软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(3):866−892 [doi: 10.13328/j.cnki.jos.005904]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

机器学习中的隐私攻击与防御*

刘睿瑄[1,2], 陈 红[1,2], 郭若杨[1,2], 赵 丹[1,2], 梁文娟[1,2], 李翠平[1,2]

[1](数据工程与知识工程教育部重点实验室(中国人民大学),北京 100872)

[2](中国人民大学 信息学院,北京 100872)

通讯作者: 陈红, E-mail: chong@ruc.edu.cn

摘 要: 大数据时代丰富的信息来源促进了机器学习技术的蓬勃发展,然而机器学习模型的训练集在数据采集、模型训练等各个环节中存在的隐私泄露风险,为人工智能环境下的数据管理提出了重大挑战.传统数据管理中的隐私保护方法无法满足机器学习中多个环节、多种场景下的隐私保护要求.分析并展望了机器学习技术中隐私攻击与防御的研究进展和趋势.首先介绍了机器学习中隐私泄露的场景和隐私攻击的敌手模型,并根据攻击者策略分类梳理了机器学习中隐私攻击的最新研究;介绍了当前机器学习隐私保护的主流基础技术,进一步分析了各技术在保护机器学习训练集隐私时面临的关键问题,重点分类总结了 5 种防御策略以及具体防御机制;最后展望了机器学习技术中隐私防御机制的未来方向和挑战.

关键词: 数据管理;机器学习;隐私保护;隐私攻击

中图法分类号: TP18

中文引用格式: 刘睿瑄,陈红,郭若杨,赵丹,梁文娟,李翠平.机器学习中的隐私攻击与防御.软件学报,2020,31(3):866−892. http://www.jos.org.cn/1000-9825/5904.htm

英文引用格式: Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020,31(3):866−892 (in Chinese). http://www.jos.org.cn/1000-9825/5904.htm

Survey on Privacy Attacks and Defenses in Machine Learning

LIU Rui-Xuan[1,2], CHEN Hong[1,2], GUO Ruo-Yang[1,2], ZHAO Dan[1,2], LIANG Wen-Juan[1,2], LI Cui-Ping[1,2]

[1](Key Laboratory of Data Engineering and Knowledge Engineering of the Ministry of Education (Renmin University of China), Beijing 100872, China)

[2](School of Information, Renmin University of China, Beijing 100872, China)

Abstract: In the era of big data, a rich source of data prompts the development of machine learning technology. However, risks of privacy leakage of models' training data in data collecting and training stages pose essential challenges to data management in the artificial intelligence age. Traditional privacy preserving methods of data management and analysis could not satisfy the complex privacy problems in various stages and scenarios of machine learning. This study surveys the state-of-the-art works of privacy attacks and defenses in machine learning. On the one hand, scenarios of privacy leakage and adversarial models of privacy attacks are illustrated. Also, specific works of privacy attacks are classified with respect to adversarial strategies. On the other hand, 3 main technologies which are commonly applied in privacy preserving of machine learning are introduced and key problems of their applications are pointed out. In addition, 5 defense strategies and corresponding specific mechanisms are elaborated. Finally, future works and challenges of privacy preserving in machine learning are concluded.

* 基金项目: 国家重点研发计划(2018YFB1004401); 国家自然科学基金(61532021, 61772537, 61772536, 61702522)

Foundation item: National Key Research and Development Program of China (2018YFB1004401); National Natural Science Foundation of China (61532021, 61772537, 61772536, 61702522)

本文由人工智能赋能的数据管理、分析与系统专刊特约编辑李战怀教授、于戈教授和杨晓春教授推荐.

收稿时间: 2019-07-19; 修改时间: 2019-09-10; 采用时间: 2019-11-25; jos 在线出版时间: 2019-12-05

CNKI 网络优先出版: 2019-12-05 14:55:00, http://kns.cnki.net/kcms/detail/11.2560.TP.20191205.1454.003.html

Key words: data management; machine learning; privacy preserving; privacy attack

机器学习作为人工智能的核心技术,旨在从数据中学习经验、构建模型,并逐步提升模型的精确程度.随着深度学习等突破性技术的兴起,机器学习迎来了阶段性的发展,得到了学界和产业界的密切关注,并在智慧医疗、商品推荐、人脸识别、网络安全、证券市场分析等各个领域得到广泛应用.

海量数据为机器学习模型提供丰富的训练数据来源,但其中不可避免地包含用户的隐私信息,机器学习中的隐私泄漏以及泄露造成的危害不可忽视,例如:医疗专家基于病人的数据构建了预测病情的模型,攻击者通过隐私攻击可以推断出训练集中病人的数据,甚至 DNA 信息,进而利用这些信息有针对性地犯罪;攻击者还可以预测某个人的数据是否在目标模型的训练集中,进而泄露个人的患病信息并引发歧视问题.同时,机器学习中的隐私泄露还会造成服务商的重大损失:使服务商在面临巨额赔偿风险的同时失去用户信任.更严重的情况下,不法分子将利用泄露的训练数据或模型信息,对机器学习系统进行安全攻击,进而干预模型的预测.

