

机器学习隐私保护研究综述*

谭作文, 张连福

(江西财经大学 信息管理学院计算机科学与技术系, 江西 南昌 330013)

通讯作者: 张连福, E-mail: zlf_jx@163.com



摘要: 机器学习已成为大数据、物联网和云计算等领域核心技术,机器学习模型训练需要大量数据,这些数据通常通过众包方式收集,里面含有大量隐私数据包括个人身份信息(如电话号码、身份证号等)、敏感信息(如金融财务、医疗健康等信息),如何低成本且高效地保护这些数据是一个重要的问题.介绍了机器学习及其隐私定义和隐私威胁,重点对机器学习隐私保护主流技术的工作原理和突出特点进行了阐述,并分别按照差分隐私、同态加密和安全多方计算等机制对机器学习隐私保护领域的研究成果进行了综述.在此基础上,对比分析了机器学习不同隐私保护机制的主要优缺点.最后,对机器学习隐私保护的发展趋势进行展望,并提出了该领域未来可能的研究方向.

关键词: 机器学习;隐私保护;差分隐私;同态加密;安全多方计算

中图法分类号: TP309

中文引用格式: 谭作文,张连福.机器学习隐私保护研究综述.软件学报.<http://www.jos.org.cn/1000-9825/6052.htm>

英文引用格式: Tan ZW, Zhang LF. Survey on privacy preserving techniques for machine learning. Ruan Jian Xue Bao/Journal of Software, 2020 (in Chinese). <http://www.jos.org.cn/1000-9825/6052.htm>

Survey on Privacy Preserving Techniques for Machine Learning

TAN Zuo-Wen, ZHANG Lian-Fu

(Department of Computer Science and Technology, School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China)

Abstract: Machine learning has become a core technology in areas such as big data, Internet of things and cloud computing. Training machine learning models requires a large amount of data, which is often collected by means of crowdsourcing and contains a large number of private data including personally identifiable information (such as phone number, id number, etc.) and sensitive information (such as financial data, health care, etc.). How to protect these data with low cost and high efficiency is an important issue. This paper first introduces the concept of machine learning, explains various definitions of privacy in machine learning and demonstrates all kinds of privacy threats encountered in machine learning, then continues to elaborate on the working principle and outstanding features of the mainstream technology of machine learning privacy protection. According to differential privacy, homomorphic encryption and secure multi-party computing, the research achievements in the field of machine learning privacy protection are summarized respectively. On this basis, the paper comparatively analyzes the main advantages and disadvantages of different mechanisms of privacy preserving for machine learning. Finally, the developing trend of privacy preserving for machine learning is prospected, and the possible research directions in this field are proposed.

Key words: machine learning; privacy-preserving; differential privacy; homomorphic encryption; secure multiparty computation

* 基金项目: 国家自然科学基金(61862028,61702238); 江西省自然科学基金(20181BAB202016); 江西省教育厅科技项目(GJJ160430); 江西省教育厅青年科技项目(GJJ180288)

Foundation item: National Natural Science Foundation of China (61862028,61702238); Natural Science Foundation of Jiangxi Province, China (20181BAB202016); Science and Technology Project of Provincial Education Department of Jiangxi (GJJ160430); Young Science and Technology Project of Provincial Education Department of Jiangxi (GJJ180288).

收稿时间: 2019-09-10; 修改时间: 2020-02-09, 2020-03-20; 采用时间: 2020-04-09; jos 在线出版时间: 2020-04-21

近年来,机器学习(machine learning,简称 ML) 发展迅速,已经成为图像处理、语音识别和网络空间安全等领域的基石.另一方面,得益于计算机技术、存储技术和网络技术的发展,政府、医院、银行等各类机构及电子商务、零售、供应链等各类平台的数据量呈指数级增长.不仅如此,物联网、社交媒体和智能手机等媒介每分钟也产生大量数据.数据持有者可以将这些数据发送给云服务提供商(cloud service provider,简称 CSP),以识别出潜在的数据模型.这些模型可能有助于支持决策,改进业务,为客户提供增值服务^[1]、预测服务和推荐服务^[2]等.

在此背景下,许多 CSP 纷纷推出机器学习即服务(Machine Learning as a Service,简称 MLaaS).这些 MLaaS 为数据持有者提供基于机器学习的数据处理、模型训练、预测服务和部署等自动化解决方案,吸引机器学习实践者在云平台部署应用程序,而无需建立自己的大规模基础设施和计算资源.著名的 MLaaS 平台包括 Google Prediction API^[3]、Amazon ML^[4]、Microsoft Azure ML^[5]和 BigML^[6]等.典型的基于云平台的机器学习体系结构如图 1 所示.这里的 CSP 可以是第三方 MLaaS 平台、合作伙伴公司甚至公司本身在场外或在某些独立设施中运行的应用程序.数据持有者是政府、银行、医院、保险公司或电子商务网站等,他们可以选择在云平台中存储、处理数据或使用云平台提供的服务.终端用户是使用部署在云平台中的服务的参与者,例如企业员工、医生和诊所员工等.终端用户将预测请求上传给 CSP,CSP 将 ML 模型的预测结果返回给终端用户.

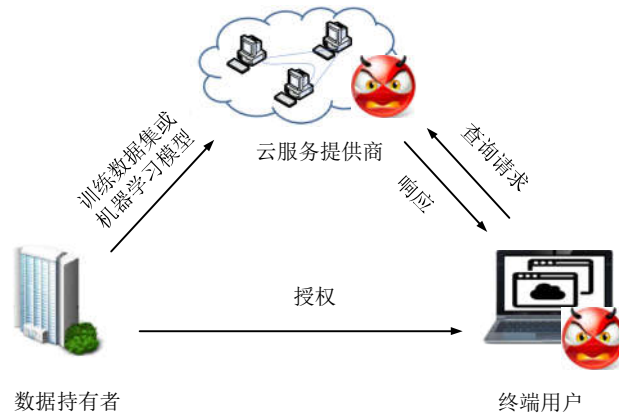


Fig.1 Architecture and privacy threat model of machine learning based on cloud platform

图 1 基于云平台的机器学习体系结构及隐私威胁模型

尽管 MLaaS 提供了诱人的好处,但也存在严重的问题,即用户数据的安全和隐私会受到各种威胁,如图 1 所示.首先,在训练阶段,恶意 CSP 只要对训练算法进行相对较小的修改,就可生成高质量模型,并且满足标准 ML 度量(如准确性和可泛化性),或者获得对它的输入-输出访问权,最终从模型中提取出关于训练数据的详细信息^[7].即使恶意 CSP 不能直接访问数据集,也可从模型参数中提取关于训练数据的敏感信息^[8].其次,预测阶段隐私泄露问题.目前已有部分研究开始关注预测数据隐私问题^[9-11].在模型预测服务中,客户需要将预先训练好的模型上传到 CSP.但模型泄露会导致数据持有者利益的损失,甚至破坏原始数据.另外,即使只有黑盒访问权限的恶意远程用户仍然可以利用精心设计的输入查询模型输出,从而获得有关训练数据的信息^[12-17].机器学习中的隐私泄露问题已成为云计算发展面临的一个重大挑战.

另一方面,隐私权作为一项基本人权,对个人和企业来说都极其重要,重视数据隐私和安全保护已经成为了世界性的趋势.欧盟于 2018 年 5 月 25 日正式实施的《通用数据保护条例》(General Data Protection Regulation, 简称 GDPR)^[18]要求企业对用户数据的处理应建立在用户的明确同意基础之上,企业应赋予用户“被遗忘权”,即用户可以随时删除或撤回其个人数据.被称为美国最严隐私法案的《加利福尼亚消费者隐私法案》(California Consumer Privacy Act, 简称 CCPA)^[19]已于 2020 年 1 月 1 日正式生效.它旨在加强消费者隐私权和数据安全保护,违反该法案的企业将遭到严厉惩罚.我国在 2017 年 6 月起实施的《中华人民共和国网络安全法》^[20]指出,任何个人和组织不得窃取或者以其他非法方式获取个人信息,未经被收集者同意,不得向他人提供个人信息.这些法规的建立不同程度上对人工智能传统的数据处理模式提出了新的挑战.

本文首先介绍机器学习隐私保护背景知识,包括机器学习概述、机器学习隐私定义、机器学习敌手模型和机器学习隐私保护场景(见第 1 节);然后讨论了机器学习中的典型隐私威胁以及机器学习隐私保护方案的分类情况(见第 2 节);接着分类研究了各种典型机器学习隐私保护机制,分析了各类隐私保护技术的相关概念、典型方案及其隐私保护场景,并对每一大类隐私保护技术进行了高层次的总结(见第 3~5 节);最后总结并展望该领域未来可能的研究方向及发展趋势(见第 6 节).

1 背景知识

1.1 机器学习概述

机器学习是一个涉及多学科的研究领域,包括计算机科学、概率与统计学、心理学和脑科学等学科.机器学习利用计算机有效地模仿人类的学习活动,通过对现有数据进行学习,产生有用的模型进而对未来的行为做出决策判断.根据用来学习的数据性质来分,机器学习可分为监督学习、半监督学习、无监督学习和强化学习四大类.

机器学习解决问题的过程分为训练阶段和预测阶段.在训练结束后获得目标模型,人们可以利用目标模型进行预测.以监督学习为例,其机器学习模型是一个参数化函数 $f_{\theta}:X \rightarrow Y$,将输入数据 $x \in X$ (特征)映射到输出数据 $y \in Y$ (标签).对于一个分类问题而言, X 是一个 d 维向量空间, Y 则是一组离散的类.根据这个函数能够对新数据准确地进行分类.机器学习模型的训练过程本质是寻找最优参数 θ 的过程,其中参数 θ 可以准确地反映 X 和 Y 的关系.拥有 N 个训练样本的数据集,可利用公式(1)所示的损失函数 ℓ 来测量真实输出和预测输出之间的误差.模型训练目的是使损失函数最小化,训练结束后可得最优模型参数 θ^* .

$$\theta^* = \arg \min_{\theta} \Omega(\theta) + \frac{1}{N} \sum_{i=1}^N \ell(y_i, f_{\theta}(x_i)) \quad (1)$$

其中, $\Omega(\theta)$ 是正则化惩罚项,用于防止过度拟合.

根据数据在模型训练前是否被集中收集,机器学习模型训练方式可分为集中式学习、分布式学习和联邦学习3类.

1) 集中式学习

在集中式学习(Centralized learning)中,各参与方训练数据集中在中央服务器,如图 2(a)所示.优点是模型训练和部署都很方便,而且大大提高了模型训练的准确性,因而在实际场景中具有广泛应用.缺点是给中央服务器的存储和运算资源带来了高负载,尤其是在大数据时代,而且所有的用户数据都将面临安全和隐私风险,即数据一旦上传到中央服务器,用户便很难再拥有对数据的控制权、知情权,即数据将被用于何处,是否未经授权便转让给第三方也不得而知.针对集中式学习模式下机器学习的隐私保护在过去几十年间已得到了广泛研究.

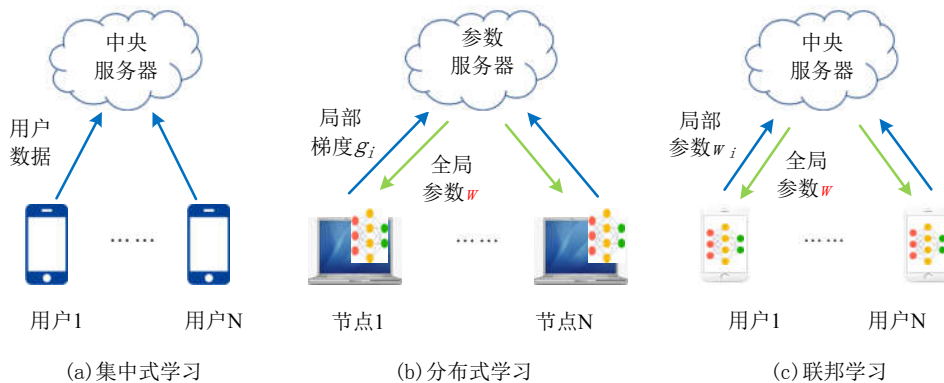


Fig.2 Model training methods in machine learning

图2 机器学习模型训练方式

2) 分布式学习

在分布式学习(Distributed learning)中,训练数据和计算负载都分布在各工作节点上,中央服务器仅维护全局参数,计算量较小.它们共同训练一个机器学习模型.参数服务器(Parameter Server)^[21]是分布式机器学习的一个典型例子,如图 2(b)所示.模型训练过程如下:首先,各工作节点在获得中心模型参数 w 后利用本地数据进行单独训练,并将训练后更新的梯度参数 g_i 上传至中央服务器;然后,中央服务器按式(2)将所有上传的梯度参数整合至中心模型,并再次将模型参数分发出去;如此迭代,直至最后收敛.在分布式学习中,中央服务器始终占据主导地位,各节点与中央服务器连接稳定,并且负载均衡,计算性能相当.

$$w' \leftarrow w - \eta \sum_{i=1}^N g_i \quad (2)$$

3) 联邦学习

联邦学习(Federated learning,简称 FL)可以看作是一种特殊的分布式机器学习.在 FL 中,多个客户端在中央服务器的协调

下联合训练一个模型,同时保持训练数据分散.联邦平均(Federated Averaging)算法^[22]是联邦学习中最流行的方法之一,如图2(c)所示.一个典型的 FL 训练过程如下^[23]:首先,服务器抽取一组满足条件的客户端;被选中的客户端从服务器下载当前模型权重参数和一个训练程序;然后,客户端在本地计算对模型参数的更新;接着,服务器收集客户端上传的参数.为了提高效率,一旦有足够数量的设备报告了结果,掉队的设备可能会在此时被丢弃;最后,服务器更新共享模型.如此迭代,直至收敛.在 FL 中,各参与方对自己的设备和数据拥有绝对的控制权,可以自主决定何时加入或退出联邦学习.各参与方的负载不平衡,并且可能需要处理非独立同分布(Non-IID)数据.因此,联邦学习面对的是一个更加复杂的学习环境^[24].

1.2 机器学习隐私定义

隐私是一个复杂的概念,目前还没有一个公认的标准定义.1890年发表在《哈佛法律评论》上的《论隐私权》^[25]将隐私定义为“不受打扰的权利”.1966年联合国大会通过的《公民权利和政治权利国际公约》^[26]将隐私定义为“任何人的私生活、家庭、住宅和通信不得任意或非法干涉,其荣誉和名誉不得加以攻击.人人有权享受法律保护,以免受非法干涉或攻击”.Saltzer等人^[27]将隐私定义为“个人(或组织)确定是否、何时、向谁公开个人(或组织)的信息的能力”.我国学者Zhou等人^[28]将隐私定义为“数据拥有者不愿意被披露的敏感数据或数据所表征的特性”.

根据机器学习隐私保护内容的不同,可将机器学习隐私分为训练数据隐私、模型隐私与预测结果隐私.

定义 1. 训练数据隐私. 训练数据隐私指机器学习中用户数据的个人信息(personally identifiable information,简称PII)和敏感信息.

个人信息是指能够唯一标识个人身份的信息,可分为标识符和准标识符.标识符包括姓名、身份证号、电话号码、电子邮件地址等主属性(key attributes).准标识符(quasi-identifier)指可以唯一地标识个体身份的属性集合,如(地址、性别、出生日期).敏感信息包括个体的人口统计学信息,如性别、薪水、犯罪记录等;财务信息,如信用卡号、帐户余额、交易记录等;健康信息,如病史、疾病症状、医学影像、医疗处方等;日常活动信息,如通话记录、活动轨迹、购物记录等.

定义 2. 模型隐私. 模型隐私指机器学习中模型训练算法、模型拓扑结构、模型权重参数、激活函数以及超参数等与机器学习模型有关的隐私信息.

如图3所示的加密预测服务(Encrypted Prediction as a Service,简称EPAAS)^[29]中,机器学习模型属于服务提供者的隐私信息,授权用户也只有使用权.但攻击者出于以下动机可能对模型发动模型提取攻击(Model Extraction Attack): 试图发起跨用户的模型提取攻击,窃取机器学习模型供后续自由免费使用;规避垃圾邮件的识别、恶意软件分类等敌对行为的检测;泄露有关敏感训练数据的信息等.

