39-49

[93] Liang Bin, Li Hongcheng, Su Miaoqiang, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction [J]. IEEE Transactions on Dependable and Secure Computing. Piscataway, NJ: IEEE, 2018: 1-1

[94] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [C/OL] //Proc of the 25th Network and Distributed System Security Symp. Reston, VA: The Internet Society, 2018 [2020-08-29]. http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-4_Xu_paper.pdf

[95] Lindell Y, Pinkas B. Privacy preserving data mining [C] //Proc of Annual Int Cryptology Conf. Berlin: Springer, 2000: 36-54

[96] Agrawal R, Srikant R. Privacy-preserving data mining [C] //Proc of the 2000 ACM SIGMOD Int Conf on Management of data. New York: ACM, 2000: 439-450

[97] Jagannathan G, Wright R N. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery in Data Mining. New York: ACM, 2005: 593-599

[98] Bunn P, Ostrovsky R. Secure two-party k-means clustering [C] //Proc of the 14th ACM Conf on Computer and Communications Security. New York: ACM, 2007: 486-497

[99] Yu H, Vaidya J, Jiang Xiaoqian. Privacy-preserving SVM classification on vertically partitioned data [C] //Proc of Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2006: 647-656

[100] Du Wenliang, Han Y S, Chen Shigang. Privacy-preserving multivariate statistical analysis: Linear regression and classification [C] //Proc of the 2004 SIAM Int Conf on Data Mining. Philadelphia, Pennsylvania: SIAM, 2004: 222-233

[101] Sanil A P, Karr A F, Lin Xiaodong, et al. Privacy preserving regression modelling via distributed computation [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 677-682

[102] Aono Y, Hayashi T, Trieu P L, et al. Scalable and secure logistic regression via homomorphic encryption [C] //Proc of the 6th ACM Conf on Data and Application Security and Privacy. New York: ACM, 2016: 142-144

[103] Nikolaenko V, Weinsberg U, Ioannidis S, et al. Privacy-preserving ridge regression on hundreds of millions of records [C] //Proc of 2013 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2013: 334-348

[104] Yao A C. How to generate and exchange secrets [C] //Proc of the 27th Annual Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 1986: 162-167

[105] Liu Jian, Juuti M, Lu Yao, et al. Oblivious neural network predictions via minionn transformations [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 619-631

[106] Sanyal A, Kusner M, Gascon A, et al. TAPAS: Tricks to accelerate (encrypted) prediction as a service [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2018: 4490-4499

[107] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 35-52

[108] Wang Ning, Xiao Xiaohui, Yang Yin, et al. Collecting and analyzing multidimensional data with local differential privacy [C] //Proc of 2019 IEEE 35th Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2019: 638-649

[109] Papernot N, Song S, Mironov I, et al. Scalable private learning with pate [OL]. arXiv preprint, 2018 [2020-07-20]. https://arxiv.org/abs/1802.08908

[110] Wagh S, Gupta D, Chandran N. SecureNN: Efficient and private neural network training [J/OL]. IACR Cryptology ePrint Archive, 2018 [2020-08-29]. https://eprint.iacr.org/2018/442

[111] Wagh S, Tople S, Benhamouda F, et al. FALCON: Honest-majority maliciously secure framework for private deep learning [OL]. arXiv preprint, 2020 [2020-07-20]. https://arxiv.org/abs/2004.02229

[112] Mandal K, Gong G. PrivFL: Practical privacy-preserving federated regressions on high-dimensional data over mobile networks [C] //Proc of the 2019 ACM SIGSAC Conf on Cloud Computing Security Workshop. New York: ACM, 2019: 57-68

[113] Gilad-Bachrach R, Dowlin N, Laine K, et al. CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2016: 201-210

[114] Hesamifard E, Takabi H, Ghasemi M. CryptoDL: Deep neural networks over encrypted data [J]. arXiv preprint, 2017 [2020-07-20]. https://arxiv.org/abs/1711.05189

[115] Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning [C] //Proc of 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 19-38

[116] Hesamifard E, Takabi H, Ghasemi M, et al. Privacy-preserving machine learning in cloud [C] //Proc of the 2017 on Cloud Computing Security Workshop. New York: ACM, 2017: 39-43

[117] Huang Zonghao, Hu Rui, Guo Yuanxiong, et al. DP-ADMM: ADMM-based distributed learning with differential privacy [J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 1002-1012

[118] Armknecht F, Boyd C, Carr C, et al. A guide to fully homomorphic encryption [J/OL]. IACR Cryptology ePrint Archive, 2015 [2020-08-29]. http://eprint.iacr.org/2015/1192