丰富多样的机器学习场景进一步提升了训练集隐私泄露的风险:目前逐步成熟的机器学习云服务 MlaaS (machine learning as a service)中,云端首先收集用户数据并训练模型,最终将训练好的模型接口提供给用户调用,或者提供模型参数以便用户下载.此过程需要用户将数据发送至不可信云平台,因此数据收集阶段就存在隐私泄露风险;即使云平台可信,如果训练好的模型接口被销售给第三方,或者被嵌入移动应用端以供所有用户访问,不可信第三方可以在预测阶段通过隐私攻击[1]窃取模型训练集的隐私.另一种新场景是协同训练,即多方在不共享数据的情况下共同训练一个模型,比如:银行希望和其他金融机构共同训练预测借贷风险的模型,医院希望和其他医疗机构共同训练预测患者身体指标的模型等.此时不需要用户上传数据,而是通过安全多方计算协议或者参数平均完成模型训练.然而安全多方计算无法防御模型在预测阶段中的隐私攻击,普通的参数平均将会面临参数服务器[2]或者其他计算参与者[3,4]在训练阶段发起的隐私攻击.

为了更好地构建人工智能环境下的数据管理标准,机器学习技术应当在保证用户隐私、合理正确使用数据的前提下发展.但是,面向传统数据收集和发布的隐私保护方法已不能适用于机器学习的保护需求,机器学习中存在的隐私泄露为采集、存储、分析等数据管理环节提出了新的挑战.其原因主要来自以下两方面.

- 首先,在使用训练数据构建机器学习模型的过程中出现了不同于传统数据收集与发布的特殊环节,这些特殊环节带来了新的隐私攻击机制.例如:在模型训练阶段,不可信的服务器或者参与者可以利用训练的中间结果构建攻击模型以侵犯用户隐私;在模型预测阶段,即使服务商没有发布模型参数,不可信攻击者也可以通过不断访问模型预测接口的方式窃取目标训练集的数据隐私.另外,深度学习模型的不可解释性也为防御此类攻击带来了难题;
- 其次,多样的机器学习场景对隐私保护提出了高效性和可用性的要求.例如:云平台提供 MLaaS 服务的初衷是提供模型参数或者访问接口供用户使用,因此从可用性角度需要确保目标模型的预测准确率;协同训练中不仅需要保证训练模型的预测准确率,还需要在通信开销以及计算开销方面保证训练过程的高效性.机器学习隐私保护中对高效性及可用性的要求比传统数据管理中更加复杂和严苛,因此亟需设计针对机器学习的隐私保护方案.

总体而言,本文综述机器学习技术中隐私攻击和隐私保护的最新研究进展和研究方向.一方面从“矛”的角度:第 1 节概述机器学习中的隐私泄露背景,首先明确攻击场景并阐述敌手假设,将隐私攻击者从敌手目标、敌手知识、敌手能力、敌手策略等 4 个角度进行抽象和建模.第 2 节基于攻击策略对现有隐私攻击进行分类阐述,重点介绍隐私攻击的适用范围、攻击模型.另一方面,从“盾”的角度:第 3 节介绍用于保护机器学习训练集隐私的主流技术,包括差分隐私(differential privacy)、同态加密、安全多方计算,阐明其基础定义、实现机制,并讨论这些技术应用于机器学习隐私保护时的关键问题.第 4 节重点对现有隐私防御方案进行分类梳理,按照防御思路总结为扰动策略、泛化策略、近似策略、对抗策略、本地策略等 5 类,并列举具体机制的相关研究.第 5 节针对机器学习中隐私保护问题的关键,总结评价现有工作并展望未来的研究方向.最后,第 6 节总结全文.

1 机器学习中的隐私风险

机器学习中数据收集、模型训练、模型预测等环节紧密结合,共同构成机器学习系统的闭环.在探讨机器学习中的隐私攻击和防御之前,明确泄露发生的场景、掌握攻击者的背景将有助于进一步理解隐私攻击、设计防御方案.本节首先对隐私攻击的场景进行分类,接着从敌手目标、敌手知识、敌手能力、敌手策略等 4 个方面阐述机器学习中的敌手模型.

1.1 攻击场景

攻击场景是指机器学习中可能造成隐私泄露的环节,是攻击者进行隐私攻击的突破口,是设计防御方案必须要明确的背景之一.目前,研究机器学习隐私泄露问题的必要性来源于:(1) 训练数据集中有敏感信息;(2) 机器学习训练或者预测阶段存在不可信参与方.如今机器学习模型的构建需要海量的训练数据,而且用户对个人敏感数据的定义范围很广,训练或预测数据中包含敏感信息是十分普遍的情况.下面介绍现有机器学习的两大类场景,并重点对不可信参与方的情况展开讨论.

1.1.1 集中式学习

集中式学习是指由中心服务器完成数据收集、模型训练、模型预测等流程的机器学习方式,其中,用户的原始数据存放与于中心服务器,中心服务器和模型访问者是用户的不可信第三方.

在数据收集阶段,虽然已有法律[5]对收集用户数据的权限做出规定,但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仍可能过度收集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直接的隐私泄露.目前,谷歌[6]、苹果[7]等公司已采用本地化差分隐私的技术保护数据收集过程.

在模型预测阶段,隐私威胁来源于不可信第三方对模型的访问请求.集中式学习的中心服务商得到训练完毕的模型之后,通过直接在用户端部署或者提供 MLaaS 平台的 API 访问接口这两种方式发布模型,因此,模型的发布对象可能是不可信的用户移动端或者购买模型接口的不可信第三方,如图 1 所示.此种情况下,攻击者可以对模型进行成员推断攻击(membership inference attack)、模型倒推攻击(model inversion attack)以及模型参数提取攻击(model extraction attack)等,若模型算法由服务器之外的不可信第三方设计[8],则其可以在模型算法中嵌入恶意模块,并在访问时实施攻击,为模型的发布带来更大的隐私隐患.