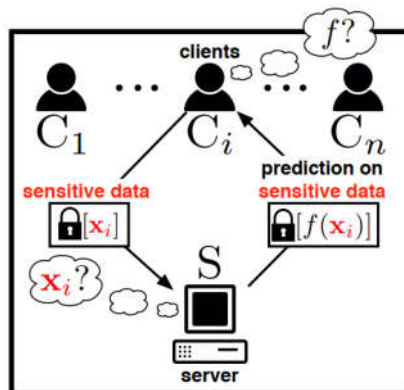


Fig.3 Architecture of EPAAS^[29]

图3 加密预测服务架构图^[29]

定义 3. 预测结果隐私. 预测结果隐私是指机器学习中模型对用户的预测输入请求反馈回来的、用户不愿意公开的敏感信息.

模型预测结果可能是用户的疾病诊断信息,例如,患某种疾病的概率.这些信息对于用户来说,属于个人隐私信息,但不可信服务提供商或者第三者可能窃取用户的此类信息.Xie等人^[30]提出的基于同态加密技术的隐私保护神经网络模型 crypto-nets,可以应用于加密数据,直接在密文上做预测,并返回加密预测结果,为在线医疗诊断模型预测结果提供了隐私保证.

训练数据隐私、模型隐私、模型预测结果隐私是在使用机器学习时需要重点保护的内容.这些信息一旦泄漏将会危及到用户敏感数据的安全或给服务提供商带来巨大的经济损失.这也是云计算发展面临的主要障碍.因此,基于云计算的机器学习服务系统应该更加重视隐私问题,不断提高隐私防护能力.

1.3 机器学习隐私攻击敌手模型

机器学习隐私攻击敌手模型包括敌手目标、敌手知识、敌手能力和敌手策略.表 1 总结了机器学习中隐私攻击敌手模型.

Table 1 Adversarial model of privacy attack

表 1 隐私攻击敌手模型

敌手目标	训练数据隐私、模型隐私、预测结果隐私	
敌手知识	白盒	掌握模型结构、模型参数、全部或部分训练数据
	黑盒	无模型的相关知识
敌手能力	强敌手	参与模型训练、收集模型或训练数据信息
	弱敌手	访问模型、收集模型信息
敌手策略	模型逆向攻击、模型提取攻击、成员推断攻击	

(1) 敌手目标

敌手针对机器学习模型的隐私攻击目标是破坏机器学习模型的机密性(Confidentiality),即敌手尽力获取机器学习中训练数据隐私、模型隐私与模型预测结果信息.机器学习隐私攻击中的敌手目标是模型的机密性.一个设计良好的 ML 系统应确保重要信息不被未经授权的用户获取.例如,一个基于 ML 的医疗诊断系统应防止敌手分析模型和恢复关于病人的信息^[31].当模型本身代表知识产权时,它要求模型及其参数是机密的,例如金融市场系统^[32].

(2) 敌手知识

敌手知识是指敌手所掌握的关于目标模型及其在目标环境中使用的信息量,包括模型训练数据集的分布情况、模型结构和参数、决策函数等.根据敌手掌握的关于机器学习模型信息量的多少,攻击方式可分为白盒攻击和黑盒攻击.在白盒攻击中,敌手掌握一些关于模型或训练数据的信息,例如机器学习模型结构、模型参数、部分或完整的训练数据;相反,黑盒攻击假定敌手没有关于模型的相关知识,敌手利用模型的脆弱性以及过去的输入来推断模型的信息.例如,敌手通过提供一系列精心设计的输入来观察模型的输出^[33].

(3) 敌手能力

敌手能力是指敌手可用的攻击内容和方式.在数据收集阶段,敌手能力为直接获取用户数据;在机器学习的训练阶段,敌手能力包括干预模型训练、访问训练数据、收集中间结果等;在机器学习的预测阶段,敌手能力是指访问模型、获取训练数据等能力.根据敌手对模型、数据控制能力和破坏力的不同可分为强敌手和弱敌手.强敌手的攻击能力包括参与模型的训练、收集模型或训练数据信息;弱敌手不直接参与模型训练,只是使用攻击来收集关于模型特征等信息^[34].

(4) 敌手策略

敌手策略是指敌手为达到攻击目标,所采取的具体攻击方式.敌手目标、敌手知识、敌手能力三者共同决定攻击者采取的敌手策略.除了数据收集阶段是直接访问数据的方式,在机器学习的训练和预测阶段,敌手策略可分为:直接攻击和间接攻击.直接攻击是指攻击者直接从模型预测结果中提取出训练数据信息或者判断某成员是否在某个模型的训练数据集中.间接攻击是指攻击者首先窃取模型参数,构建一个替代模型,然后利用该替代模型提取出模型的训练数据集相关信息.

敌手策略具体包括:模型逆向攻击、模型提取攻击和成员推断攻击.其中模型逆向攻击和成员推断攻击为直接攻击策略,模型提取攻击为间接攻击策略.

1.4 机器学习隐私保护场景

隐私保护场景是指机器学习中可能造成隐私泄露、需要采取措施进行隐私防护的特定场景.不同的隐私保护技术适用于不同的隐私保护场景,了解隐私保护场景是设计隐私保护方案的前提.机器学习所处的阶段、模型的训练方式、训练数据的分布与参与方的可信程度等因素决定了机器学习隐私保护场景.集中式学习^[35]中数据收集阶段、联邦学习^[36]中模型训练阶段、以病人为中心的在线医疗系统^[30]模型预测服务期间、基于云平台的在线金融系统^[32]在线服务期间都是典型的机器学习隐私保护场景.下面重点介绍一下前 2 个场景.

(1) 集中式学习

集中式学习最大的特点就是系统部署简单,无需考虑如何对服务进行多节点部署,不用考虑众多节点之间的分布式协作问题.在集中式学习中,中心服务器完成训练数据收集、机器学习模型训练、模型发布和模型预测等整个机器学习流程.在这些流程中,都可能存在各种隐私攻击.因此,在集中式学习中,整个机器学习环节都是需要重点关注的隐私保护场景.其中,在数据收集阶段,由于目前缺乏有关数据收集的统一标准,不可信的数据收集者可能过度收集用户数据并贩卖用户隐私.这种窃取用户原始数据的方式是机器学习系统中最典型的隐私保护场景.目前苹果和谷歌^[35]等公司已采用本地化差分隐私(local differential privacy,简称 LDP)技术来保护用户在数据收集阶段的数据隐私.

(2) 联邦学习

联邦学习不需要在云端集中存储用户数据,在隐私保护上具有更大的价值,但在模型训练阶段,它还可能遭受各种恶意攻击. Shokri 等人^[37]研究表明:在联邦学习中,一个好奇的参数服务器甚至一个参与者可以对其他参与者实施精确得惊人的成员资格推断攻击.对于运行在 CIFAR100 数据集上的 DenseNet 模型,好奇的中央参数服务器通过从所有参与者那里接收单个参数更新,可以实现 79.2%的成员推断准确性;本地参与者通过观察参数服务器的聚合参数更新,也可以获得 72.2%的成员推断精度.敌手还可以主动利用 SGD 泄露参与者训练数据的更多信息.中央服务器通过在参数更新过程中隔离参与者,在 DenseNet 模型上的主动推理攻击准确率可以提高到 87.3%.恶意参与者通过观察全局参数变化以及自己的对抗性参数更新,可以获得关于其他参与者的训练数据信息.因此,在联邦学习中,模型训练阶段是需重点关注的隐私保护场景.

2 机器学习典型隐私威胁与隐私保护方案

机器学习已经形成了一个商业模式,尽管 MlaaS 给用户带来极大的便利,但同时也将数据持有者的隐私数据暴露在了攻击者的各种攻击之下,因此有必要了解目前机器学习中典型的隐私威胁.

2.1 机器学习典型隐私威胁

现在流行的 MlaaS 一般包括机器学习模型训练和提供模型预测服务两个阶段.在这两个阶段,可能面临的典型隐私攻击主要有:模型逆向攻击、模型提取攻击和成员推断攻击.表 2 给出了机器学习中典型的隐私威胁.

Table 2 Typical privacy threats in machine learning

表 2 机器学习中典型的隐私威胁

阶段	模型逆向攻击	模型提取攻击	成员推断攻击
训练阶段	[7, 8, 31, 38]	[7]	
预测阶段	[12-17][39-40]	[41]	[40] [42]

(1) 模型逆向攻击(Model Inversion Attack)

模型逆向攻击是指攻击者从模型预测结果中提取和训练数据有关的信息^[12].这种攻击手段结合生成对抗网络后,尤为见效.Fredrikson 等人^[13]对基于线性回归算法的定制药物医疗系统实施了一种反向攻击,不仅泄露了病人的隐私,还可能导致药物的错误配置,从而危及患者生命. Fredrikson 等人^[12]分析了从已知模型中检索原始学习数据的可行性,他们成功地利用基于神经网络的人脸识别模型重建了人脸图像. Hitaj 等人^[38]的研究表明,分布式或联邦机器学习结构很难保护诚实参与者的训练数据集免遭基于 GAN 的攻击(GAN-based attack).一个基于 GAN 的对手可能愚弄受害者,让他们透露出更多的隐私信息. Ateniese 等人^[8]构建了一个新的元分类器(meta-classifier),并对其进行训练,使其能够攻击其他的分类器,从而获得它们训练数据集的敏感信息.例如供应商利用这种信息泄露,可以直接从竞争对手的设备上获取贸易证书,侵犯竞争对手知识产权.

(2) 模型提取攻击(Model Extraction Attack)

模型提取攻击是指攻击者获得对某个目标模型的黑盒访问权后,取得模型内部的参数或结构,或是试图构造出一个与目标模型近似甚至完全等价的机器学习模型^[41]. Song 等人^[7]证实了恶意机器学习算法可以创建满足精度和泛化要求的高质量模型,同时泄漏大量关于其训练数据集的信息,即使对手只有该模型的黑盒访问权,并指出机器学习模型不能盲目地应用于敏感数据,特别是如果模型训练代码是由另一方提供的. Florian Tramer 等人^[41]发现,敌手通过有限次访问预测服务的 API 接口,可以提取出模型的信息.对于一个 N 维的线性模型,理论上通过 $N+1$ 次查询访问就能够窃取到这个模型.

(3) 成员推断攻击(Membership Inference Attack)

成员推断攻击是指攻击者通过访问模型预测 API,从预测结果中获知某个特征数据是否包含在模型的训练集中^[40].在这种攻击中,攻击者仅需要得到预测分类的置信度,不需要知道模型结构、训练方法、模型参数、训练数据集分布等信息.对于

过拟合的模型,这种攻击尤其有效^[40]. Shokri 等人^[40]利用成员推理攻击,推测出某一数据是否在训练数据集中.Melis 等人^[42]证明了在协作机器学习和联合学习中,敌手不仅可在其他参与者的训练数据中推断出准确的数据点(如特定的位置)的存在(成员推理攻击),还可推断出其他参与者的训练数据的属性(属性推断攻击),并且可推断出某个属性在训练期间什么时候在数据中出现和消失.例如,确定某个特定的人何时第一次出现在用于训练通用性别分类器的照片中.

2.2 机器学习隐私保护方案分类

机器学习隐私保护可按机器学习模型的种类、机器学习过程、模型训练方式和隐私保护技术等进行分类,如表 3 所示.

Table 3 Classification of privacy protection schemes in machine learning

表 3 机器学习隐私保护方案分类

分类		典型方案				
按机器学习模型的种类分	监督学习中的隐私保护	线性回归	Ref [13]	Ref [43]		
		逻辑回归	Ref [44]			
		支持向量机	Ref [9]			
		决策树与随机森林	Ref [45]			
		极限学习	Ref [46]			
		贝叶斯算法	Ref [47]	Ref [48]		
	神经网络	DP-GANs ^[49]	AdLM ^[50]	Ref [51]	Ref [52]	
		Ref [53]	crypto-nets ^[30]	CryptoNets ^[54]	Ref [55]	
		CryptoDL ^[56]	pCDBN ^[57]	LPP-CNN ^[58]	OPSR ^[59]	
	半监督学习中的隐私保护		PATE ^[60]			
无监督学习中的隐私保护	k-Means	Ref [61]	Ref [62]			
强化学习中的隐私保护	Q-learning	LiPSG ^[63]				
按机器学习过程分	训练阶段中的隐私保护		Ref [48]	Ref [52]	Ref [53]	Ref [64]
			PPDL ^[65]	Ref [66]		
预测阶段中的隐私保护			Ref [9]	Ref [47]	crypto-nets ^[30]	CryptoNets ^[54]
			Ref [55]	CryptoDL ^[56]	TAPAS ^[29]	Ref [67]
			Ref [68]	Ref [69]	FHE-DiNN ^[70]	
按模型训练方式分	集中式学习中的隐私保护		Ref [9]	Ref [53]	Ref [64]	Ref [66]
	分布式学习中的隐私保护		Ref [21]	Ref [43]	Ref [45]	Ref [46]
			Ref [61]	Ref [71]	Ref [72]	Ref [73]
联邦学习中的隐私保护		Ref [22]	Ref [23]	Ref [24]	Ref [37]	
按隐私保护技术分	差分隐私	输入扰动	DP-GANs ^[49]	DPGAN ^[74]	Ref [75]	
		中间参数扰动	AdLM ^[50]	Ref [36]	Ref [76]	Ref [77]
		目标扰动	Ref [44]	dPAs ^[78]	pCDBN ^[57]	Ref [51]
		输出扰动	Ref [73]	PATE ^[60]	Ref [79]	
	同态加密	无需多项式近似	Ref [9]	Ref [47]	Ref [48]	Ref [52]
			Ref [53]	TAPAS ^[29]	Ref [67]	Ref [68]
		多项式近似	Ref [69]	FHE-DiNN ^[70]	Ref [80]	Ref [81]
			crypto-nets ^[30]	CryptoNets ^[54]	Ref [55]	CryptoDL ^[56]
	安全多方计算	传统分布式学习	Ref [45]	Ref [46]	Ref [61]	Ref [71]
			Ref [72]	Ref [82]	Ref [83]	Ref [84]
			Ref [85]	Ref [62]		
		基于 2PC 架构	Ref [43]	SecureML ^[86]	DeepSecure ^[87]	MiniONN ^[88]
EzPC ^[89]			Chameleon ^[90]	GAZELLE ^[91]	TASTY ^[92]	
LPP-CNN ^[58]			POR ^[93]	OPSR ^[59]	LiPSG ^[63]	

(1) 按机器学习模型的种类分类

机器学习的类型通常分为监督学习、半监督学习、无监督学习和强化学习等四类.典型的监督学习有:线性回归(Linear Regression)、逻辑回归(Logistic Regression)、支持向量机(Support Vector Machines (SVMs)、决策树与随机森林(Decision Trees and Random Forests)、神经网络(Neural networks)等,半监督学习有生成式模型 (Generative semi-supervised models)等,无监督学习有 k-Means 等,强化学习有 Q-learning 等.由于深度学习目前应用广泛,在各领域深受追捧,所以基于神经网络模型及其变种的隐私保护方法是本文论述的重点,另外,生成对抗网络在近两年发展迅速,所以在本文也占了一定的篇幅.

(2) 按机器学习过程分类

机器学习的整个过程包括两个阶段:机器学习模型的训练阶段和模型预测阶段.在机器学习的不同阶段,面临不同的隐私威胁,加之机器学习本身技术的原因,所采用的保护方法也不同,这是我们研究的重点.例如,目前同态加密技术多用于深度神经网络的预测阶段,而很少用于训练阶段.其原因是:由于深度学习本身是一项计算密集型的任务,计算以及通信开销大,即使没有加密,也需要高吞吐量的计算单元,而同态加密的计算和通信开销也很大.所以,同态加密目前一般用在神经网络预测阶段.研究训练时基于加密技术的高效机器学习隐私保护方法仍是一个公开问题.

(3) 按模型训练方式分类

机器学习模型训练方式可分为集中式学习、分布式学习和联邦学习 3 类.集中式学习中训练数据集由单机、集群或云端中央服务器统一收集、管理,其优点是训练、部署方便,模型训练准确率高;分布式学习中各参与方的训练数据无需集中到中央服务器,各参与方与中央服务器连接稳定,并且负载均衡,训练数据在各参与方的分布可能是水平分割的、垂直分割的或者是任意分割的.联邦学习也是多个客户端在中央服务器的协调下联合训练一个模型,同时保持训练数据分散.不过联邦学习面对的是一个更加复杂的学习环境,但联邦学习更加重视用户数据隐私的保护,因而目前倍受学界和产业界的关注.