[119] Gentry C. Fully homomorphic encryption using ideal lattices [C] //Proc of the 41st Annual ACM Symp on Theory of Computing. New York: ACM, 2009: 169-178

[120] Gentry C, Boneh D. A fully Homomorphic Encryption Scheme [M]. Stanford: Stanford University, 2009

[121] Van Dijk M, Gentry C, Halevi S, et al. Fully homomorphic encryption over the integers [C] //Proc of Annual Int Conf on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2010: 24-43

[122] Coron J S, Naccache D, Tibouchi M. Public key compression and modulus switching for fully homomorphic encryption over the integers [C] //Proc of Annual Int Conf on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2012: 446-464

[123] Coron J S, Mandal A, Naccache D, et al. Fully homomorphic encryption over the integers with shorter public keys [C] //Proc of Annual Cryptology Conf. Berlin: Springer, 2011: 487-504

[124] Cheon J H, Coron J S, Kim J, et al. Batch fully homomorphic encryption over the integers [C] //Proc of Annual Int Conf on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2013: 315-335

[125] Coron J S, Lepoint T, Tibouchi M. Scale-invariant fully homomorphic encryption over the integers [C] //Proc of Int Workshop on Public Key Cryptography. Berlin: Springer, 2014: 311-328

[126] Rivest R L, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems [J]. Communications of the ACM, 1978, 21(2): 120-126

[127] ElGamal T. A public key cryptosystem and a signature scheme based on discrete logarithms [J]. IEEE Transactions on Information Theory, 1985, 31(4): 469-472

[128] Fan J, Vercauteren F. Somewhat practical fully homomorphic encryption [J/OL]. IACR Cryptology ePrint Archive, 2012 [2020-08-29]. http://eprint.iacr.org/2012/144

[129] Gentry C, Halevi S, Smart N P. Better bootstrapping in fully homomorphic encryption [C] //Proc of Int Workshop on Public Key Cryptography. Berlin: Springer, 2012: 1-16

[130] Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE [J]. SIAM Journal on Computing, 2014, 43(2): 831-871

[131] Brakerski Z. Fully homomorphic encryption without modulus switching from classical GapSVP [C] //Proc of Annual Cryptology Conf. Berlin: Springer, 2012: 868-886

[132] Chabanne H, de Wargny A, Milgram J, et al. Privacy-preserving classification on deep neural network [J]. IACR Cryptology ePrint Archive, 2017 [2020-08-29]. http://eprint.iacr.org/2017/035

[133] Chillotti I, Gama N, Georgieva M, et al. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds [C] //Proc of Int Conf on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2016: 3-33

[134] Bourse F, Minelli M, Minihold M, et al. Fast homomorphic evaluation of deep discretized neural networks [C] //Proc of Annual Int Cryptology Conf. Berlin: Springer, 2018: 483-512

[135] Henecka W, Kögl S, Sadeghi A R, et al. TASTY: Tool for automating secure two-party computations [C] //Proc of the 17th ACM Conf on Computer and Communications Security. New York: ACM, 2010: 451-462

[136] Demmler D, Schneider T, Zohner M. ABY-aframework for efficient mixed-protocol secure two-party computation [C/OL] //Proc of Network and Distributed System Security Symp. Reston, VA: Internet Society, 2015: 1-15. DOI: 10.14722/ndss.2015.23113

[137] Rouhani B D, Riazi M S, Koushanfar F. DeepSecure: Scalable provably-secure deep learning [C] //Proc of the 55th Annual Design Automation Conf. New York: ACM, 2018: 1-6

[138] Juvekar C, Vaikuntanathan V, Chandrakasan A. GAZELLE: A low latency framework for secure neural network inference [C] //Proc of the 27th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2018: 1651-1668

[139] Riazi M S, Weinert C, Tkachenko O, et al. Chameleon: A hybrid secure computation framework for machine learning applications [C] //Proc of the 2018 on Asia Conf on Computer and Communications Security. New York: ACM, 2018: 707-721

[140] Agrawal N, Shahin S A, Kusner M J, et al. QUOTIENT: Two-party secure neural network training and prediction [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 1231-1247

[141] Mishra P, Lehmkuhl R, Srinivasan A, et al. DELPHI: A cyptographic inference service for neural networks [C] //Proc of the 29th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2020: 2505-2522

[142] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1223-1231

[143] Abadi M, Barham P, Chen Jianmin, et al. Tensorflow: A system for large-scale machine learning [C] //Proc of the 12th USENIX Conf on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2016: 265-283