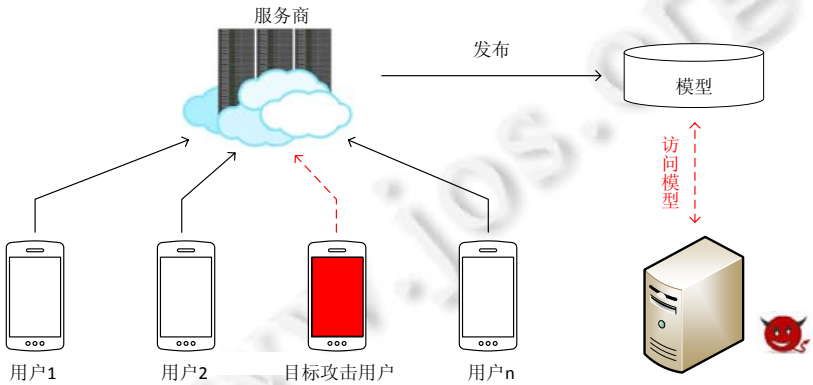

Fig.1 Privacy leakage of centralized learning in predicting stage

图 1 集中式学习中的隐私泄露-模型发布阶段

1.1.2 联合式学习

联合式学习即多个数据所有者在不向中心服务器上传本地数据的前提下,共同学习同一目标模型,以实现移动端计算或者数据共享的需求.此时,攻击者可能是中心服务器或者是任意一个训练参与方.联合学习中没有数据收集阶段,各方保留本地数据且独立地参与模型训练.现有的联合学习中隐私攻击的研究主要集中在模型训练阶段.

1.2 敌手模型

敌手模型刻画了隐私攻击的假设背景,是设计防御机制的首要假设.下面将从敌手目标、敌手知识、敌手能力、敌手策略这 4 个方面[9]对隐私攻击的敌手行为进行分类和分析,表 1 总结了机器学习中隐私攻击的敌手模型,由于攻击者在数据收集阶段直接获取用户数据,因此本节重点讨论训练阶段和预测阶段的敌手模型.

Table 1 Adversarial model

表 1 敌手模型

<table>
<tr><th>攻击阶段</th><th>敌手目标</th><th>敌手知识</th><th>敌手能力</th><th>敌手策略</th></tr>
<tr><td>数据收集阶段</td><td rowspan="4">机密性</td><td>–</td><td>直接获取数据</td><td>–</td></tr>
<tr><td>训练阶段</td><td rowspan="2">白盒/
黑盒</td><td>训练过程干预训练
训练阶段收集中间结果</td><td>模型倒推
成员推断</td></tr>
<tr><td rowspan="2">预测阶段</td><td rowspan="2">访问模型/提取
其他辅助信息</td><td>成员推断
模型倒推
模型参数提取</td></tr>
<tr><td>黑盒</td></tr>
</table>

1.2.1 敌手目标

机密性与完整性、可用性一同构成机器学习模型的评价指标,机密性威胁是指攻击者获取模型或模型训练集数据的隐私信息;完整性威胁是指攻击者有目的性地诱导模型的输出结果;可用性威胁是指攻击者阻止或妨碍普通用户对模型的正常请求.隐私攻击中的敌手目标是模型的机密性.

从被攻击的具体效果来看,敌手目标大致可以分为以下 3 个方面.

(1) 判断某条个人数据是否在目标模型的训练集中.例如:目标模型是基于癌症患者的基因数据训练的,一旦攻击者判断出某条数据存在于训练集,则可推断该条数据拥有者的患癌情况;

(2) 推断训练数据中某列或若干列的敏感属性值.例如:训练集是基因数据,且若干基因序列和疾病直接相关,一旦攻击者掌握相关背景知识,并通过隐私攻击推断出目标攻击对象的敏感基因序列,则会侵犯患者隐私;

(3) 重建分类模型训练集中某一类数据.若目标模型为人脸识别模型,攻击者通过隐私攻击可重建出某人的人脸图片,从而将该个体的姓名和外貌联系起来,侵犯个人隐私.

现有的隐私攻击中,目标模型主要为监督学习,并涵盖判别模型和生成模型两大类.判别模型是指由数据学习联合概率分布 $P(Y|X)$,然后求出条件概率分布 $P(Y|X)$作为预测的模型;生成模型是指由数据直接学习决策函数 $f(X)$或者条件概率分布 $P(Y|X)$.

1.2.2 敌手知识

敌手知识是指攻击者掌握的关于目标模型的背景知识,包括模型训练集的分布假设、其他辅助统计信息、模型结构和参数、决策函数等.其中,是否掌握模型结构和参数,是决定攻击方式和攻击力度的关键.因此,本文将攻击者掌握模型结构和参数的隐私攻击划分为白盒攻击,将攻击者没有掌握模型结构和参数的攻击划分为黑盒攻击.具体攻击方案中,对攻击者知识的假设有可能介于两者之间,例如:攻击者掌握模型的结构,不知道模型的参数.并且具体攻击中可以先提升攻击者背景,再发起隐私攻击.例如:有研究[10]在攻击训练集前先提取模型参数,将攻击难度大的黑盒场景转化为更易被攻击的白盒场景.

黑盒攻击下敌手知识最弱,因此如果某一模型的训练集隐私能被黑盒攻击窃取,则该模型的隐私防御能力很弱,在面对白盒攻击时更容易泄露训练集隐私.白盒攻击下敌手知识较强,因此如果某一模型能抵御白盒攻击,则该模型的防御能力很强,且能抵御同类的黑盒攻击.

1.2.3 敌手能力

敌手能力是指攻击者对目标模型的操作权限.在机器学习的数据收集阶段,敌手能力指直接获取数据;在机器学习的训练阶段,敌手能力包括干预模型训练、收集中间结果的能力;在机器学习的预测阶段,敌手能力是指访问模型、提取模型或部分数据等辅助信息的能力.