(4) 按隐私保护技术分类

机器学习常见的隐私保护技术可以分成 3 类:基于差分隐私的隐私保护技术、基于同态加密的隐私保护技术和基于安全多方计算的隐私保护技术.其中差分隐私技术,属于数据失真的方法,它是通过生成人工合成数据或者在模型训练过程中给梯度参数、权重参数、目标函数或模型输出中添加噪声扰动,以保证模型或训练数据隐私.同态加密和安全多方计算技术,属于密码学方法,它们通过安全协议保护运算过程中数据隐私.上述方法往往组合起来使用,例如安全多方计算结合差分隐私、同态加密与安全多方计算组合使用.

3 基于差分隐私的机器学习隐私保护机制

差分隐私(differential privacy,简称 DP)是一种被广泛认可的严格的隐私保护技术.这一概念最早由微软的 Dwork^[94]提出.DP 技术使得恶意敌手即使知道用户发布的结果,也不能推断出用户的敏感信息.将 DP 应用于 ML 模型,可以在模型参数释放时保护训练数据不受模型逆向攻击.因此,有许多研究将 DP 应用到 ML 模型中.

3.1 相关概念

定义 4^[95]. (ϵ, δ) -差分隐私. 一个随机算法 $M: D \rightarrow R$ 满足 (ϵ, δ) -差分隐私,当且仅当对于任意相差仅一条数据的相邻数据集 $d, d' \in D$ 和任意输出 $S \subseteq R$,满足如下条件:

$$\Pr[M(d) \in S] \leq e^\epsilon \Pr[M(d') \in S] + \delta \quad (3)$$

其中, $M(d)$ 和 $M(d')$ 分别代表算法 M 在数据集 d, d' 上的输出; \Pr 为算法的输出概率; ϵ 为隐私预算,用于控制隐私保护级别. ϵ 越小,提供的隐私保护能力越强; δ 为另一个隐私预算,代表可容忍的隐私预算超出 ϵ 的概率.如果 δ 等于 0,我们就称 M 满足 ϵ -差分隐私.

为了更好地控制深度学习模型训练过程中的全局隐私损失, Abadi 等人^[76]引入了 Moments accountant 机制,用于对每次访问训练数据时所产生的隐私损失进行更精确的核算.其定义如下:

定义 5^[76]. MA(Moments accountant). 给定数据集 D , 设有一个随机算法 $M: D \rightarrow R$, 算法 M 满足 (ϵ, δ) -差分隐私, aux 为辅助输入,那么 λ 时刻的隐私损失定义为:

$$\alpha_M(\lambda) \triangleq \max_{aux, d, d'} \alpha_M(\lambda; aux, d, d') \quad (4)$$

λ 时刻隐私损失可认为是时刻生成函数的最大值,即遍历所有可能的 aux 、 d 、 d' 后取最大值.其中,时刻生成函数 $\alpha_M(\lambda; aux, d, d') \triangleq \log \mathbb{E}[\exp(\lambda c(o; M, aux, d, d'))]$;如式(5)所示的随机变量 c 表示输出空间 o 点的隐私损失.

$$c(o; M, aux, d, d') \triangleq \log \frac{\Pr[M(aux, d) = o]}{\Pr[M(aux, d') = o]} \quad (5)$$

目前 MA 机制已实现且在 TensorFlow 隐私库^[96]开源,因此被广泛应用于差分隐私深度学习中.

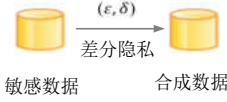
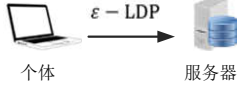
差分隐私是一种鲁棒模型,具有以下性质:

性质 1^[97]. 后处理免疫性. 对于同一数据集 D ,如果机制 M 满足 ϵ -差分隐私,那么对于任意随机算法 A (不一定满足差分隐私定义),新的机制 $M'=A(M(D))$ 仍然满足 ϵ -差分隐私.

性质 2^[98]. 序列组合性. 如果一系列算法 M_1, M_2, \dots, M_k ,均满足 (ϵ, δ) -差分隐私,那么对于同一数据集 D ,由这些算法构成的组合算法 $\varphi(M_1(D), M_2(D), \dots, M_k(D))$ 提供 $(k\epsilon, k\delta)$ -差分隐私保护.

考虑到隐私性、可用性要求以及不同的敌手威胁场景,可以选择在机器学习的不同阶段,部署不同差分隐私扰动.例如,在模型训练前可以选择部署输入扰动,在模型训练中,可以选择部署中间参数和目标函数扰动,在模型训练结束时,可以选择部署输出参数扰动.具体如表 4 所示.

Table 4 Differential privacy perturbation methods in machine learning
表 4 机器学习中差分隐私扰动方法

扰动方式		扰动时机	核心公式或模型	参数说明
输入扰动	生成合成数据	训练前	 <p>敏感数据 $\xrightarrow{(\epsilon, \delta)}$ 合成数据 差分隐私</p>	ϵ 为隐私预算; δ 代表可容忍 ϵ 超出隐私预算的概率
	本地化差分隐私扰动		 <p>个体 $\xrightarrow{\epsilon\text{-LDP}}$ 服务器</p>	ϵ -LDP 为本地化差分隐私
中间参数扰动	扰动梯度参数	训练中	$\theta_{t+1} \leftarrow \theta_t - \eta(\nabla \ell(\theta_t) + \beta)$	ℓ 为损失函数; β 为高斯噪声; ∇ 是标准梯度算子; η 为学习率
	扰动特征参数		$\bar{x}_{ij} \triangleq x_{ij} + \frac{1}{ L } \text{Lap}(\Delta h_0 / \epsilon_j)$ $\epsilon_j = \beta_j \times \epsilon$	x_{ij} 样本 i 第 j 个特征; ϵ 为噪声总量; β_j 为特征参数对输出的贡献
目标扰动	扰动目标函数	训练中	$\ell_{priv}(\theta) = \ell(\theta) + \beta$ $\theta^* = \arg \min_{\theta} \ell_{priv}(\theta)$	β 为 Laplace 噪声; $\ell_{priv}(\theta)$ 为目标函数; θ^* 为扰动后目标函数最优参数
	扰动目标函数展开式系数		$\ell(\theta) = \sum_{j=0}^J \sum_{\phi \in \Phi_j} \sum_{x_i \in D} \lambda_{\phi x_i} \phi(\theta)$ $\bar{\lambda}_{\phi} = \sum_{x_i \in D} \lambda_{\phi x_i} + \text{Lap}(\Delta / \epsilon)$ $\theta^* = \arg \min_{\theta} \hat{\ell}(\theta)$	D 为训练数据集; x_i 为单个样本; J 为展开式阶数; $\phi(\theta)$ 为权重参数; $\lambda_{\phi x_i}$ 为权重参数系数; θ^* 为扰动后目标函数最优参数
输出扰动	扰动输出参数	训练结束时	$\theta^* = \arg \min_{\theta} \ell(\theta)$ $\theta_{priv} = \theta^* + \beta$	θ^* 为最优参数; β 为 Laplace 或指数噪声; θ_{priv} 为扰动后的参数
	扰动集成输出结果	预测输出时	$n_j(x) = \{i : i \in [t], f_i(x) = j\} $ $f(x) = \arg \max_j \{n_j(x) + \text{Lap}(1 / \epsilon)\}$	x 为待标记样本; t 为教师数; $n_j(x)$ 为 j 类得票数; $f(x)$ 为含噪输出标签

3.2 典型方案分析

(1) 基于输入扰动的隐私保护方案

输入扰动(Input Perturbation)是指为避免模型接触到用户真实数据,在模型训练前,先对训练数据进行一定程度的随机扰动.这种在模型训练前即对数据前进行保护的方法,大大减少了敏感信息的泄漏,从隐私性角度来讲比其他阶段的扰动更加可靠.现有文献中常采用差分隐私数据合成和本地化差分隐私扰动两种方法.

差分隐私数据合成,可以看作是训练数据的预处理过程.这种方法生成具有与原始输入数据相似统计特性和相同格式的

人工合成数据,从而达到保护原始数据隐私的目的.本地化差分隐私下的保护模型关注的是个人与不可信服务器之间通信的隐私.在该模型中,每个用户首先在本地对原始数据进行差分隐私扰动,再将处理后的数据发送给数据收集者^[99].

近年来,生成对抗网络(GAN)及其变体作为生成模型很好地解决了数据稀缺的问题,但由于GANs可能泄露训练数据隐私.为解决这一问题,Beaulieu-Jones等人^[49]提出了一种利用DP-SGD训练AC-GANs (auxiliary classifier generative adversarial networks)的模型DP-GANs,利用深度神经网络在DP下生成合成数据,为共享临床研究数据并保持患者隐私提供了解决方案.如图4所示,该模型使用两个神经网络:一个称为生成器(Generator)的神经网络 G 被训练从一组随机数 z 中生成与原始数据 x 足够相似的新数据;另一个称为判别器(Discriminator)的神经网络 D 用于判断一个样本是真实的还是生成器生成的样本.该模型价值函数如式(6)所示,通过构造一个两方player的minmax game,经过对抗训练,最终达到纳什均衡 (Nash equilibrium).在模型学习训练过程中,通过向判别器梯度中添加 (ϵ, δ) -差分隐私保护,根据差分隐私的后处理免疫性^[97],从而生成器也获得 (ϵ, δ) -差分隐私保护.此外,在该DPGAN框架中,判别器是唯一能访问真实、私有数据的组件,因此,敌手即使获得生成器本身,也无法获取训练数据的隐私.

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)} [\log(D(x))] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (6)$$

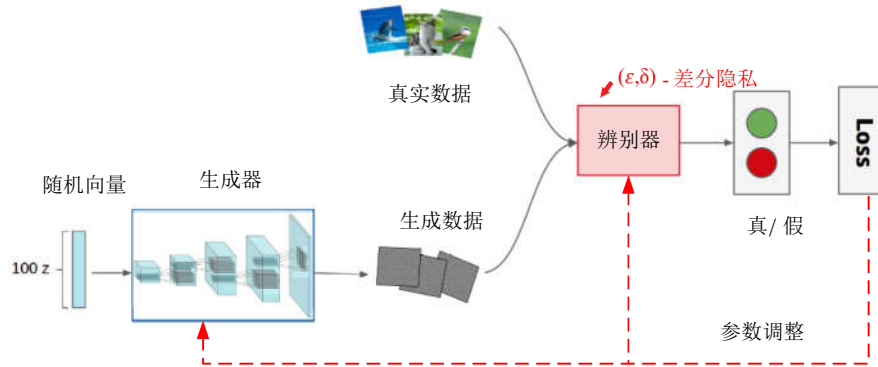


Fig.4 Framework for DP-GANs

图 4 DP-GANs模型框架

针对文献[49]中使用GAN存在训练不稳定、梯度消失和缺乏多样性等问题,Xie等人^[74]提出了一种差分隐私生成对抗网络(differentially private GAN,简称DPGAN)模型.该模型基于Wasserstein GAN(简称WGAN)网络,构造了另一个两方player的minmax game.相比GAN网络中的KL散度、JS散度的优势在于,即便两个分布没有重叠,Wasserstein距离仍然能够反映它们的远近,有效克服了GAN模型的训练不稳定、梯度消失等问题.该模型训练过程中使用 (ϵ, δ) -差分隐私保护训练数据的隐私,使用Moments Accountant机制精确控制模型训练过程中的隐私损失,确保了模型的可用性.

为了解决生成数据效用下降问题,Bindschaedler等人^[75]利用可信否认(plausible deniability)标准^[100]来度量生成数据隐私性,为高维敏感数据的发布提出了一种正式的隐私保障方法.满足plausible deniability标准的机制包括两个独立模块:生成模块和隐私测试模块.该方法先生成数据,然后只发布满足隐私要求的输出子集,因此能够生成高实用性的合成数据.通过隐私测试的思想来实现差分隐私,从而拒绝坏的样本,实现可信否认性. (k, γ) -plausible deniability机制定义如式(7)所示,这种机制导致输入的不可分辨性,意味着通过观察输出集(即生成数据),敌手无法确定某个特定的数据记录是否在输入集中(即真正的数据).隐私参数 k 越大,不可分辨性的输入数据集就越大;隐私参数 γ 越接近1,输入数据记录的不可分辨性越强.

$$\gamma^{-1} \leq \frac{\Pr\{y = M(d_i)\}}{\Pr\{y = M(d_j)\}} \leq \gamma \quad (7)$$

其中, $\forall i, j \in \{1, 2, \dots, k\}$; d 为原始输入数据; $M(d)$ 为概率生成模型; y 为生成的数据; k, γ 为隐私参数.

(2) 基于中间参数扰动的隐私保护方案

这种方案是在模型训练过程中给梯度参数或特征参数添加拉普拉斯噪声或高斯噪声,以防止敌手获取模型或训练数据

隐私.在最近的研究中,学者们提出了一些创新性的改进措施,如更精确地添加噪声和更严格地测量隐私损失,这对模型优化具有非常重要的意义.

针对深度学习中直接共享训练数据集将可能导致用户隐私泄露的问题,Shokri 和 Shmatikov^[36]提出了一种分布式选择性随机梯度下降算法(distributed selective SGD,简称 DSSGD).多方在不共享真实训练数据的情况下,通过并行异步训练过程,共同学习精确的目标模型.DSSGD 算法框架如图 5 所示,其服务器参数更新规则如式 8 所示,其中, α 为学习率; W_{global} 为中央服务器的全局参数,并被广播给所有参与者供其下载更新;向量 G 包含各参与者大约 1%-10%的梯度参数.为了确保参数更新不会泄漏关于训练数据集的太多信息,算法将 ϵ -差分隐私噪声(Laplace noises)添加到梯度参数中.各参与者之间不必交互,它们在本地图使用各自训练好的模型.实验表明,对于许多参与者,当参与者共享很大一部分梯度时,联合训练模型的准确性优于独立训练模型的准确性.

$$W_{global} \leftarrow W_{global} - \alpha G_{local}^{selective} \tag{8}$$

Liu 等人^[77]在文献[36]的基础上提出了一种移动环境下不共享局部原始数据的协同隐私保护深度学习系统,仅通过共享部分参数,就可以实现多个站点学习深度学习模型.移动设备在本地数据上进行训练,并通过循环和异步参数交换协议将训练后的参数上传到 XMPP (global server).

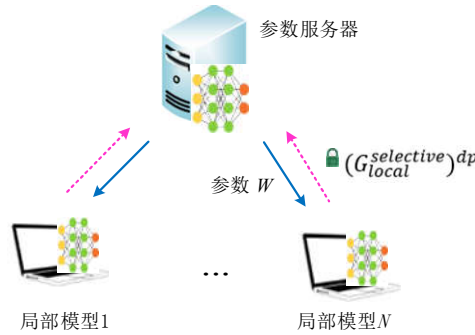


Fig.5 Framework for DSSGD

图 5 DSSGD算法框架

文献[36]中注入噪声的大小和隐私预算按训练周期数和共享参数数量的比例累积.因此,它可能会消耗不必要的大量隐私预算,因为训练迭代次数和多方共享参数的数量通常很大.为了改进这一点,跟踪训练过程中隐私损失,Abadi 等人^[76]基于组合定理(composition theorem)提出了一种 MA(Moments Accountant)机制.该机制允许对隐私损失进行自动跟踪分析,可以得到对整体隐私损失的更严格的估计,其性能目前已经优于高级组合定理(advanced composition theorems).其具体思路是:基于一种差分隐私随机梯度下降算法(Differentially Private SGD),在每个训练步骤中将噪声引入到“梯度”参数中,并利用 MA 机制对训练过程中总体隐私损失支出进行精细自动化的跟踪分析,以帮助每个参与者控制特别敏感的梯度参数,从而确保参数共享不会泄漏太多隐私.在隐私成本可控情况下,可对多达数百万个参数的深层模型进行训练,可应对强大的对手,允许对手控制部分甚至全部其余训练数据.在 MNIST 的实验中,实现了 97%的训练准确度.