[144] Chaudhari H, Choudhury A, Patra A, et al. ASTRA: High throughput 3PC over rings with application to secure prediction [C] //Proc of the 2019 ACM SIGSAC Conf on Cloud Computing Security Workshop. New York: ACM, 2019: 81-92

[145] Byali M, Chaudhari H, Patra A, et al. FLASH: Fast and robust framework for privacy-preserving machine learning [J]. Proceeding on Privacy Enhancing Technologies, 2020, 2020(2): 459-480

[146] Chaudhari H, Rachuri R, Suresh A. Trident: Efficient 4PC framework for privacy preserving machine learning [C] // Proc of the 27th Annual Network and Distributed System Security Symp(NDSS 2020). San Diego,CA: ISOC, 2020: 23-26

[147] Patra A, Suresh A. BLAZE: Blazing fast privacy-preserving machine learning [J]. arXiv preprint, arXiv: 2005.09042, 2020 [2020-07-20]. https://arxiv.org/abs/2005.09042

[148] Gordon S D, Ranellucci S, Wang Xiao. Secure computation with low communication from cross-checking [C] //Proc of Int Conf on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2018: 59-85

[149] Damgård I, Escudero D, Frederiksen T, et al. New primitives for actively-secure MPC over rings with applications to private machine learning [C] //Proc of 2019 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 1102-1120

[150] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of Theory of Cryptography Conf. Berlin: Springer, 2006: 265-284

[151] Liu Junxu, Meng Xiaofeng. Survey on privacy-preserving machine learning [J]. Journal of Computer Research and Development, 2020, 57(2): 346-362 (in Chinese)
(**刘俊旭, 孟小峰. 机器学习的隐私保护研究综述**[J]. **计算机研究与发展**, 2020, 57(2): 346-362)

[152] McSherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the 48th Annual IEEE Symp on Foundations of Computer Science (FOCS'07). Piscataway, NJ: IEEE, 2007: 94-103

[153] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias [J]. Journal of the American Statistical Association, 1965, 60(309): 63-69

[154] Greenberg B G, Abul-Ela A L A, Simmons W R, et al. The unrelated question randomized response model: Theoretical framework [J]. Journal of the American Statistical Association, 1969, 64(326): 520-539

[155] Wang Qian, Zhang Yan, Lu Xiao, et al. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy [J]. IEEE Transactions on Dependable and Secure Computing, 2016, 15(4): 591-606

[156] Su Sen, Tang Peng, Cheng Xiang, et al. Differentially private multi-party high-dimensional data publishing [C] // Proc of 2016 IEEE 32nd Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2016: 205-216

[157] Zhang Tao, Zhu Quanyan. Dynamic differential privacy for ADMM-based distributed classification learning [J]. IEEE Transactions on Information Forensics and Security, 2016, 12(1): 172-187

[158] Bun M, Steinke T, Ullman J. Make up your mind: The price of online queries in differential privacy [C] //Proc of the 28th Annual ACM-SIAM Symp on Discrete Algorithms. Philadelphia, Pennsylvania: SIAM, 2017: 1306-1325

[159] Yuan Ganzhao, Yang Ying, Zhang Zhenjie, et al. Convex optimization for linear query processing under approximate differential privacy [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 2005-2014

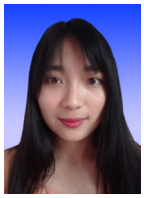
[160] Bindschaedler V, Shokri R, Gunter C A. Plausible deniability for privacy-preserving data synthesis [J]. arXiv preprint, arXiv: 1708.07975, 2017 [2020-07-20]. https://arxiv.org/abs/1708.07975

Wei Lifei, born in 1982. PhD. Associate professor and master supervisor. Senior member of CCF. His main research interests include information security, privacy preserving and cryptography.

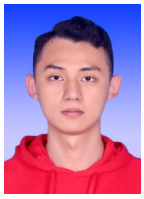
Chen Congcong, born in 1996. Master candidate. Student member of CCF. His main research interests include machine learning, secure computation and information security. (chencongcong0302@163.com)

Zhang Lei, born in 1983. PhD. Assistant professor. Her main research interests include applied cryptography, big data security and access control.

Li Mengsi, born in 1994. Master candidate. Student member of CCF. Her main research interests include machine learning and information security. (lims1021@163.com)



Chen Yujiao, born in 1996. Master candidate. Student member of CCF. Her main research interests include machine learning and information security. (270835899@qq.com)



Wang Qin, born in 1996. Master candidate. Student member of CCF. His main research interests include information security and secure computation. (913377391@qq.com)