根据攻击者的介入能力,可以将隐私攻击分为主动攻击和被动攻击:主动攻击中,攻击者的敌手能力包括参

与模型的训练,甚至恶意使用特定策略诱导目标模型泄露更多信息;被动攻击中,对敌手能力的假设控制在不影响模型完整性和可用性的范围内,即攻击者不直接参与模型训练,而是通过访问模型、观察输出、获取辅助信息等方式达到攻击目的.

1.2.4 敌手策略

敌手目标、敌手知识、敌手能力这三者共同决定了攻击者采取的敌手策略.除了数据收集阶段直接获取数据的方式,敌手策略可分为:

(1) 直接攻击:攻击者构建攻击模型直接攻击目标模型的训练集数据隐私,包括判断某个用户数据是否在训练集中以及倒推用户数据;

(2) 间接攻击:首先构建攻击模型窃取模型参数,利用该参数作为直接攻击训练集数据的背景知识,增大攻击模型训练集成功率,进一步攻击机器学习模型训练集.

具体而言,现有 3 种具体的策略:成员推断攻击、模型倒推攻击和模型参数提取攻击.成员推断攻击和模型倒推攻击为直接攻击策略,提取模型参数为间接攻击策略.

2 机器学习中的隐私攻击

面对上述不同场景、不同敌手模型下机器学习技术中存在的威胁,诸多研究通过设计攻击模型证实了机器学习中隐私威胁的破坏力,典型攻击的研究成果总结见表 2.本节将以表 2 中的敌手策略为主线,分别阐述成员推断攻击、模型倒推攻击以及参数提取攻击的主要研究,并对比分析不同攻击之间的关联.

Table 2 Classic privacy attacks in machine learning

表 2 机器学习中典型隐私攻击

	敌手知识			敌手能力			敌手策略			敌手目标	
	模型结构	模型参数	类型	请求模型	训练模型	设计模型	模式	具体策略	类型	模型类型	场景类型
Shokri, 2017[1]	√	×	黑盒	√	×	×	被动	成员推断	直接攻击	神经网络	集中式
	×	×		√	×	×				任意模型	
Nasr, 2019[3]	√	√	白盒	√	×	×	被动			神经网络	集中式
	√	√		√	√	×					联合式
	√	√		√	√	×	主动				
Hayes, 2017[10]	×	×	黑盒	√	×	×	被动			GAN/VAE	集中式
	√	√	白盒	√	×	×				GAN	
Fredrikson, 2014[11]	√	×	黑盒	√	×	×	被动	模型倒推		线性回归	集中式
Fredrikson, 2015[12]	√	×	黑盒	√	×	×	被动			决策树	集中式
	√	√	白盒	√	×	×				神经网络	
	√	√		√	×	×					
	√	×	黑盒	√	×	×					
Hitaj, 2017[4]	√	√	白盒	√	√	×	主动				联合式
Wang, 2019[2]	√	√	白盒	√	√	√	主动				联合式
	√	√		√	√	×	被动				
Song, 2017[8]	√	√	白盒	√	√	√	主动				集中式
	√	×	黑盒	√	√	√					
Tramèr, 2017[13]	√	×	黑盒	√	×	×	被动	参数提取	间接攻击	逻辑回归 决策树 SVM 神经网络	集中式

2.1 成员推断攻击

成员推断攻击是指攻击者试图判断某条个人信息是否存在于目标模型的训练数据集.当训练数据包含医疗数据等敏感信息,数据的拥有者并不想暴露个人数据在特定训练集中存在与否,然而成员推断攻击泄露了这类隐私.

Shokri 等人[1]提出,可以利用机器学习模型在训练集和非训练集上表现的差别进行黑盒下的成员推断攻

击,如图 2 所示.假设目标模型的类型和结构是公开的,攻击者拥有和目标模型训练集同分布的数据集.根据上述敌手知识,攻击者首先训练出 k 个模仿目标模型预测行为的影子模型(shadow model).随后,他们用影子模型的训练集以及非训练集数据分别请求影子模型的输出,并标记为“在数据集”和“不在数据集”两类.例如以在影子模型训练集 $D^{train}_{shadow^i}$ 中的数据(x,y)为输入,攻击者得到影子模型的输出为 $y'=f^i_{shadow}(x)$,并将(y,y',in)添加至攻击模型训练集 D^{train}_{attack}.得到攻击模型训练集之后,攻击者将此训练集 D^{train}_{attack} 分为若干份,每一份对应目标模型输出的一个类别.最后,对于目标模型的每一个类别,训练一个输入为模型预测 y'、输出为 x 的二分类器作为攻击模型.

另外,文献[1]还指出,当攻击者没有与目标模型训练集数据同分布的数据集时,仍可能通过 3 种方案训练影子模型:(1) 通过不断请求目标模型,使人造数据尽可能接近训练集在模型上的表现;(2) 利用有关于目标模型训练数据的统计信息(比如不同特征的边缘分布),合成训练集;(3) 借助含有噪声的相似数据集.因此,不论是否有额外数据集作为敌手知识,攻击者总可以得到类似目标模型的影子模型,并通过观察影子模型在训练集和非训练集上的不同表现,训练一个有监督的二分类器作为攻击模型,预测某条数据存在于训练集中的可能性,实现成员推断攻击.