然而,文献[76]中的方法仍然依赖于训练周期的数量.当只有很少隐私预算时,将只有少量迭代次数可用于模型训练.当需要大量训练迭代次数来保证模型精度时,这可能会潜在地影响模型效用.此外,现有技术另一个缺点是,所有参数注入的噪声量都是相同的,这在实际场景中可能并不理想,因为不同特征和参数通常对模型输出有着不同的影响.因此,Phan 等人^[50]基于逐层相关传播(layer-wise relevance propagation,简称 LRP)^[101]算法提出了一种自适应拉普拉斯机制(adaptive laplace mechanism,简称 AdLM),以实现深度神经网络的差分隐私保护. LRP 算法框架如图 6 所示.

AdLM 实现思路是:首先,根据 LRP 算法原理、仿射变换(affine transformation)及反向传播理论,来评估每个输入特征 x_{ij} 与模型输出 $\mathcal{F}_m(\theta)$ 之间的相关性,如式(9)所示;然后,基于预训练好的神经网络计算数据集 D 上每个特征的平均相关性 \bar{R}_i ,并添加拉普拉斯噪声,如式(10)所示;最后,根据每个特征 x_{ij} 对输出贡献不同自适应地向特征中注入噪声,在与模型输出关系不大的特征中注入更多的拉普拉斯噪声,如式(11)所示.该机制中,每个训练步骤中注入的噪声和隐私预算消耗不会积累,因此隐私预算消耗完全独立于训练迭代的次数.

$$\mathcal{F}_{x_i}(\theta) = \sum_{m \in h_k} R_m^{(k)}(x_i) = \dots = \sum_{x_{ij} \in x_i} R_{x_{ij}}(x_i) \tag{9}$$

$$\bar{R}_j \triangleq \frac{1}{|D|} \sum_{x_i \in D} R_{x_j}(x_i) + Lap(\Delta R / \varepsilon_1) \tag{10}$$

$$\hat{x}_{ij} \triangleq x_{ij} + \frac{1}{|L|} Lap(\Delta h_0 / \varepsilon_j) \tag{11}$$

其中,噪声系数 $\varepsilon_j = \beta_j \times \varepsilon$, ε 为这一步注入的噪声总量, β_j 为神经元第 j 个特征对输出的贡献系数.

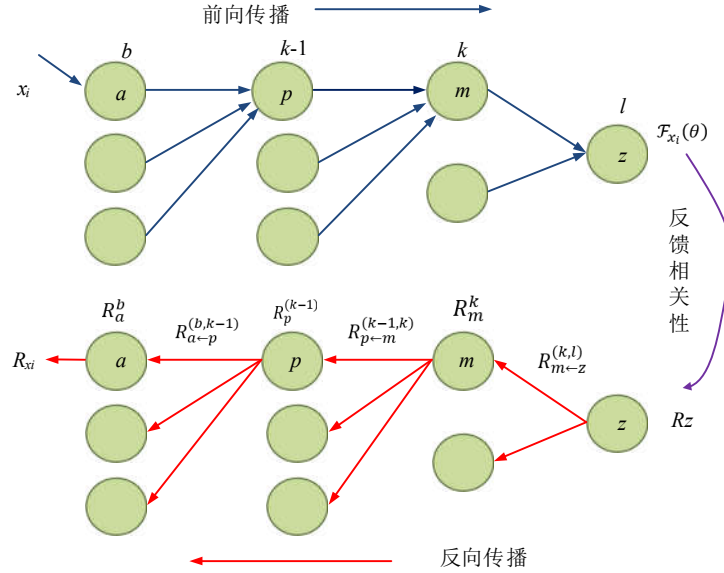


Fig.6 Framework for LRP^[101]

图 6 LRP算法框架^[101]

(3) 基于目标扰动的隐私保护方案

目标扰动(objective perturbation) 也称函数扰动,是指在机器学习模型的目标函数或目标函数展开式的系数中添加拉普拉斯噪声,并最小化此目标函数的方法.与参数扰动方法不同,目标扰动方法其隐私损失是由目标函数本身决定的,与训练迭代次数无关.已有研究^[102]表明,目标扰动方法在理论保证下优于输出扰动方法的有效性.不过目标扰动要求目标函数连续可微且为凸函数,因而直接扰动目标函数的方法具有一定的局限性,不适用于神经网络等非凸模型.

另一种扰动方法是在目标函数展开式的系数中添加拉普拉斯噪声.为了在系数中注入噪声,目标函数应该是权重的多项式表示.如果目标函数不是多项式形式,则目标函数应使用泰勒(Taylor)或切比雪夫展开式(Chebyshev expansion)等逼近技术将其近似为多项式表示,然后将噪声添加到各系数中.然而由于求解近似多项式方法仅针对特定的目标函数,故该方法难以拓展到更通用的模型.表 5 给出了基于函数扰动机制的差分隐私保护方案对比.

Table 5 Comparison of differential privacy schemes based on function mechanism

表 5 基于函数扰动机制的差分隐私保护方案比较

方案	底层模型	目标函数	逼近方法
dPAs ^[78]	Deep Auto-Encoders	交叉熵误差目标函数	泰勒级数展开
pCDBN ^[57]	Convolutional Deep Belief Network	能量函数	切比雪夫级数展开
Ref[51]	Deep Neural Network	交叉熵误差目标函数	麦克劳林级数展开

Chaudhuri等人^[44]首先基于函数敏感性(sensitivity-based)思想^[103]设计了一种隐私保护的逻辑回归算法.使用这种方法需要限定要学习的函数类的灵敏度,然后利用与灵敏度成正比的噪声干扰学习分类器.该方法中的 ε -差分隐私模型可以限制敌手获得关于特定数据的隐私信息,但对于某些机器学习函数来说,这可能很困难,因此,作者提出了另一种保护隐私的逻辑回归方法.该方法基于扰动目标函数(perturbed objective function),不依赖于函数的敏感性,并且该方法在模型中是私有的.实验证明,后一种方法具有更好的学习性能.

针对深度学习中可能存在的模型反演攻击,Phan 等人^[78]以深度学习的基础组件—自动编码器为研究对象,提出了一种深度私有自编码器(Deep Private Auto-Encoders,简称 dPAs)方案.该方案通过 ϵ -差分隐私来扰动深度自动编码器的交叉熵误差目标函数(cross-entropy error function),在数据重建过程中添加噪声干扰从而保护训练数据的隐私.当目标函数的多项式形式包含无限次项时,利用 Taylor 展开式进行近似.

现有 DP 算法在深度学习中的适用性问题引起了广泛关注.例如 dPAs 方案^[78]是为特定的深度学习模型所设计.为此,Phan 等人^[57]提出了一个私有卷积深度信念网络(private Convolutional Deep Belief Network,简称 pCDBN).而卷积深度信念网络是一种典型的基于能量的深度学习模型,其结构比 Auto-Encoders 更为复杂.pCDBN 本质上是一个基于差分隐私的 CDBN,它利用 Chebyshev expansion 将非线性目标函数近似为多项式,将噪声注入到多项式系数中.每个隐藏层在训练阶段都满足 ϵ -差分隐私.pCDBN 框架隐私预算独立于训练 epochs 数量,使其可应用于大型数据集,大大促进了隐私保护在深度学习中的应用.

现有的许多基于隐私保护的DNN模型,其准确性比非隐私保护模型要低得多,从而限制了隐私保护DNN模型在工业界的使用.针对这一现象,Adesuyi 等人^[51]提出了一种基于DP和逐层相关传播(LRP)的隐私保护深度神经网络训练方法.该方法通过麦克劳林级数(Maclaurin series)对交叉熵误差函数进行多项式逼近,利用差分隐私噪声扰动交叉熵误差目标函数系数,采用 LRP 算法确定噪声添加的位置.其分类精度接近于非隐私保护神经网络模型的精度.

(4) 基于输出扰动的隐私保护方案

输出扰动(output perturbation)是在模型训练结束时扰动模型输出参数以及在模型预测输出后扰动集成输出结果.前一种方法是直接在训练好的模型参数上添加噪声的扰动方法.由于直接在模型参数上添加扰动,可以有效防止模型提取攻击,从而为攻击者进一步利用模型逆向攻击窃取训练数据造成障碍.但这种方式仅仅实现了模型发布阶段的隐私保护,攻击者仍有可能在前期阶段通过多次请求,攻击训练数据隐私;后一种方法往往发生在师生框架的知识转移阶段,即在利用教师模型训练学生模型时,在教师模型的预测输出投票结果上添加拉普拉斯噪声.其目的是增强模型的泛化度,防止敌手对模型进行成员推断攻击和模型逆向攻击.

Jayaraman 等人^[73]提出了一种对分布式学习输出进行差分隐私扰动的方法.各方基于安全多方计算协议共同学习一个机器学习模型,然后在全局模型中添加 Laplace 噪音进行输出扰动,作者证明了在安全多方计算场景下,在聚合后的模型中加入噪声比其他扰动方案的噪声小,并且可以防止敌手对最终模型的推理攻击.KDDCup98 数据集上的实验表明,该方法能达到与非隐私方法相近的精度.

Papernot 等人^[60]基于半监督知识迁移的思想提出一种称为教师群体私有集成(private aggregation of teacher ensembles,简称 PATE)的模型,用于解决机器学习中训练数据隐私泄露问题.PATE 将敏感数据分割成 N 个不相交的数据子集,在每个数据子集上分别训练一个教师模型.对于待标记的公共数据,在教师模型集体投票结果上添加差分隐私噪声扰动,以得票数最多的类标签为预测结果.之后再教师标注的数据集训练学生模型,最终使用学生模型进行预测服务.这样能够防止模型逆向攻击对原始敏感数据的窃取.然而,由于 PATE 模型的隐私损失与公共数据集中带标记数据量成正比,可能导致无法承受的隐私损失,因此,PATE 只可应用于简单分类任务.

后来Papernot等人^[79]将PATE扩展到大规模环境,可用于图像分类任务.改进后的PATE在各性能指标上均优于原PATE,通过引入一种新噪声聚集机制RDP(Rényi Differential Privacy)^[104],需要比传统Differential Privacy更低的隐私成本,提供了更严格的差分隐私保证. PATE框架的关键约束是假定学生模型可以访问未标记的、非敏感的公共数据,其统计特性与训练教师模型的数据一样,但在医疗及其他应用领域找到这种数据不太现实. 表6给出了Differential Privacy与RDP的性质比较.

Table 6 Comparison of properties provided by Differential Privacy and RDP

表 6 Differential Privacy与RDP的性质比较

性质	Differential Privacy	RDP
输出概率变换	$\Pr[f(d) \in S] \leq e^\epsilon \Pr[f(d') \in S]$ $\Pr[f(d) \in S] \geq e^{-\epsilon} \Pr[f(d') \in S]$	$\Pr[f(d) \in S] \leq (e^\epsilon \Pr[f(d') \in S])^{(a-1)/a}$ $\Pr[f(d) \in S] \geq e^{-\epsilon} \Pr[f(d') \in S]^{a/(a-1)}$
贝叶斯因子变化	$\frac{R_{post}(d, d')}{R_{prior}(d, d')} \leq e^\epsilon$	$E \left\{ \left[\frac{R_{post}(d, d')}{R_{prior}(d, d')} \right]^{a-1} \right\} \leq \exp[(a-1)\epsilon]$
后处理	若 f 满足 ϵ -DP (或 (α, ϵ) -RDP), 则 $g \circ f$ 满足 ϵ -DP (或 (α, ϵ) -RDP)	
顺序组合	若 f, g 满足 ϵ -DP (或 (α, ϵ) -RDP), 则 (f, g) 满足 2ϵ -DP (或 $(\alpha, 2\epsilon)$ -RDP)	

3.3 综合分析

与加密技术相比,差分隐私仅通过随机化和利用随机噪声扰动数据便可以实现,所以在机器学习中部署差分隐私技术并不会带来过多额外的计算开销,与非隐私保护的传统算法相比,其运行时长差不多^[105],但一定程度上会影响模型的可用性,导致模型的预测准确性下降.最严格的差分隐私机制可以更好地保证机器学习模型不受成员推理攻击或模型逆向攻击.理论上可以实现攻击者已知数据集中除一条记录之外的全部数据时仍能提供隐私保护,但这种做法将导致模型不可用^[13].一种解决思路是适当降低隐私保护要求,让算法满足一种更为宽松的差分隐私约束,但这样将造成更大可能的泄露隐私^[106].

本地化差分隐私技术可以在一定程度上保证用户隐私数据在数据采集过程中被窃取的风险^[35, 99].在本地化模型中,每个用户对即将上传至服务器的数据或者中间结果进行扰动,可以避免服务器直接收集或接触到用户本地原始数据,同时又不影响对用户数据进行统计分析.

与传统机器学习模型相比,深度学习模型由于其目标函数是非凸函数,且参数多、结构更加复杂,因而需要更多次访问敏感训练数据集,更多次训练迭代才可能收敛至最优解,且常常是局部最优解.如果每次参数更新都要求满足差分隐私保证,则整个训练过程的全局隐私开销将很大,从而导致该技术面临难以合理地权衡隐私性与模型可用性的难题.

基于差分隐私保证的生成对抗网络生成的人工数据,缺乏严格的隐私保护,并且非常接近真实样本,在细节上差别很小,使得这种技术可能不能完全保护隐私,安全保护强度比加密机制弱.另外,由于新样本仍然保持了原有样本的特征,因此无法抵抗对抗统计特性的推理攻击.

4 基于同态加密的机器学习隐私保护机制

密码学和机器学习之间的联系已经被研究了很长时间,普遍认为它们是相互对立的.在某种意义上,密码学的目的是为了防止对信息的访问,而机器学习则试图从数据中提取信息^[107].在机器学习领域,为了实现用户数据机密性,一种方法是利用传统的密码学方法,但需要加密和解密阶段,这使得它在现实世界中不切实际,因为它的计算复杂性非常大.不过密码学的最新研究成果允许在加密数据上执行任意操作,而无需解密,全同态加密(full homomorphic encryption)即属于此类方法.下面先简要介绍同态加密技术,然后介绍基于同态加密的机器学习隐私保护研究进展.

4.1 相关概念

(1) 同态加密(Homomorphic Encryption, HE)是一种允许用户直接在密文上进行运算的加密形式,其得到的结果仍是密文,解密结果与对明文运算的结果一致.同态加密方案满足等式(12).

$$Dec(k_s, Enc(k_p, m_1) \diamond Enc(k_p, m_2)) = m_1 \circ m_2 \quad (12)$$

其中, m_1 、 m_2 为明文, k_s 、 k_p 分别为私钥与公钥, $Enc()$ 是加密运算, $Dec()$ 是解密运算, \circ 、 \diamond 分别为明文域和密文域上的运算.按照其发展阶段、支持密文运算的种类和次数, HE 分为部分同态加密、类同态加密和完全同态加密^[108].

(2) 部分同态加密(partially homomorphic encryption, 简称 PHE): 是最早设计的同态方案,只支持加法或乘法运算,且运算次数不受限制.可进一步分为加法同态加密方案(additive homomorphic encryption, 简称 AHE)如 Paillier 方案、乘法同态加密方案(multiplication homomorphic encryption, 简称 MHE)如 El-Gamal 方案等.

(3) 类同态加密(somewhat homomorphic encryption, 简称 SHE): 是一种只支持有限次加法和乘法运算的同态方案. SHE 比 FHE 方案稍弱,但也意味着开销更小,更容易实现.而层次型全同态加密方案(levelled full homomorphic encryption, levelled-FHE), 又称深度有界同态加密,也属于 SHE 方案^[109].所谓深度有界是指,只能处理有限数量的电路深度,因而, levelled-FHE 方案不适合训练深度神经网络. levelled-FHE 支持单指令多数据(single instruction multiple data, 简称 SIMD)批处理技术,因而 levelled-FHE 方案的性能较高.

(4) 完全同态加密(fully homomorphic encryption, 简称 FHE): Gentry 基于理想格(ideal lattices)理论提出的研究成果^[110], 它支持密文上任意算法,并且执行运算次数不限(unlimited number of times). FHE 方案安全可靠,然而,自举(bootstrapping)是一个非常昂贵的过程,计算开销太大,导致 FHE 依然不能成为一个实用的方案,也更无法直接应用在大数据环境中.近年来各种改进版 FHE 方案^[111-114]相续被提出,这些研究大都致力于噪声的减少和效率的提升.