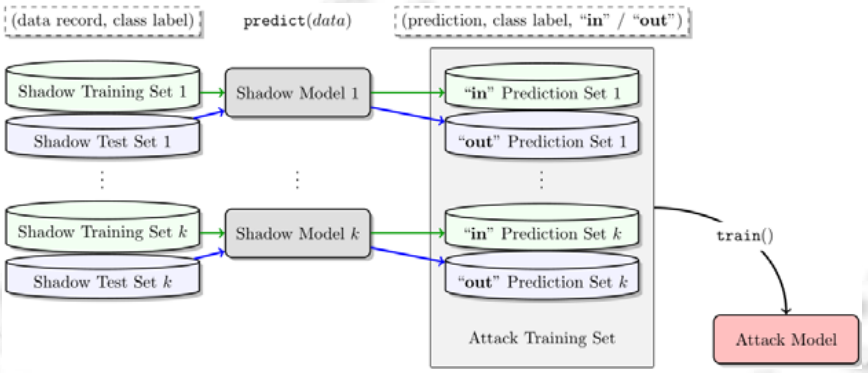

Fig.2 An example of membership inference[1]

图 2 成员推断攻击举例[1]

上述攻击利用了目标模型泛化能力有限这一缺点,而判断一个模型是否容易受到成员推断攻击不仅仅受到模型泛化能力这一个因素的影响.对于泛化能力相对较好的神经网络模型,Nasr 等人[3]假设攻击者以神经网络模型的结构和参数为敌手知识进行成员推断攻击.他们首次提出了分层的攻击方式,把神经网络每层的输出和梯度信息作为攻击模型的输入特征,分别输入到若干全连接层和卷积层中,构建攻击模型,在 CIFAR 数据集上达到了 75.1%的白盒攻击准确率,证明泛化能力很强的神经网络模型仍有隐私泄露风险.他们还在攻击者没有掌握目标模型训练集成员标签的无监督情况下设计了白盒下的成员攻击,其中假设攻击者拥有一组和目标模型训练集部分相交的数据,利用编码器(encoder)生成每个数据点对应的编码向量,输入至解码器(decoder)提取关键输入特征,完成对编码器的训练.最后,将想要判别成员信息的所有目标数据输入该攻击模型,通过对结果聚类完成对成员信息的判别.

以上研究的目标模型聚焦于集中式学习的判别模型,在联合式学习中,也存在成员推断攻击.Nasr 等人[3]首次提出这种情况下的主动成员推断攻击,由于中心服务器或者参与方都能观察到每一轮参数的变化,若攻击者是参与方之一,通过在目标数据上进行反向梯度更新,并观察多次反向更新之后梯度的变化,即可判断该数据在其他参与方训练集中是否存在.若攻击者是联合分布式机器学习模型的中心服务器,则可通过修改发送给目标攻击者的模型参数进行攻击.

上述研究攻击的目标模型均为判别模型,因此模型在训练集和测试集上表现的差异可以通过模型输出的置信度衡量.但是在生成模型中,并不容易判断模型是否过拟合,因而也不容易发现生成模型是否存在成员隐私泄露的风险.Hayes 等人[10]基于生成对抗网络首次提出了以生成模型为目标模型的成员推断攻击.在白盒攻击

中,他们假设攻击者可以直接获取生成对抗模型的判别器,根据目标模型在自身训练集上发生过拟合时判别器输出的置信度很高这一现象,没有训练额外的攻击模型就完成了对目标模型的成员推断攻击,并达到100%的准确率.在黑盒攻击中,他们通过目标生成模型产生的样本数据构造了一个等同于目标模型的生成对抗网络,遂将黑盒攻击转化为白盒成员推断攻击,并达到了80%的攻击准确率.

2.2 模型倒推攻击

模型倒推是指通过模型的输出反推训练集中某条目标数据的部分或全部属性值,本文中,此概念包含部分研究提到的属性推断及模型重建.

当攻击者采取被动攻击的方式,在不干预模型训练过程的情况下进行模型倒推,其基本思路是找到使输出中某一类对应的可能性最大的输入.Fredrikson 等人[11]的研究中将病人的人口统计信息作为辅助信息,以预测药物剂量的线性回归模型作为目标模型,根据模型输出恢复出患者的部分基因组信息.这项研究证明,即使攻击者仅有对模型预测接口的访问能力,也可以通过反复请求目标模型得到训练集中用户的敏感数据.然而当敏感属性有更大的维数和更多取值情况,按此类方法遍历所有可能的取值需要消耗大量的计算开销,因此限制了攻击者进行模型倒推的能力.

Fredrikson 等人[12]又提出了适用更多模型、有更强攻击力度的模型倒推攻击.对于决策树模型,他们利用MLaaS 平台输出的置信度以及决策树的结构和参数信息,在白盒情况下推断模型训练集的敏感属性取值.他们的结论和上述攻击一致,即攻击某个属性的请求次数与目标的敏感属性可能取值的个数成线性关系.另外,他们对人脸识别模型进行倒推攻击的结果显示:该方法结合一些图像技术之后,可以非常近似地还原出训练集中某个标签对应的数据.如图3中,攻击模型根据右图标签生成左图.该攻击表明:不仅倒推出某一列属性会侵犯训练集中用户隐私,获取某一类数据的平均值或者近似值也是泄露训练集隐私的一种方式.

Fig.3 An example of model inversion[12]

图3 模型倒推攻击举例[12]

上述攻击方法采用的是被动策略,如果攻击者采取主动策略直接干扰模型训练过程,攻击者将对目标模型拥有更大的掌控权.Hitaj 等人[4]在白盒场景下对深度学习模型发起模型倒推攻击,还原出用户的人脸数据.不过与上述工作不同的是,他们是基于联合分布式学习的场景,假设攻击者是参与联合学习的某一方,并采用主动的策略参与目标模型的学习过程,诱导联合分布式学习的各参与方提供更加精确的数据.最后他们指出:只要参与方本地的模型精度足够,就可以实现较高的攻击准确率.不同于 Hitaj 的研究,Wang 等人[2]提出的白盒模型倒推攻击中假设攻击者是负责平均参与方梯度的中心服务器,设计了一种多任务的生成对抗模型作为攻击模型,成功还原了某一参与方训练集中的图片.

上述隐私攻击并没有影响模型的可用性和完整性,如果模型算法的设计是恶意,不仅会暴露更多、更准确的训练集的信息,还将破坏完整性和可用性.Song 等人[8]设想了如下攻击场景:数据拥有者从模型提供者处获得模型的算法代码,然后在自己的数据集上运行.假设模型提供者是潜在的攻击者,其在设计算法的时候对训练算法稍作修改,即可在模型的发布阶段获取想要得到的具体训练数据.在白盒攻击中,攻击者把经过编码的训练集数据隐藏在最终模型的参数中;在黑盒攻击中,攻击者在算法设计阶段用假数据对原训练数据进行扩充,并把部分真实训练集数据编码为假数据的“标签”,在最终的训练阶段同时完成数据拥有者要求的真实训练任务和用

于攻击的恶意训练任务.在模型发布阶段,当攻击者用假数据对模型进行请求的时候,返回的即是攻击者想要获取的已编码在“标签”中的真实训练集.

2.3 参数提取攻击

模型参数提取是指当目标模型参数不公开,攻击者已知部分模型结构信息和标签信息,试图通过访问目标模型得到模型参数的攻击.总体而言,攻击者发起模型参数提取的动机包括:避免向模型训练服务缴费;规避恶意邮件分类等模型的检测,发起安全攻击;掌握模型参数之后,增加对模型训练集的攻击成功率.前两者均为机器学习技术在模型安全层面需要讨论的问题,本文提到的模型参数提取主要关注的是第 3 点,即训练集数据隐私的问题.

Tramer 等人[14]提出并扩展了等式求解的模型提取方法,并对仅输出标签的目标模型进行讨论,如图 4 中通过反复请求模型预测标签得到模型的参数.此方法对于多分类的逻辑回归和神经网络模型也适用.设目标模型为 f,攻击模型可等价为随机算法$\mathcal{A}$,其目标是通过请求 f 以获得一个和 f 非常类似的模型 $\hat{f}$.他们定义了测试误差 R_{test} 和均一误差 R_{unif},并用 $1-R_{test}$ 和 $1-R_{unif}$ 作为评估攻击准确率的指标,定义如下:

$$R_{test}(f,\hat{f})=\sum_{(x,y)\in D}\frac{d(f(x),\hat{f}(x))}{|D|} \tag{1}$$

$$R_{unif}(f,\hat{f})=\sum_{x\in U}\frac{d(f(x),\hat{f}(x))}{|U|} \tag{2}$$

测试误差的含义是测试集 D 上的平均误差,其值较小,表示 $\hat{f}$ 和 f 在输入数据的分布上十分接近;均一误差的含义是对于从训练集分布中均匀采样的向量 U,其评估了 $\hat{f}$ 和 f 在特征空间上不同的比例.若用 x 表示模型输入,用 $h_{theta}(x)$ 表示模型的输出,这种方法可以归结为已知$(x;h_{theta}(x))$求解模型参数 $theta$,即:对于一个 N 维的线性模型,理论上在 N+1 次查询之后就可以提取它的参数.

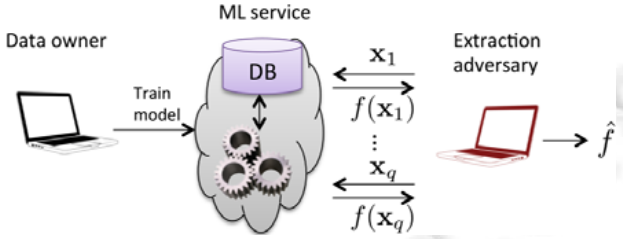

Fig.4 An example of model extraction attack[13]

图 4 参数提取攻击举例[13]

2.4 各类隐私攻击的比较分析

表 2 中,在敌手知识方面,已知模型参数的攻击均为白盒攻击;在敌手能力方面,攻击者参与到模型设计的攻击均为主动攻击;在敌手目标方面,现有隐私攻击的目标模型已从线性模型扩展到深度模型;在具体敌手策略方面,成员推断和模型倒推攻击可以发生在黑盒或白盒场景下,攻击者可采取主动或者被动的攻击模式,而参数提取攻击的敌手背景仅限于黑盒,攻击模式为被动攻击,且仅存在于集中式机器学习场景中.这是由于参数提取攻击的目标就是已有模型的参数,因而假设攻击者不参与模型训练过程.

表 2 中,我们将成员推断攻击和模型倒推攻击归类为直接攻击的敌手策略,模型参数提取归类为间接敌手策略.实质上,间接攻击与直接攻击有着紧密的联系.一旦攻击者通过模型提取攻击得到模型参数,敌手知识将从原先的黑盒扩展为白盒,因此参数提取攻击为成员推断攻击和模型倒推攻击做出了充分的准备.目前,已有研究证实了模型提取攻击对这两种攻击的促进作用,例如:如果攻击者掌握了例如 SVM[14]或隐马尔可夫模型[15]等机器学习模型的参数,则可以倒推获取训练集数据,比如在语音识别模型中获取演讲者的口音.Shokri 等人[1]也指出,如果模型的训练算法和模型结构已知,白盒成员推断攻击的准确率和效率将远远比黑盒成员推断攻击

的准确率和效率高.因此,防御模型参数提取攻击也是机器学习的隐私防御中需要考虑的重要隐私攻击之一.本文第 4 节中众多防御方案都是从这一角度出发,通过防止对参数的推断降低成员推断和模型倒推这两种直接攻击的可能.