表7从关键技术、是否支持深层模型、是否支持批处理、是否支持非线性运算和分类准确性等角度对基于同态加密的机器学习隐私保护方案进行分析对比.

Table 7 Comparison of machine learning privacy protection schemes based on cryptography

表 7 基于同态加密的机器学习隐私保护方案对比

分类	方案	关键技术	是否支持 深层模型	是否支持 批处理	是否支持非 线性运算	分类准确性	
						数据集	准确性(%)
无需多项式近似	Ref[52]	HE	否	否	是	N/A	N/A
	Ref [53]	HE	否	N/A	是	UCI Datasets	99.8
	Ref [9]	AHE	N/A	N/A	N/A	WBC	98.24
	Ref [47]	AHE	否	否	N/A	N/A	N/A
	Ref[48]	FHE	是	是	N/A	N/A	N/A
	TAPAS ^[29]	FHE,BNNs	是	是	是	MNIST	99.04
多项式近似	crypto-nets ^[30]	Leveled-FHE	否	是	否	N/A	N/A
	Ref [64]	FHE	是	N/A	否	STL-10	85.5
	Ref [66]	HE	是	是	否	MNIST	99.10
	PPDL ^[65]	AHE	否	是	N/A	MNIST	99
	CryptoNets ^[54]	Leveled-FHE	否	Y	否	MNIST	98.95
	Ref [55]	FHE	是	是	否	MNIST	99.30
	CryptoDL ^[56]	Leveled-FHE	否	Y	否	MNIST	99.52

4.2 典型方案分析

(1) 无需多项式近似的同态加密隐私保护方案

同态加密方案虽然安全可靠,但只支持加法和乘法等多项式运算,而不支持机器学习过程中使用的非线性运算,如神经网络中的sigmoid和ReLU等激活函数.解决方法之一是依靠数据持有者来完成非线性运算.例如,Barni等人^[52]提出了一种基于神经网络的数据隐私保护方法.数据持有者利用HE加密数据并将其发送到云平台.云平台计算数据与第一层权重之间的内积,并将结果发送给数据持有者.数据持有者解密,进行非线性转换,将转换结果加密后发送回云平台.云平台计算数据与第二层权重之间的内积并将输出发送回数据持有者.这个过程一直持续到所有层都计算完为止.数据所有者必须保持在线,并且共享中间结果.因此,机器学习过程中,神经网络大部分权重信息会泄露给数据持有者.为了克服这一弊端,Orlandi等人^[53]提出了另一种隐私保护方法.该方法仍然利用HE加密数据,确保提供给神经网络的数据是保密的.作者通过在数据所有者和模型所有者之间创建一个交互式的协议来解决非线性激活函数的问题.在该方案中,每个非线性转换都由数据所有者计算.模型以加密形式将输入发送到数据所有者进行非线性转换,数据所有者解密消息、应用转换、加密结果并将其发送回来.不幸的是,这种交互需要很大的延迟,并增加了数据所有者方面的复杂性,实际上使其不切实际.此外,它泄露了关于模型的信息.因此,Orlandi等人不得不引入安全机制,如随机执行顺序,来缓解这个问题.

为了防止分类过程中发生隐私泄露,Rahulathavan等人^[9]提出了一种利用Paillier加密系统将SVM决策函数转换为密文形式的方案.分类样本也进行加密处理.所有的计算都在密文上进行,只有持有私钥的测试人员才能解密获得分类结果.Prasad^[47]利用AHE算法以及朴素贝叶斯算法,研究了完全分布式通信环境下,连续数据和离散数据的朴素贝叶斯分类器隐私保护问题.Aslett等人^[48]利用贝叶斯分类器(Naive bayes classifier)、随机森林(Random forest)及其变体等对基于FHE加密的数据训练机器学习模型.其模型在某些任务上工作得很好,但在图像识别等领域效果不如神经网络.

为了在有限的内存和计算资源的设备(如移动电话)上训练和测试深度学习模型,近年来,二值神经网络(BNNs)^[68, 69]深受欢迎.它通过二值化 weights 和 activations(即取值+1或-1),使得原来 32bit 浮点数,只需要 1bit 表示,大幅度地降低了内存的占用. BNNs 结合同态加密技术,可用对密文数据进行高效和准确的预测.Chillotti 等人^[67]提出了自举全同态加密方案,通过引入 TFHE 库中的优化算法,将自举时间减少到 0.1 秒以下.该方案只支持对二值数据的操作,可以用来执行 BNNs 的所有操作.Bourse 等人^[70]提出 FHE-DiNN 模型,利用二值神经网络执行加密预测.此模型通过 MNIST 上的测试,预测的准确性一般.由于加密方案参数依赖于模型的结构,所以服务提供者若更新模型,那么用户将需要重新加密数据.Sanyal 等人^[29]提出了 TAPAS 系统,用于对 FHE 加密数据进行机器学习模型的预测.方案基于二值化和稀疏化技术,修改对神经网络的设计,实现了复杂模型上的加速和并行计算,并且允许服务提供者随时更新模型,在 MNIST 数据集上取得 99.04%的准确率.上述 FHE-DiNN 和 TAPAS 都利用了 BNNs 的概念,预测速度均快于基于 leveled-FHE 的批处理预测方法.

针对外包计算环境下,有些方案无法保证机器学习模型隐私^{[88] [91]}或中间运算结果隐私^[115], Li 等人^[80]提出了一种新的隐私保护卷积神经网络(Convolutional Neural Network,简称 CNN)预测方案.该方案利用同态加密、秘密共享和混淆电路技术,将预测数据和模型初始参数以秘密的形式存储在两个不共谋的服务器上,随后这两个服务器协同完成模型预测服务.由于两个服务器各自拥有一部分的秘密,在不共谋的情况下,无法得知用户数据以及模型参数的明文,因此保护了用户查询数据、模型、任何中间结果和最终预测结果隐私.对于 ReLU 等非线性激活函数,没有采用多项式近似的方法,而是使用基于混淆电路(GC-based)的方法,达到了与明文计算相同的精度.方案采用数据打包(data packing)、单指令多数据(SIMD)和异步计算等技术,减少了计算和通信开销,提高了计算速度.Liu 等人^[81]也提出了类似的方案,既保护了外包数据、中间查询以及结果的隐私,也保护了模型的隐私.

(2) 基于多项式近似的同态加密隐私保护方案

针对同态加密方案不支持机器学习中非线性运算问题,研究者们提出的另一种解决方案是利用多项式逼近技术.例如 Xie 等人^[30]提出了一种基于 leveled-FHE 技术的隐私保护神经网络模型 crypto-nets.作者研究了如何利用已训练好的神经网络直接在密文上做预测,并返回加密预测结果,如图7所示.实现这一解决方案的主要挑战是,神经网络中常用的激活函数(比如 sigmoid、ReLU 等)不是多项式形式的,但它们都是闭区间上连续的.因此,作者从理论的角度,利用 Stone-Weierstrass 定理^[116]得到 $\sup_{x \in X} \|N(x) - P(x)\| < \epsilon$, 其中 N 为一个神经网络, X 为 N 的非空连续实值空间, P 为多项式, ϵ 为大于 0 的任意实数.从而证明了可用多项式近似模拟神经网络.根据同态加密的性质, HE 方案满足 $P(m_1, \dots, m_n) = D(P'(E(m_1), \dots, E(m_n)))$, 其中, P 、 P' 为多项式函数, (E, D) 分别为加密函数和解密函数.这表明 HE 方案可以在不首先解密的情况下,对加密消息计算任意有界多项式函数 P .由此容易得到: $\sup_{x \in X} \|N(x) - D(N'(E(x)))\| < \epsilon$.这表明,现有神经网络可以应用于加密数据.这是通过两个阶段的过程完成的:首先,神经网络 N 被一个多项式 N' 近似;然后,这个多项式被加密,即利用同态加密函数 \oplus 、 \otimes 分别替代多项式 P 中的加法和乘法,利用常量的加密版来代替多项式 P 中的常量.

对于神经网络的学习过程,由于所有的非线性变换和损失函数都是多项式,这意味着梯度(权重的导数)也是多项式,因此它可以在密文数据上进行计算,即反向传播算法可以学习到在对应明文数据上学习的系数的加密版本.与 Orlandi 等人^[53]的方案相比, crypto-nets 模型由于不需要数据所有者参与任意中间计算,所有计算由模型完成,因此 crypto-nets 模型没有复杂的通信过程,从而允许异步通信,并且不会泄漏关于模型的信息.

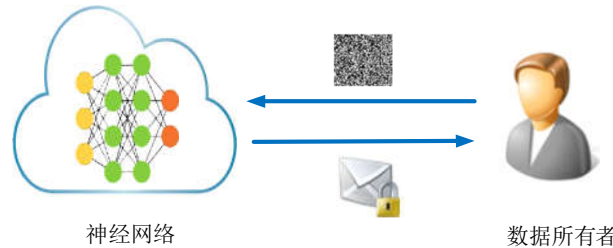


Fig.7 Privacy preserving neural network prediction on encrypted data^[30]

图 7 神经网络对加密数据进行安全预测^[30]

Zhang 等人^[64]提出利用 BGV 全同态加密方案,在密文上直接训练深度计算模型.作者利用 Taylor 公式对激活函数等非多项式函数进行模拟,支持高阶反向传播算法的高效安全计算.为了避免乘法深度过大,每次迭代后更新的权值被发送给各方进行解密和再加密.这样导致通信复杂度非常高.Hesamifard 等人^[66]在训练神经网络时,利用 Chebyshev 多项式近似模拟激活函数.当用多项式近似代替 ReLU 激活函数时,预测精度为 99.10%,近似代替 Sigmoid 激活函数时,预测精度为 99.00%.

针对 Shokri 和 Shmatikov 方案^[36]中可能存在的缺陷:如果参数服务器是好奇的,通过模型逆向攻击,即使只上传一小部分梯度信息也可能间接泄露用户数据隐私, Le Trieu Phong 等人^[65]提出了一种基于同态加密的隐私保护深度学习系统 PDDL.在模型训练过程中,各参与者利用加法同态将梯度参数加密后发送给中央服务器,防止了潜在的隐私泄露给不可信中央服务器. PDDL 网络框架如图 8 所示.基于 AHE 加法同态性,各处理单元按式(13)进行权重参数更新.对于每个参与者,下载并用密钥 sk 解密后得到权重参数 $w^{(j)}$ (其中 $j \in [1, n_{pu}]$),进而可得权重向量 W_{global} .利用 n_{pu} 个处理单元同时对梯度进行并行更新计算,显著加快了深度神经网络的训练.不过,采用同态加密虽然提高了数据和模型的隐私,但也付出了更多的通信成本.

$$W_{global}^{(i)} \leftarrow W_{global}^{(i)} - a \cdot G^{(i)} \quad (13)$$

微软研究院的 Gilad-Bachrach 等人^[54]基于 leveled-FHE 技术(YASHE^[117])提出了一种近似神经网络模型 CryptoNets.作者假设在云端已经有应用明文训练好的神经网络模型,使用低次多项式近似非线性激活函数(平方激活函数),使目标模型用于密文预测,将加密预测结果返回给用户.虽然是近似模拟神经网络模型,但 MNIST 分类性能达到了 98.95%的准确率.中间结果不共享,云端泄露给数据持有者的信息更少.多项式代替了平方激活函数,预测模型得到的结果与训练模型得到的结果有较大的差异.由于采用了 leveled-FHE 技术,增加乘法深度,将大大增加计算复杂度,当非线性层的数目很小如 2 时,效率和准确性得到了证实,但对于较深的神经网络,模型变得无效.由于使用了 SIMD 批处理技术,CryptoNets 支持高吞吐量计算,但对单个图像进行分类时,这个特性没有优势.客户端需根据模型结构生成加密参数,因此可对模型进行推断,这也导致模型隐私被泄露.

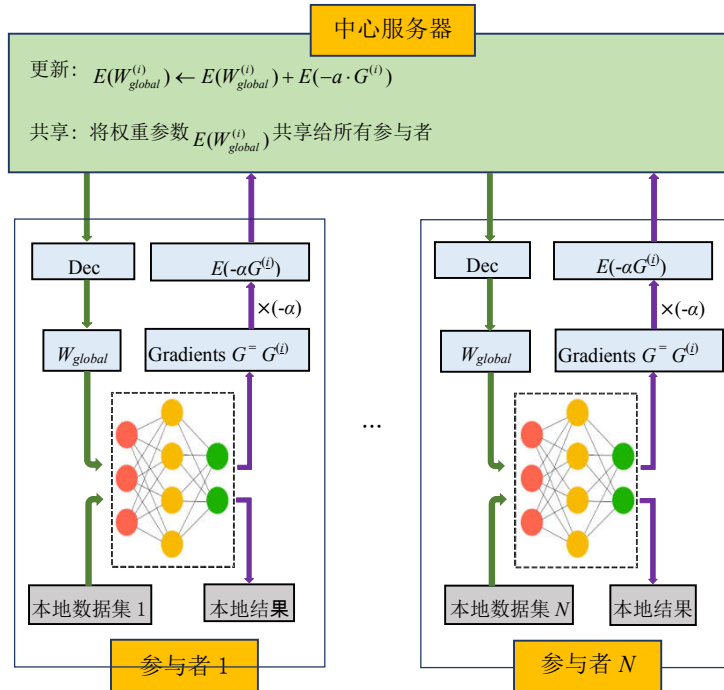


Fig.8 Framework for PPDL^[65]
图 8 PPDL网络框架^[65]

Chabanne 等人^[55]基于全同态加密技术提出了一种将 ReLU 激活函数的多项式逼近与批处理规范化相结合的深度神经网络分类方案.与 Cryptonets^[54]相比,该方法可应用于更深的神经网络,同时保持较高的精度.ReLU 的多项式逼近与批处理规范化相结合,减少了实际训练模型与转换模型之间的精度差距.不过,存在与 Cryptonets^[54]同样的问题,客户端需要根据模型的结构生成加密参数,泄露了模型隐私.

Hesamifard等人^[56]提出了对密文数据进行分类的深度学习CryptoDL.利用已用明文训练好的模型对Leveled-FHE加密的数据进行分类,采用低阶多项式逼近CNNs中常用的激活函数.由于采用SIMD批处理技术,提高了数据分类效率.

4.3 综合分析

同态加密是真正的端到端加密系统,有望从根本上解决当今数据模型的信任问题,它使用户能更好地控制其数据,同时受益于远程服务器提供的计算服务.例如在集中式机器学习中,用户将训练数据以密文形式上传至服务器,服务器进行模型训练但并不知道用户原始训练集因而保护了用户数据隐私;在联邦学习中,各个参与方将模型参数或者梯度加密后上传至中央服务器,中央服务器在不知道每个参与方上传的原始模型参数或者梯度的同时完成了模型训练迭代,从而保护了模型和用户原始数据的隐私.

由于任何计算都可以写成二元多项式,这意味着利用全同态加密方案执行密文计算时多项式的次数没有限制.机器学习模型训练过程中涉及的数据和参数通常是浮点数的形式,而同态加密技术只支持整数运算,因此,全同态加密不支持机器学习中激活函数等非线性运算,必须利用多项式来近似.然而正是这种近似,造成了精度和效率上的下降.同态加密技术计算和通信开销很大,对当前的计算资源和通信设施是一大挑战.

leveled-FHE 方法支持 SIMD 批处理技术,因而 leveled-FHE 方案的性能较高.leveled-FHE 方案经过仔细的参数调优后,

加密方案安全可靠,只需使用尽可能少的操作就能对其进行近似,并且许多计算任务只需计算低阶多项式.因此,加密数据常常采用 leveled-FHE 或 SHE 方案^[118].但训练过程中计算 sigmoid 或 softmax 等非线性激活函数代价较大,导致 leveled-FHE 无法执行许多嵌套乘法,因此,密文数据集上的深度神经网络的训练依然是一个公开的难题.

5 基于安全多方计算的机器学习隐私保护机制

安全多方计算(secure multiparty computation,简称 SMC)起源于姚期智^[119]的百万富翁问题,主要用于解决一组互不信任的参与方之间保持隐私的协同计算问题.下面先介绍几个相关概念.