另外,成员推断攻击和模型倒推攻击也有着紧密的关联.Long 等人[16]和 Yeom 等人[17]就两者关系进行了分析,并得出结论:(1) 模型倒推攻击中隐含着成员推断攻击,反之却没有充分的依据,因为模型倒推攻击正是对在训练集中数据的倒推,只有当某条数据在训练集中,模型倒推攻击才有意义;(2) 若模型过拟合问题严重,在训练集上的预测表现明显优于测试集,则该模型面对这两种攻击的抵抗力都很薄弱.

3 机器学习中的隐私防御技术

为应对上述隐私攻击,目前已有很多研究对机器学习中的隐私防御方案进行讨论,其中,主流的三大类技术分别是差分隐私、同态加密、安全多方计算.不同背景下的不同防御方案都基于这 3 种技术展开,为方便后续介绍具体防御方案,本节将分别介绍这 3 种技术的基础定义、实现机制以及应用在机器学习模型中的关键问题.

3.1 差分隐私

3.1.1 基础定义

差分隐私使得某一条数据是否在数据集中,几乎不影响算法的计算结果,其定义如下.

定义 1(差分隐私)[18]. 设有随机算法 M,P_M 为 M 所有可能的输出构成的集合,对于任意两个只有一条数据不同的数据集 D 和 D'以及 P_M 的任何子集 S_M,若算法 M 满足:

$$\Pr[M(D)\in S_M]\leqslant \exp(\varepsilon)\times \Pr[M(D')\in S_M]+\delta \tag{3}$$

则称算法 M 提供(ε,δ)-差分隐私保护,其中,参数ε为隐私保护预算,δ代表可容忍的隐私预算不成立的概率.若δ=0,则称算法 M 提供ε-差分隐私保护.

由于在数据采集过程中就存在用户隐私数据被窃取的风险,无需任何可信方的本地化差分隐私技术也逐步成为近年的研究热点.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,因此避免了服务器直接收集或接触到本地原始数据,同时还能完成对总体数据的统计分析.本地化差分隐私定义如下:

定义 2(本地化差分隐私)[19]. 设隐私算法 M 的定义域为 D,值域为 R,若 M 由任意两个输入 t 和 $t'(t,t'\in D)$得到相同输出 $t^*(t^*\subseteq R)$满足以下不等式,则 M 满足ε-本地化差分隐私:

$$\Pr[M(t)=t^*]\leqslant \exp(\varepsilon)\times \Pr[M(t')=t^*] \tag{4}$$

3.1.2 实现机制

差分隐私中主要存在两种直接添加噪声的方法:拉普拉斯机制(Laplace mechanism)[18]、高斯机制(Gaussian mechanism)[20],这两种机制适用于保护数值型数据.保护非数值型数据常用的方法是指数机制(exponential mechanism)[21].以上方法的定义如下(其中,添加噪声的大小取决于敏感度(sensitivity)的计算,敏感度是指改变数据集中任一记录对查询结果造成的最大影响):

定义 3(拉普拉斯机制)[18]. 给定数据集 D,设有函数 $f:D\to R^d$,其敏感度为Δf,则称随机算法 $M(D)=f(D)+Y$ 提供ε-差分隐私保护,其中,$Y\sim Lap(\Delta f/\varepsilon)$为随机噪声,服从尺度参数为$\Delta f/\varepsilon$的拉普拉斯分布.

定义 4(指数机制)[21]. 设随机算法 M 的输入为数据集 D,输出对象 $r\in R$,$q(D,r)$为可用性函数,Δq 为可用性函数的敏感度.若算法 M 以正比于 $\exp(\varepsilon q(D,r)/2\Delta q)$的概率从 R 中选择并输出 r,则称算法 M 提供ε-差分隐私保护.

本地化差分隐私的实现方法主要基于随机扰动(randomized response)[22],下面以一个具体例子[23]简要阐述随机扰动机制的原理.假设服务器想要统计所有 n 个用户中属于某一群体 A 的比例π,因此向每个用户发起查询“请问你是否属于群体 A?”,为了在收集查询结果时不侵犯用户的个人隐私,规定用户根据如下规则响应该查询:以 p 的概率回答真实答案,以 $1-p$ 的概率回答真实答案的相反答案.最终服务器对收集到的响应进行矫正,即可估计 n 个用户中属于群体 A 的比例,其中,n_1 是收集结果中回答为“是”的数目:

$$\hat{\pi}=\frac{p-1}{2p-1}+\frac{n_1}{(2p-1)n} \tag{5}$$

3.1.3 关键问题

差分隐私技术通过严格的隐私定义[24]使攻击者几乎不能通过模型输出分辨某一条数据是否被用于训练机器学习模型,保护了机器学习模型训练集中每个个体的数据隐私.得益于差分隐私后处理性和组合性的支持,现有多种方式在机器学习中添加噪声,如生成噪声数据[25−27]、在模型训练中对目标函数添加噪声[28]、在模型训练中对参数或梯度添加噪声[29−33]、在模型的输出结果上添加噪声[34]等方式.例如图 5 中目标模型在训练过程中给梯度信息添加了噪声,随着隐私预算从左到右降低,添加的噪声增大,攻击者通过模型倒推攻击[4]得到数据的可用性也逐步降低.目前,该技术已经被用于 SVM[30]、回归[30,33,35],决策树[36]和神经网络[29,32,37]等机器学习模型的隐私防御.