5.1 相关概念

安全多方计算形式化描述为:假定有 m 个参与方 P_1, P_2, \dots, P_m ,他们拥有各自的数据集 d_1, d_2, \dots, d_m ,在无可信第三方的情况下,如何安全地计算一个约定函数 $y=(d_1, d_2, \dots, d_m)$,同时要求每个参与方除了计算结果外不能得到其他参与方任何输入信息. SMC 具有输入独立性、计算正确性、去中心化等特征.

SMC 基础密码协议包括 OT 协议(oblivious transfer protocol,简称 OT)、GC 协议(garbled circuits,简称 GC)、SS 协议(secret sharing,简称 SS)、GMW 协议(Goldreich-Micali-Wigderson,简称 GMW)等.这些协议都是重要的密码学工具,可以看作特殊的安全多方计算问题. SMC 是多种密码学基础工具的综合应用,因此在实现安全多方计算时也广泛地应用了同态加密技术.

1) OT 协议

OT协议又称不经意传输协议、遗忘传输协议或茫然传输协议,最早由Rabin^[120]于1981年提出. OT协议是一个两方计算协议,其中一方是发送方,另一方是接收方.接收方获得了部分信息,但发送方不知道他收到了哪些消息.在恶意敌手模型下, SMC 所需执行的OT次数需数百万次.例如,在计算隐私集合求交电路时,需要 2^{30} 次OT计算,这导致OT通常成为两方计算的瓶颈^[121]. 为了提高效率,应使用尽可能少的OT调用,或使用OT扩展技术,仅使用少量的基本OT协议来实现大量OT实例^[122].

2) GC 协议

GC协议又称混乱电路协议,是Yao^[119, 123]在1982年提出的一种通用高效的安全两方计算协议.2009年,Lindell等人^[124]给出了安全性证明.2012年,Bellare等人^[125]给出了GC的标准化定义.GC协议只需常数轮交互(不管电路大小),是最有效的安全两方计算解决方案之一,但总的通信量很高^[90].

3) SS协议

SS协议又称秘密共享、秘密分割协议,最早由Shamir和Blakley在1979年分别基于Lagrange插值多项式和线性几何投影理论独立提出来的.SS协议有Shamir秘密共享协议、Blakley秘密共享协议和中国剩余定理等.Shamir提出的 (t,n) 门限秘密共享协议 $(t < n)$ ^[126],是将秘密信息 K 拆分为 n 个份额 $\{p_1, \dots, p_n\}$,每个份额叫做 K 的“影子”或共享,利用任意 t 个 $(2 \leq t \leq n)$ 或更多个共享份额才可以恢复秘密信息 K .

4) GMW协议

GMW协议是Goldreich等人^[127]在1987年提出的一种通用高效的安全多方计算协议.与GC类似,它需要将函数描述为一个布尔电路.与GC不同,GMW评估电路的每一层布尔门都需要一轮交互.与GC相比,GMW需要更少的数据通信.如果只考虑在线成本,GMW中的大部分计算和通信可以转移到预处理阶段,在线阶段将非常高效.

通过大量的文献调研,我们发现目前机器学习隐私保护领域主要有两类方案与多方相关.第一类是基于传统分布式学习的方案.在这类方案中,各方能够参与 ML 模型的训练或测试,而无需披露其数据或模型;另一类是基于 HE、OT 或 GC 等技术的 2PC 架构的方案.该方案主要包含有两个参与方:一方是数据提供方,另一方是基于提供的数据实现机器学习的服务器.

5.2 典型方案分析

(1) 基于传统分布式学习的 SMC 方案

这种方案其实质是一种加密的分布式机器学习技术,参与各方在不披露自己数据隐私的情况下,通过交换必要信息,进而整个数据集上联合构建统一的机器学习模型.

Vaidya 等人^[61]针对任意划分的数据,提出一种基于安全多方计算的 k -means 聚类算法.各方在不向对方披露各自数据的情况下,交换必要信息,在整个数据上协同执行 k -means 计算.Bansal 等人^[72]针对任意分割训练数据集,提出了一种基于 HE 的神经网络学习算法.除了双方都知道最终训练权重外,没有泄露任何数据隐私,包含中间运算结果.Samet 等人^[46]针对水平分割或垂直分割训练数据,实现了一个极限学习机(extreme learning machine).由于数据持有者直接参与那些不受部分同态支持的

操作,因此可能会导致关于学习模型的敏感信息泄漏.Mehnaz 等人^[71]提出了一个基于安全和计算的通用框架,使多方能以隐私保护的方式对分割数据进行模型训练.作者设计了两种安全梯度下降算法,一种用于水平分割数据,另一种用于垂直分割数据.这个框架能够抵抗共谋攻击,适用于大型数据集多方计算情形,也适用于各种机器学习算法.

目前机器学习中提高 SMC 算法计算效率是大家的主要关注点.Li 等人^[45]基于改进的 C4.5 决策树提出了一种外包计算解决方案.为了减少计算开销,作者运用 OPPWAP 和 OSSIP 协议实现通用 SMC 计算,借助密码算法把计算任务外包给服务器端.将水平分割数据上的分布式 C4.5 决策树规约到权值平均问题,将垂直分割数据上的分布式 C4.5 决策树规约到安全交集问题,从而把用户端计算复杂度降低到亚线性级别.Abbasi 等人^[83]提出了一种安全聚类多方计算 (Secure Clustered Multi-Party Computation, SCMC) 方法.SCMC 允许类中存在一定的隐私泄露,实现了效率与隐私保护之间的平衡.Asharov 等人^[85]在不同的 SMC 模型中使用扩展的 OT 协议,降低了通信和计算复杂度.实验表明,改进的 OT 算法确实提高了 SMC 系统的效率.Gheid 等人^[62]针对大数据集直接运行 k -means 聚类算法会导致隐私泄露问题,提出了一种改进的安全多方求和协议.该算法操作简单,解决了密码学解决方案导致的性能下降问题.Dani 等人^[82]利用 quorum 概念设计了同步、异步 SMC 协议,解决了 SMC 系统的通信开销和计算开销随着参与人数的增加而线性增长、难以在大规模分布式系统中实现的问题.在保证安全的同时,将 SMC 的通信和计算从线性复杂度降低到亚线性复杂度.Bogdanov 等人^[84]利用 Sharemind 模型的优点实现了大数据集的安全计算,解决了一般 SMC 模型无法处理大数据集的问题.然而,Sharemind 只支持三方计算,不支持更多方参与者的安全计算.

(2) 基于 2PC 架构的 SMC 方案

基于 2PC 架构的 SMC 方案是另一种典型的多方计算隐私保护方案.这些机器学习隐私保护方案由若干个安全多方计算基础密码协议组合构建的.其中经典的两方计算方案有: HE + GC^[43]、HE + GC+SS+OT^[86]、GC + OT^[87]、HE + GC+SS^[88] 和 GC+SS+OT^[89] 等.一方为提供数据的用户,另一方为对数据进行计算的服务器.

表 8 从关键技术、是否支持非线性运算、是否支持批处理、运行耗时、通信量和准确性等角度对基于 2PC 架构的 SMC 隐私保护方案进行分析对比.

Table 8 Comparison of SMC privacy-preserving schemes based on 2PC

表 8 基于 2PC 架构的 SMC 隐私保护方案对比

框架	关键技术	是否支持非线性运算	是否支持批处理	数据集	运行耗时(s)			通信量(MB)			准确性 (%)
					离线	在线	合计	离线	在线	合计	
SecureML ^[86]	Linearly HE, GC,SS,OT	否	否	MNIST	4.7	0.18	4.88	N/A	N/A	N/A	93.4
DeepSecure ^[87]	GC,OT	是	否	MNIST	N/A	9.67	9.67	N/A	791	791	98.95
MiniONN ^[88]	AHE,GC,SS	是	否	MNIST	3.58	5.74	9.32	20.9	636.6	657.5	99
EzPC ^[89]	GC,SS,OT	是	N/A	MNIST	N/A	N/A	5.1	N/A	N/A	501	99.2
Chameleon ^[90]	GC,GMW,A-S S	是	否	MNIST	1.25	0.99	2.24	5.4	5.1	10.5	99
GAZELLE ^[91]	AHE,GC	是	是	MNIST	0.48	0.33	0.81	47.5	22.5	70.0	N/A
LPP-CNN ^[58]	A-SS	是	否	MNIST	0.09	0.21	0.3	1.57	0.99	2.56	99.14
OPSR ^[59]	A-SS	是	否	TIMIT	N/A	N/A	0.39518	N/A	N/A	2.435	N/A

Nikolaenko 等人^[43]提出了一种基于 leveled-FHE 和 GC 的水平分割数据隐私保护线性回归算法.数百万个样本集实验表明,其性能明显优于仅基于 leveled-FHE 或 GC 的隐私保护方案,且可根据用户数量和特征进行扩展,同时确保结果的准确性.

Mohassel 等人^[86]基于 SMC、SS 和乘法三元组(multiplication triplets)等设计了一种双服务器机器学习模型 SecureML.将 leveled-FHE 加密的数据发送到两个互不合谋的服务器,使用安全两方计算训练神经网络等各种模型.该方案重点支持模型训练,也支持隐私保护预测.在训练阶段,非线性激活函数用多项式近似,并通过预计算减少在线预测阶段计算成本.该模型比文献^[43]中的协议快 1100-1300 倍,且可以扩展到数百万个大数据样本,但预测输出泄露了一些模型信息.

Chandran 等人^[89]提出了一个安全两方计算框架 EzPC,实现了从高水平、易于编写的程序生成高效两方计算协议.EzPC 框架将算术共享和混乱电路结合起来,服务器无法获得客户端的输入和输出信息.除了服务器端的输出外,客户端也无法获得服务器端的模型信息.EzPC 框架所生成的协议比目前具有安全预测和矩阵分解等功能的协议快 19 倍.Henecka 等人^[92]提出了另一种自动化工具 TASTY.它基于 HE 和 GC 技术,为 PSI 和隐私保护人脸识别 (privacy-preserving face recognition) 等特定应用问题自动生成有效的安全两方计算协议.TASTY 的自动化体现在集描述、生成、执行、基准测试和比较于一体.

针对CryptoNets^[54]、SecureML^[86]框架在训练阶段用多项式近似非线性激活函数,从而改变了神经网络训练方式,导致模型精度下降的问题,Rouhani等人^[87]提出了遗忘神经网络预测的框架DeepSecure.该框架对加密数据进行遗忘预测,是第一个可扩展具有可证明安全的深度学习框架.与SecureML^[86]相比,DeepSecure消除了双服务器合谋攻击.由于基于GC技术,该框架支持任何非线性激活函数,无需改变神经网络训练方式,保证了模型的精度.为了减少GC协议的开销,此框架引入了预处理步骤.Liu等人^[88]提出了另一个基于遗忘神经网络(oblivious neural networks,简称ONN)的两方计算框架MiniONN.在离线预计算阶段引入了HE方法,在在线预测阶段使用秘密共享等轻量级密码原语,确保了模型和数据隐私.该框架使用真实的sigmoid激活函数进行训练,没有改变神经网络训练方式.

M.SadeghRiazi等人^[90]也提出了一个减少GC协议开销的混合安全计算框架Chameleon.该框架利用加法秘密共享(Additive Secret Sharing,简称A-SS)协议执行线性操作,利用GMW或GC协议执行非线性操作.与SecureML^[86]框架类似,Chameleon需要一个额外的非合谋方,即半诚实第三方(STP).基于STP生成的相关随机性(correlated randomness)^[128],Chameleon将几乎所有繁重的密码操作在离线阶段完成,显著降低了计算和通信开销,提高了分类效率.

Juvekar等人^[91]基于AHE与GC提出了一种新的安全神经网络推理方案GAZELLE.客户端在不向服务器公开其输入的情况下获取加密分类结果,同时保证了神经网络的隐私性.作者利用AHE执行线性运算,利用GC执行非线性运算.采用SIMD操作,避免了密文-密文乘法,降低了噪声增长.与纯同态方案Cryptonets^[54]相比,延迟降低了3个数量级,带宽降低了2个数量级.

针对Chameleon^[90]、GAZELLE^[91]等框架由于使用了计算密集型的密码原语而导致不能充分利用CNN高效的并行数据结构,CNN模型不易于部署到资源受限型的移动传感器上等问题,Huang等人^[58]提出了一种轻量级隐私保护框架LPP-CNN,用于基于边缘计算的移动传感器中CNN特征的提取,系统架构如图9所示.作者基于A-SS和乘法三元组设计了一系列高效的安全交互子协议,并利用两个边缘服务器和一个可信第三方协同执行CNN特征提取.可信第三方负责在脱机阶段生成随机值.由于不需要对CNN结构做任何近似处理,因此确保了CNN模型的准确性.由于不依赖于计算密集型的加密原语,极大地减少了计算和通信开销.

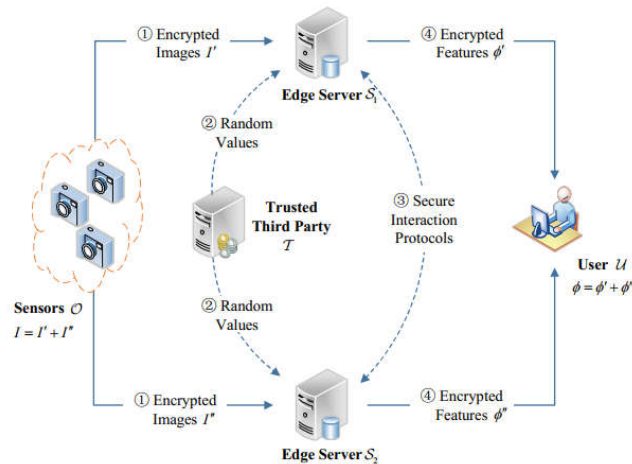


Fig.9 System architecture for LPP-CNN^[58]

图 9 LPP-CNN系统架构图^[58]

针对边缘计算的隐私保护问题,Ma等人^[93]也设计了一种轻量级隐私保护框架POR.该方案基于A-SS和边缘计算技术,利用两个服务器协同完成基于AdaBoost集成的人脸识别分类任务.针对AdaBoost的不同训练阶段,设计了一系列交互协议.实验表明,与现有的基于差分隐私的框架相比,POR可以减少约58%的计算误差.文献[59]提出了一种双服务器轻量级隐私保护框架OPSR,用于面向智能物联网设备的隐私保护语音识别.该方案基于长时记忆(LSTM)神经网络、边缘计算和A-SS技术,以实现轻量级的外包计算任务.与基于HE和GC的框架相比,OPSR大大减少了计算和通信开销.文献[63]提出了一种轻量级的隐私保护强化学习框架LiPSG,用于智能电网的能源管理策略制定.LiPSG基于A-SS和边缘计算技术,每个供电区域的电力数据在发送到控制中心之前,首先被安全外包给第三方双服务器进行Q-learning模型计算,在完成Q-learning计算任务过程中,数据始终保持随机共享格式,避免了敌手滥用用户数据.

5.3 综合分析

安全多方计算(SMC)协议允许多个参与者通过使用同态加密、秘密共享和不经意传输等加密技术,在不实际共享输入的地方对数据进行聚合计算.因此,SMC 协议被重点应用在高效并行分布式机器学习中.在某些环境中,这种方法已经被证明可以扩展到具有数亿条记录的学习任务中^[129].但是,与模型上使用差分隐私的方法不同,这些方法只在学习过程中保护了训练数据的隐私,而不能防止对结果模型的推理攻击^[73].

利用安全多方计算协议构造支持多方共同训练机器学习模型的关键在于:

(1)根据不同协议的特点,选用合适的基础密码学工具以保证安全性.例如,同态加密技术擅长线性运算,主要是矩阵-向量乘法运算,总体通信复杂性较低^[91].因此,同态加密技术适合可对较大数值进行特殊加法和乘法运算.

(2)对机器学习模型中的非线性函数设计高效的替代表达式. 现有的定点或浮点乘法技术需要位级操作,这些操作使用布尔电路效率最高^[86].理论上,任何可以表示为布尔电路的函数都可以使用混淆电路技术协议进行安全计算^[90].因而混淆电路技术更适合于 DNN 模型中近似非线性函数.基于 2PC 架构的 SMC 隐私保护方案,例如 HE + GC+SS+OT,提供了强大的隐私保护,但它们大多只适用于两方计算场景中,扩展到多方场景将导致显著的通信开销.另外,由于它们基于复杂的技术,这些技术速度较慢,通常不能用于大型数据集^[71].

6 总结与展望

6.1 典型的隐私保护技术对比

差分隐私、同态加密和安全多方计算等技术具有不同的技术特点、优点和缺点,相应地具有不同的应用场景.表 9 对比分析了机器学习中不同的隐私保护技术.

Table 9 Comparison of different privacy-preserving technologies

表 9 各种隐私保护技术对比

名称	技术特点	优点	缺点	隐私保护场景
差分隐私	噪声扰动数据	隐私性和效率较高	影响分类器的性能,可用性不高	算力较弱环境
同态加密	密文计算	隐私性高	计算、存储开销大,效率、可用性较低	算力强大、隐私要求高
安全多方计算	不暴露隐私联合计算	隐私性和可用性好	通信开销大,效率低	分布式协同学习环境

在实际使用隐私保护技术时,需要考虑用户设备的硬件性能、传输成本和时间约束等诸多因素.例如,当医院或银行等拥有大量敏感数据的组织充当用户时,需要使用同态加密技术来保证模型的安全性.当计算能力较弱的个体作为用户时,则需要使用差分隐私技术保证模型的效率.当分布式环境下,多方协同训练一个机器学习模型,则可能使用安全多方计算技术来保证各方的隐私.越来越多的研究致力于将 SMC、HE 和 DP 等方法结合起来,以达到数据隐私和效用之间的合理权衡.

6.2 研究展望

目前,机器学习已成为大数据、物联网、云计算和人工智能的核心技术.机器学习的各种隐私威胁以及相应的防御机制受到了学术界和工业界越来越多的关注.机器学习隐私保护的研究仍处于起步阶段,仍有许多问题亟待解决,其中以下五个问题值得我们开展进一步研究.

(1) 研究训练阶段基于密文的高效机器学习隐私保护方法

目前,基于加密技术的机器学习隐私保护方法多用于预测阶段,而很少用于训练阶段.原因如下:首先,同态加密生成的密文更大、更复杂,随着运算次数增多,计算电路深度加深,一旦超过阈值,将无法解密得到正确的结果;其次,深度学习本身是一项计算密集型的任务,计算资源以及通信带宽开销大,即使没有加密,也需要高吞吐量的计算单元.而最直接有效保护隐私的方法是使用加密技术.因此,研究训练时基于加密技术的高效机器学习隐私保护方法是一个亟待解决的问题.

(2) 设计适用于机器学习各个阶段的通用隐私保护体系结构

一方面,云平台中现有的许多应用程序无法处理加密数据,必须重新编写应用程序.另一方面,现有 SMC 方法中使用的 HE、GC 和 OT 等技术有其固有的缺陷.例如,HE 只适用于有限类型的运算,不能直接处理机器学习中的非线性运算.GC 由于

需要对电路中的每个门进行几个对称密钥操作,计算复杂度高,只适用于 2 方或 3 方的安全计算,不容易扩展到与更多方用户协作参与.OT 协议需要昂贵的公钥操作,不适合于大数据.因此,研究设计面向大数据的适用于机器学习各个阶段的通用隐私保护体系结构是一个重大挑战.

(3) 提出针对半结构化、非结构化数据隐私保护的切实可行的解决方案

现有的隐私保护机制几乎都是针对结构化数据的,半结构化、非结构化数据隐私泄露严重.而大数据绝大部分由半结构化、非结构化数据组成,结构化数据只占一小部分.最新研究表明:深度学习方法可在社交网络上自动收集和处理用户照片或视频,以惊人准确率检测出物体的类型、识别出个人;反演攻击模型可从人脸识别系统重建图像.传统的隐私保护机制不适合该领域,即使利用密码学方法仍会泄露隐私.因此,保护半结构化、非结构化数据隐私,且不影响在线社交网络等用户的使用体验,是一个很有前途的研究方向.

(4) 实现隐私性、高效性和可用性之间的最佳平衡

机器学习中训练数据的隐私性、模型的高效性和可用性之间相互矛盾.例如,基于差分隐私的防御方法隐私性和效率较高,但由于添加了噪声扰动导致可用性不高;基于同态加密的防御方法隐私性较高,但密文计算中利用多项式近似导致可用性不高;基于安全多方计算的防御方法隐私性和可用性高,但由于各参与方之间交互多、通信开销大导致效率低下.因此,建立隐私保护机制多维评估体系十分必要,在不同模型、不同攻击方式下对三者之间的关系进行建模,实现三者在不同应用场景下的权衡最优化.

(5) 建立统一的隐私泄露度量标准

机器学习隐私保护研究中,如何度量机器学习模型的隐私泄露风险,是风险评估体系中的重要问题.目前已有一些学者关注隐私量化问题,并已开展了一些初步研究工作,但还较为零散,更多地是针对某一特定领域,其应用范围也受到限制.加之隐私泄露涉及因素众多,目前尚未形成统一的模型及体系.因此,建立统一的隐私泄露衡量标准以及完善的隐私泄露风险分析与评估机制是机器学习中有待进一步深入研究的课题.

6.3 总结

以大数据为驱动力的第四次工业革命即将开启人类智能化时代,机器学习已经成为我们日常生活中不可分割的技术,然而机器学习隐私泄露给我们带来巨大威胁.

本文总结分析了机器学习中几种典型的隐私攻击及其防御机制,对机器学习隐私保护主流技术的工作原理和突出特点进行了阐述,对机器学习隐私保护领域的最新研究成果进行了综述.机器学习的隐私泄露及防御是一个动态的攻防过程.随着技术的不断发展,特别是联邦学习、MLaaS 模式的流行,针对模型的隐私攻击手段会越来越多样化,防御所面临的挑战也越来越大.特别是在数据隐私性、模型高效性和可用性这一本质矛盾的前提下,如何提供符合特定场景隐私保护方法,最小化机器学习中的用户隐私泄露风险,将是个长期挑战.

References:

- [1] Ducange P, Pecori R, Mezzina P. A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing*, 2018, 22 (1):325-342. [doi: 10.1007/s00500-017-2536-4]
- [2] Yin Y, Zhang W, Xu Y, Zhang H, Mai Z, Yu L. QoS prediction for mobile edge service recommendation with auto-encoder. *IEEE Access*, 2019, 7:62312-62324. [doi: 10.1109/ACCESS.2019.2914737]
- [3] Google Prediction API. <https://cloud.google.com/prediction>
- [4] Amazon ML. <https://aws.amazon.com/cn/machine-learning/>
- [5] Azure ML. <https://studio.azureml.net/>
- [6] BigML. <https://bigml.com/>
- [7] Song CZ, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proc. of the the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017. 587-601. [doi: 10.1145/3133956.3134077]
- [8] Atenièse G, Felici G, Mancini LV, Spognardi A, Villani A, Vitali D. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. arXiv preprint arXiv:13064447, 2015. [doi: 10.1504/ijsn.2015.071829]
- [9] Rahulamathavan Y, Phan RC-W, Veluru S, Cumanan K, Rajarajan M. Privacy-preserving multi-class support vector machine for outsourcing the data classification in cloud. *IEEE Transactions on Dependable and Secure Computing*, 2014, 11 (5):467-479. [doi: 10.1109/TDSC.2013.51]
- [10] Wilber MJ, Boulton TE. Secure remote matching with privacy: Scrambled support vector vaulted verification (s 2 v 3). In: Proc. of the the IEEE Workshop on the Applications of Computer Vision. Piscataway, NJ: IEEE, 2012. 169-176. [doi: 10.1109/WACV.2012.6163018]
- [11] Bost R, Popa RA, Tu S, Goldwasser S. Machine learning classification over encrypted data. In: Proc. of the the 22nd Annual Network and

- Distributed System Security Symposium. Rosten: The Internet Society, 2015. [doi: 10.14722/ndss.2015.23241]
- [12] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015. 1322-1333. [doi: 10.1145/2810103.2813677]
- [13] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In: Proc. of the the 23rd USENIX Security Symposium. Berkeley, CA: USENIX Association, 2014. 17-32.
- [14] Hayes J, Melis L, Danezis G, De Cristofaro E. LOGAN: evaluating privacy leakage of generative models using generative adversarial networks. arXiv preprint arXiv:170507663, 2017.
- [15] Liu KS, Li B, Gao J. Generative model: Membership attack, generalization and diversity. arXiv preprint arXiv:180509898, 2018.
- [16] Long YH, Bindschaedler V, Wang L, Bu DY, Wang XF, Tang HX, Gunter CA, Chen K. Understanding membership inferences on well-generalized learning models. arXiv preprint arXiv:180204889, 2018.
- [17] Salem A, Zhang Y, Humbert M, Fritz M, Backes M. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint arXiv:180601246, 2018.
- [18] Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Official Journal of the European Union (OJ), 2016, 59 (1-88):294.
- [19] California Consumer Privacy Act (CCPA) Fines and Consumer Damages. <https://www.clarip.com/data-privacy/california-consumer-privacy-act-fines/>
- [20] China's Cyber Security Law. http://www.xinhuanet.com/politics/2016-11/07/c_1119867015.htm
- [21] Li M, Andersen DG, Park JW, Smola AJ, Ahmed A, Josifovski V, Long J, Shekita EJ, Su B-Y. Scaling distributed machine learning with the parameter server. In: Proc. of the 11th USENIX Symposium on Operating Systems Design and Implementation. 2014. 583-598. [doi: 10.1145/2640087.2644155]
- [22] McMahan HB, Moore E, Ramage D, Hampson S. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:160205629, 2016.
- [23] Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R. Advances and open problems in federated learning. arXiv preprint arXiv:191204977, 2019.
- [24] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10 (2):1-19. [doi: 10.1145/3298981]
- [25] Brandeis L, Warren S. The right to privacy. Harvard Law Review, 1890, 4 (5):193-220.
- [26] Joseph, Sarah, Castan, Melissa. The International Covenant on Civil and Political Rights and United Kingdom law. International Covenant On Civil And Political Rights. Oxford: Clarendon Press. 1995.
- [27] Saltzer JH, Schroeder MD. The protection of information in computer systems. Proceedings of the IEEE, 1975, 63 (9):1278-1308. [doi: 10.1109/PROC.1975.9939]
- [28] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32 (5):847-861(in Chinese). [doi: 10.3724/SP.J.1016.2009.00847]
- [29] Sanyal A, Kusner MJ, Gascon A, Kanade V. TAPAS: Tricks to accelerate (encrypted) prediction as a service. arXiv preprint arXiv:180603461, 2018.
- [30] Xie PT, Bilenko M, Finley T, Gilad-Bachrach R, Lauter K, Naehrig M. Crypto-nets: Neural networks over encrypted data. arXiv preprint arXiv:14126181, 2014.
- [31] Rindfleisch TC. Privacy, information technology, and health care. Communications of the ACM, 1997, 40 (8):92-100. [doi: 10.1145/257874.257896]
- [32] Bolton RJ, Hand DJ. Statistical fraud detection: A review. Statistical science, 2002:235-249. [doi: 10.1214/ss/1042727940]
- [33] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:160202697, 2016, 1 (2):3.
- [34] Laskov P. Practical evasion of a learning-based classifier: A case study. In: Proc. of the 2014 IEEE symposium on security and privacy. IEEE, 2014. 197-211. [doi: 10.1109/SP.2014.20]
- [35] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proc. of the Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. 2014. 1054-1067. [doi: 10.1145/2660267.2660348]
- [36] Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proc. of the the 22nd ACM SIGSAC conference on computer and communications security. New York: ACM, 2015. 1310-1321. [doi: 10.1145/2810103.2813687]
- [37] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. arXiv preprint arXiv:181200910, 2018.
- [38] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. In: Proc. of the the ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017. 603-618. [doi: 10.1145/3133956.3134012]
- [39] Melis L, Song CZ, De Cristofaro E, Shmatikov V. Inference attacks against collaborative learning. arXiv preprint arXiv:180504049, 2018.

- [40] Shokri R, Stronati M, Song CZ, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the the IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2017. 3-18. [doi: 10.1109/sp.2017.41]
- [41] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing Machine Learning Models via Prediction APIs. In: Proc. of the the USENIX Security Symposium. Berkeley, CA: USENIX Association, 2016. 601-618.
- [42] Melis L, Song CZ, De Cristofaro E, Shmatikov V. Exploiting Unintended Feature Leakage in Collaborative Learning. In: Proc. of the the IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2019.
- [43] Nikolaenko V, Weinsberg U, Ioannidis S, Joye M, Boneh D, Taft N. Privacy-preserving ridge regression on hundreds of millions of records. In: Proc. of the Security and Privacy (SP), 2013 IEEE Symposium on. IEEE, 2013. 334-348. [doi: 10.1109/SP.2013.30]
- [44] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. In: Proc. of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009. 289-296. [doi: 10.12720/jait.6.3.88-95]
- [45] Li Y, Jiang ZL, Yao L, Wang X, Yiu S, Huang ZA. Outsourced privacy-preserving C4. 5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. Cluster Computing, 2017:1-13. [doi: 10.1007/s10586-017-1019-9]
- [46] Samet S, Miri A. Privacy-preserving back-propagation and extreme learning machine algorithms. Data & Knowledge Engineering, 2012, 79:40-61. [doi: 10.1016/j.datak.2012.06.001]
- [47] Prasad KD, Reddy KaN, Vasumathi D. Privacy-Preserving Naive Bayesian Classifier for Continuous Data and Discrete Data. In: Proc. of the the First International Conference on Artificial Intelligence and Cognitive Computing. Berlin, Heidelberg: Springer-Verlag, 2019. 289-299. [doi: 10.1007/978-981-13-1580-0_28]
- [48] Aslett LJ, Esperança PM, Holmes CC. Encrypted statistical machine learning:new privacy preserving methods. arXiv preprint arXiv:150806845, 2015.
- [49] Beaulieu-Jones BK, Wu ZS, Williams C, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. BioRxiv, 2017:159756. [doi: 10.1101/159756]
- [50] Phan NH, Wu XT, Hu H, Dou DJ. Adaptive laplace mechanism: differential privacy preservation in deep learning. In: Proc. of the the IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2017. 385-394. [doi: 10.1109/ICDM.2017.48]
- [51] Adesuyi TA, Kim BM. A layer-wise Perturbation based Privacy Preserving Deep Neural Networks. In: Proc. of the the International Conference on Artificial Intelligence in Information and Communication. Piscataway, NJ: IEEE, 2019. 389-394. [doi: 10.1109/ICAII.2019.8669014]
- [52] Barni M, Orlandi C, Piva A. A privacy-preserving protocol for neural-network-based computation. In: Proc. of the the 8th workshop on Multimedia and security. New York: ACM, 2006. 146-151. [doi: 10.1145/1161366.1161393]
- [53] Orlandi C, Piva A, Barni M. Oblivious neural network computing via homomorphic encryption. EURASIP Journal on Information Security, 2007, 2007 (1):1-11. [doi: 10.1155/2007/37343]
- [54] Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: Proc. of the the 33rd International Conference on Machine Learning. New York: IMLS, 2016. 201-210.
- [55] Chabanne H, De Wargny A, Milgram J, Morel C, Prouff E. Privacy-preserving classification on deep neural network. IACR Cryptology ePrint Archive, 2017, 2017:35.
- [56] Hesamifard E, Takabi H, Ghasemi M. CryptoDL: Deep Neural Networks over Encrypted Data. arXiv preprint arXiv:171105189, 2017.
- [57] Phan NH, Wu XT, Dou DJ. Preserving differential privacy in convolutional deep belief networks. Machine Learning, 2017, 106 (9-10):1681-1704. [doi: 10.1007/s10994-017-5656-2]
- [58] Huang K, Liu X, Fu S, Guo D, Xu M. A lightweight privacy-preserving CNN feature extraction framework for mobile sensing. IEEE Transactions on Dependable and Secure Computing, 2019. [doi: 10.1109/TDSC.2019.2913362]
- [59] Ma Z, Liu Y, Liu X, Ma J, Li F. Privacy-Preserving Outsourced Speech Recognition for Smart IoT Devices. IEEE Internet of Things Journal, 2019, 6 (5):8406-8420. [doi: 10.1109/JIOT.2019.2917933]
- [60] Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:161005755, 2016.
- [61] Vaidya J, Clifton C. Privacy-preserving k-means clustering over vertically partitioned data. In: Proc. of the the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2003. 206-215. [doi: 10.1145/956750.956776]
- [62] Gheid Z, Challal Y. Efficient and privacy-preserving k-means clustering for big data mining. In: Proc. of the 2016 IEEE Trustcom/BigDataSE/ISPA. Piscataway, NJ: IEEE, 2016. 791-798. [doi: 10.1109/TrustCom.2016.0140]
- [63] Wang Z, Liu Y, Ma Z, Liu X, Ma J. LiPSG: Lightweight Privacy-Preserving Q-Learning Based Energy Management for the IoT-Enable Smart Grid. IEEE Internet of Things Journal, 2020. [doi: 10.1109/JIOT.2020.2968631]
- [64] Zhang QC, Yang LT, Chen ZK. Privacy preserving deep computation model on cloud for big data feature learning. IEEE Transactions on Computers, 2016, 65 (5):1351-1362. [doi: 10.1109/TC.2015.2470255]
- [65] Trieu PL, Aono Y, Hayashi T, Wang LH, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 2018, 13 (5):1333-1345. [doi: 10.1109/TIFS.2017.2787987]
- [66] Hesamifard E TH, Ghasemi M, Et Al. Privacy-preserving machine learning in cloud. the 2017 on Cloud Computing Security Workshop, 2017:39-43. [doi: 10.1145/3140649.3140655]

- [67] Chillotti I, Gama N, Georgieva M, Izabachene M. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In: Proc. of the the International Conference on the Theory and Application of Cryptology and Information Security. Berlin, Heidelberg: Springer-Verlag, 2016. 3-33. [doi: 10.1007/978-3-662-53887-6_1]
- [68] Courbariaux M, Hubara I, Soudry D, El-Yaniv R, Bengio Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint arXiv:160202830, 2016.
- [69] Kim M, Smaragdis P. Bitwise neural networks. arXiv preprint arXiv:160106071, 2016.
- [70] Bourse F, Minelli M, Minihold M, Paillier P. Fast homomorphic evaluation of deep discretized neural networks. In: Proc. of the the Annual International Cryptology Conference. Berlin, Heidelberg: Springer-Verlag, 2018. 483-512. [doi: 10.1007/978-3-319-96878-0_17]
- [71] Mehnaz S, Bellala G, Bertino E. A secure sum protocol and its application to privacy-preserving multi-party analytics. In: Proc. of the the 22nd ACM on Symposium on Access Control Models and Technologies. New York: ACM, 2017. 219-230. [doi: 10.1145/3078861.3078869]
- [72] Bansal A, Chen TT, Zhong S. Privacy preserving back-propagation neural network learning over arbitrarily partitioned data. *Neural Computing and Applications*, 2011, 20 (1):143-150. [doi: 10.1007/s00521-010-0346-z]
- [73] Jayaraman B, Wang L, Evans D, Gu Q. Distributed learning without distress: Privacy-preserving empirical risk minimization. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 6343-6354.
- [74] Xie LY, Lin KX, Wang S, Wang F, Zhou JY. Differentially Private Generative Adversarial Network. arXiv preprint arXiv:180206739, 2018. [doi: 10.475/123_4]
- [75] Bindschaedler V, Shokri R, Gunter CA. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment*, 2017, 10 (5):481-492. [doi: 10.14778/3055540.3055542]
- [76] Abadi M, Chu A, Goodfellow I, Mcmahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016. 308-318. [doi: 10.1145/2976749.2978318]
- [77] Liu MH, Jiang HT, Chen J, Badokhon A, Wei XT, Huang MC. A Collaborative Privacy-Preserving Deep Learning System in Distributed Mobile Environment. In: Proc. of the the International Conference on Computational Science and Computational Intelligence. Piscataway, NJ: IEEE, 2017. 192-197. [doi: 10.1109/CSCI.2016.42]
- [78] Phan N, Wang Y, Wu XT, Dou DJ. Differential Privacy Preservation for Deep Auto-Encoders: an Application of Human Behavior Prediction. In: Proc. of the the thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016. 1309-1316.
- [79] Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson Ú. Scalable Private Learning with PATE. arXiv preprint arXiv:180208908, 2018.
- [80] Li M, Chow SS, Hu S, Yan Y, Du M, Wang Z. Optimizing Privacy-Preserving Outsourced Convolutional Neural Network Predictions. arXiv preprint arXiv:200210944, 2020.
- [81] Liu L, Su J, Liu X, Chen R, Huang K, Deng RH, Wang X. Toward Highly Secure Yet Efficient KNN Classification Scheme on Outsourced Cloud Data. *IEEE Internet of Things Journal*, 2019, 6 (6):9841-9852. [doi: 10.1109/JIOT.2019.2932444]
- [82] Dani V, King V, Movahedi M, Saia J, Zamani M. Secure multi-party computation in large networks. *Distributed Computing*, 2017, 30 (3):193-229. [doi: 10.1007/s00446-016-0284-9]
- [83] Abbasi S, Cimato S, Damiani E. Toward secure clustered multi-party computation: a privacy-preserving clustering protocol. In: Proc. of the Information and Communication Technology-EurAsia Conference. Berlin, Heidelberg: Springer-Verlag, 2013. 447-452. [doi: 10.1007/978-3-642-36818-9_49]
- [84] Bogdanov D, Niiitsoo M, Toft T, Willemsen J. High-performance secure multi-party computation for data mining applications. *International Journal of Information Security*, 2012, 11 (6):403-418. [doi: 10.1007/s10207-012-0177-2]
- [85] Asharov G, Lindell Y, Schneider T, Zohner M. More efficient oblivious transfer extensions. *Journal of Cryptology*, 2017, 30 (3):805-858. [doi: 10.1007/s00145-016-9236-6]
- [86] Mohassel P, Zhang YP. SecureML: A system for scalable privacy-preserving machine learning. In: Proc. of the the 38th IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2017. 19-38. [doi: 10.1109/SP.2017.12]
- [87] Rouhani BD, Riazi MS, Koushanfar F. Deepsecure: Scalable provably-secure deep learning. In: Proc. of the the 55th ACM/ESDA/IEEE Design Automation Conference. Piscataway, NJ: IEEE, 2018. 1-6. [doi: 10.1145/3195970.3196023]
- [88] Liu J, Juuti M, Lu Y, Asokan N. Oblivious neural network predictions via miniomn transformations. In: Proc. of the the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017. 619-631. [doi: 10.1145/3133956.3134056]
- [89] Chandran N, Gupta D, Rastogi A, Sharma R, Tripathi S. EzPC: Programmable, Efficient, and Scalable Secure Two-Party Computation. ePrint Report, 2017, 1109.
- [90] Riazi MS, Weinert C, Tkachenko O, Songhori EM, Schneider T, Koushanfar F. Chameleon: A hybrid secure computation framework for machine learning applications. In: Proc. of the the Asia Conference on Computer and Communications Security. New York: ACM, 2018. 707-721. [doi: 10.1145/3196494.3196522]
- [91] Juvekar C, Vaikuntanathan V, Chandrakasan A. Gazelle: A low latency framework for secure neural network inference. the 27th USENIX Security Symposium, 2018:1651-1669.
- [92] Henecka W, Sadeghi A-R, Schneider T, Wehrenberg I. TASTY: tool for automating secure two-party computations. In: Proc. of the the 17th

- ACM conference on Computer and communications security. New York: ACM, 2010. 451-462. [doi: 10.1145/1866307.1866358]
- [93] Ma Z, Liu Y, Liu X, Ma J, Ren K. Lightweight privacy-preserving ensemble classification for face recognition. *IEEE Internet of Things Journal*, 2019, 6 (3):5778-5790. [doi: 10.1109/JIOT.2019.2905555]
- [94] Dwork C. Differential privacy. *Encyclopedia of Cryptography and Security*, 2011:338-340. [doi: 10.1007/11787006_1]
- [95] Dwork C, Roth A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 2014, 9 (3-4):211-407. [doi: 10.1561/04000000042]
- [96] Andrew G, Chien S, Papernot N. TensorFlow Privacy. <https://github.com/tensorflow/privacy>.
- [97] Kifer D, Lin BR. Towards an axiomatization of statistical privacy and utility. In: *Proc. of the Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. New York: ACM, 2010. 147-158. [doi: 10.1145/1807085.1807106]
- [98] Dwork C, Rothblum GN, Vadhan S. Boosting and differential privacy. In: *Proc. of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. Piscataway, NJ: IEEE, 2010. 51-60. [doi: 10.1109/focs.2010.12]
- [99] Ye QQ, Meng XF, Zhu MJ, Huo Z. Survey on local differential privacy. *Ruan Jian Xue Bao/Journal of Software*, 2018, 29 (7):1981-2005(in Chinese).<http://www.jos.org.cn/1000-9825/5364.htm>. [doi: 10.13328/j.cnki.jos.005364]
- [100] Bindschaedler V, Shokri R. Synthesizing plausible privacy-preserving location traces. In: *Proc. of the 2016 IEEE Symposium on Security and Privacy (SP)*. Piscataway, NJ: IEEE, 2016. 546-563. [doi: 10.1109/SP.2016.39]
- [101] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015, 10 (7):e0130140. [doi: 10.1371/journal.pone.0130140]
- [102] Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 2011, 12 (Mar):1069-1109. [doi: 10.1109/MIS.2011.2]
- [103] Dwork C, Meshery F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: *Proc. of the the Theory of cryptography conference*. Berlin, Heidelberg: Springer-Verlag, 2006. 265-284. [doi: 10.1007/11681878_14]
- [104] Mironov I. Rényi differential privacy. In: *Proc. of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. Piscataway, NJ: IEEE, 2017. 263-275. [doi: 10.1109/CSF.2017.11]
- [105] Wu X, Li F, Kumar A, Chaudhuri K, Jha S, Naughton J. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In: *Proc. of the Proceedings of the 2017 ACM International Conference on Management of Data*. 2017. 1307-1322. [doi: 10.1145/3035918.3064047]
- [106] Jayaraman B, Evans D. Evaluating Differentially Private Machine Learning in Practice. *arXiv preprint arXiv:190208874*, 2019.
- [107] Graepel T, Lauter K, Naehrig M. ML confidential: Machine learning on encrypted data. In: *Proc. of the the International Conference on Information Security and Cryptology*. Berlin, Heidelberg: Springer-Verlag, 2012. 1-21. [doi: 10.1007/978-3-642-37682-5_1]
- [108] Li ZY, Gui XL, Gu YJ, Li XS, Dai HJ, Zhang XJ. Survey on homomorphic encryption algorithm and its application in the privacy-preserving for cloud computing. *Ruan Jian Xue Bao/Journal of Software*, 2018, 29 (7):1827-1851(in Chinese).<http://www.jos.org.cn/1000-9825/5354.htm>. [doi: 10.13328/j.cnki.jos.005354]
- [109] Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 2018, 51 (4):79. [doi: 10.1145/0000000.0000000]
- [110] Gentry C. Fully homomorphic encryption using ideal lattices. In: *Proc. of the the 41st annual ACM Symposium on theory of computing*. New York: ACM, 2009. 169-178. [doi: 10.1109/TIFS.2013.2287732]
- [111] Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory*, 2014, 6 (3):13. [doi: 10.1145/2090236.2090262]
- [112] El-Yahyaoui A, El Kettani MDE-C. An Efficient Fully Homomorphic Encryption Scheme. *IJ Network Security*, 2019, 21 (1):91-99. [doi: 10.6633/IJNS.201901 21(1).11]
- [113] Ichibane Y, Gahi Y, Guennoun M, Guennoun Z. Fully Homomorphic Encryption Without Noise. *International Journal of Smart Security Technologies (IJSST)*, 2019, 6 (2):33-51. [doi: 10.4018/IJSST.2019070102]
- [114] Chillotti I, Gama N, Georgieva M, Izabachène M. TFHE: fast fully homomorphic encryption over the torus. *Journal of Cryptology*, 2020, 33 (1):34-91.
- [115] Baryalai M, Jang-Jaccard J, Liu D. Towards privacy-preserving classification in neural networks. In: *Proc. of the 2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 2016. 392-399. [doi: 10.1109/PST.2016.7906962]
- [116] Stone MH. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 1948, 21 (5):237-254. [doi: 10.2307/3029750]
- [117] Bos JW, Lauter K, Loftus J, Naehrig M. Improved security for a ring-based fully homomorphic encryption scheme. In: *Proc. of the the IMA International Conference on Cryptography and Coding*. Berlin, Heidelberg: Springer-Verlag, 2013. 45-64. [doi: 10.1007/978-3-642-45239-0_4]
- [118] Naehrig M, Lauter K, Vaikuntanathan V. Can homomorphic encryption be practical? In: *Proc. of the the 3rd ACM workshop on Cloud computing security workshop*. New York: ACM, 2011. 113-124. [doi: 10.1145/2046660.2046682]
- [119] Yao AC. Protocols for secure computations. In: *Proc. of the the 23rd Annual Symposium on Foundations of Computer Science*. Piscataway, NJ: IEEE, 1982. 160-164. [doi: 10.1109/SFCS.1982.38]
- [120] Rabin MO. How to exchange secrets with oblivious transfer. *IACR Cryptology ePrint Archive*, 2005, 2005:187.
- [121] Jiang H, Xu QL. Secure Multi-party computation in cloud computing. *Journal of Computer Research and Development*, 2016, 53

- (10):2152-2162(in Chinese). [doi: 10.7544/issn1000-1239.2016.20160685]
- [122] Ishai Y, Kilian J, Nissim K, Petrank E. Extending oblivious transfers efficiently. In: Proc. of the Annual International Cryptology Conference. Berlin, Heidelberg: Springer-Verlag, 2003. 145-161. [doi: 10.1007/978-3-540-45146-4_9]
- [123] Yao AC. How to generate and exchange secrets. In: Proc. of the the 27th Annual Symposium on Foundations of Computer Science. Piscataway, NJ: IEEE, 1986. 162-167. [doi: 10.1109/SFCS.1986.25]
- [124] Lindell Y, Pinkas B. A proof of security of Yao's protocol for two-party computation. Journal of cryptology, 2009, 22 (2):161-188. [doi: 10.1007/s00145-008-9036-8]
- [125] Bellare M, Hoang VT, Rogaway P. Foundations of garbled circuits. In: Proc. of the the 2012 ACM conference on Computer and communications security. New York: ACM, 2012. 784-796. [doi: 10.1145/2382196.2382279]
- [126] Shamir A. How to share a secret. Communications of the ACM, 1979, 22 (11):612-613. [doi: 10.1007/978-3-642-15328-0_17]
- [127] Goldreich O, Micali S, Wigderson A. How to play any mental game. In: Proc. of the the nineteenth annual ACM symposium on Theory of computing. New York: ACM, 1987. 218-229. [doi: 10.1145/28395.28420]
- [128] Huang Y. Practical Secure Two-Party Computation [Ph.D. Thesis]. Charlottesville: University of Virginia, 2012.
- [129] Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Privacy-preserving distributed linear regression on high-dimensional data. Proceedings on Privacy Enhancing Technologies, 2017, 2017 (4):345-364. [doi: 10.1515/popets-2017-0053]

附中文参考文献:

- [28] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847-858. [doi: 10.3724/SP.J.1016.2009.00847]
- [99] 叶青青,孟小峰,朱敏杰,霍峥.本地化差分隐私研究综述.软件学报,2018,29(7):1981-2005. <http://www.jos.org.cn/1000-9825/5364.htm> [doi: 10.13328/j.cnki.jos.005364]
- [108] 李宗育,桂小林,顾迎捷,李雪松,戴慧珺,张学军.同态加密技术及其在云计算隐私保护中的应用.软件学报,2018,29(7):1830-1851. <http://www.jos.org.cn/1000-9825/5354.htm> [doi: 10.13328/j.cnki.jos.005354]
- [121] 蒋瀚,徐秋亮.基于云计算服务的安全多方计算.计算机研究与发展,2016,53(10):2152-2162. [doi: 10.7544/issn1000-1239.2016.20160685]