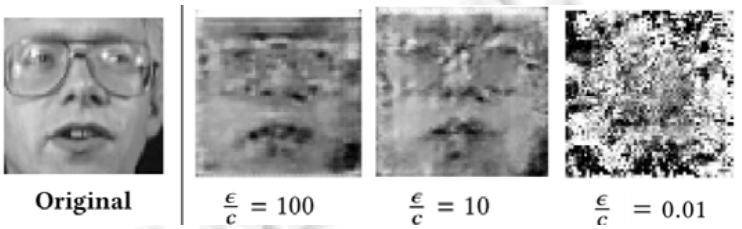

Fig.5 Example of how differential privacy defends model inversion attack[4]

图 5 差分隐私防御模型倒推攻击举例[4]

差分隐私和成员攻击的定义都与攻击者推断出一条数据是否存在于模型训练集的可能性直接相关,因此,差分隐私的隐私开销直接决定了防御成员推断攻击的效果.当前,关于成员推断攻击的研究[1,3,38]也明确指出了这一点.但是对于模型倒推攻击和参数提取攻击,隐私预算与防御效果并不直接相关,因此,防御这两种攻击时隐私预算和防御效果之间的权衡成为了关键问题.当前可以确定的是,降低隐私开销可以降低这 3 种隐私攻击的准确性,但同时,这将在目标模型中引入大量噪声,降低目标模型可用性.若隐私开销过大,即添加的噪声过小,攻击者仍有可能获得较接近真实值的参数,达不到防御效果.因此,隐私性和可用性之间的平衡是差分隐私技术在防御所有机器学习中隐私攻击时面临的关键问题.

在机器学习模型中实际部署差分隐私技术并不会带来过多额外的计算开销,Wu 等人[34]的研究表明:在模型输出上添加差分隐私噪声的方法和非隐私的传统算法相比,其运行时长几乎一致.因此,部署的关键在于:

(1) 需重新设计已有系统或者算法.除了在模型输出中添加扰动以外,其他在目标函数以及参数或梯度上添加噪声的方法都需要修改机器学习的内部算法;

(2) 为算法调试带来困难.差分隐私算法会引入额外的超参数,增加了机器学习模型的算法调试工作量.

3.2 同态加密

3.2.1 基础定义

同态加密是一种允许用户对密文进行特定代数运算并得到密文结果的加密形式,保护了数据存储以及运算过程中的数据隐私.其同态的含义在于:对用户在密文上进行代数运算的结果解密后,与其在明文上进行相同计算得到的结果相同.密码体系中的安全基础是计算困难问题,同态加密通常基于的计算困难问题包括整数分解问题、离散对数问题、判定合数剩余问题、近似最大公因子问题、系数子集求和问题、二次剩余问题等.根据在加密状态下可以完成的操作,可将同态加密技术分为加法同态、乘法同态以及全同态[39].

3.2.2 实现机制

目前有许多同态加密的实现机制:RSA 机制[39]、El Gamal 体制[40]、Paillier 机制[41]等,下面以广泛应用于机器学习隐私保护机制[42,43]的 Paillier 体制为例,介绍同态加密的实现机制.Paillier 机制是基于合数剩余问题的公钥密码体制.设 E_{pk} 是以 $pk=(N,g)$为公钥的加密函数,其中 N 是两个大素数的乘积,整数 $g \in Z_{n^2}^*$;D_{sk} 是以 sk 为私钥的解密函数,则给定两个整数值 $a,b \in Z_N$,Paillier 加密体制满足以下性质.私钥 sk 的拥有者对最终得到的密文

结果解密即可得到目标运算结果:

(1) 同态加法特性: $E_{pk}(a+b)\leftarrow E_{pk}(a)\times E_{pk}(b) \bmod N^2$

(2) 同态乘法特性: $E_{pk}(a\times b)\leftarrow E_{pk}(a)^b \bmod N^2$

3.2.3 关键问题

同态加密技术可以在密文上完成运算的特性使其可以应用于服务商与用户的交互,例如在集中式学习中,用户将数据以密文形式上传至服务器,服务器训练模型的同时并不知道原始训练集的数据因而保护了用户数据隐私;在联合学习中,各个参与者将模型参数或者梯度以密文的形式上传至服务器,服务器在不知道每个参与者真实的上传内容的同时完成参数汇总与总体模型迭代,保护了用户的中间计算结果,因而保护了用户原始数据的隐私.

虽然理论上全同态加密可以进行任意计算,但是目前,该方案在机器学习领域的应用存在诸多约束,例如:(1) 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,而同态加密算法只支持整数类型的数据[44,45];(2) 同态加密方案中需要固定乘法深度,不能无限进行加法和乘法运算[46],因而不能应用于包含较多复杂运算的神经网络模型;(3) 全同态加密不支持比较和取最大值等运算,因此不能支持机器学习中常见的幂运算(如 sigmoid 激活函数)等运算;(4) 同态加密方案计算量大,当前的计算硬件和通信设施难以满足实际需求.另外,Hesamifard 等人[47]的研究表明:在密文上训练简单神经网络的时间随着梯度下降的批数据数量的增大而减小;同时,批数据数目的增长需要更大的存储开销,因此需要选取合适的学习参数以平衡性能.

3.3 安全多方计算

3.3.1 基础定义

安全多方计算是一种无需可信第三方参与即可协助多方完成密文计算的技术,其形式化定义如下.

定义 5(安全多方计算)[48]. 假设有 n 个参与方 $P_1,\ldots,P_n$,每个参与方都有一个秘密输入 $m_i(i=1,2,\ldots,n)$.这 n 个参与者共同执行一个协议π来计算函数 $f(m_1,\ldots,m_n)$,且不泄露每个参与者 P_i 的输入信息.

由于安全多方计算中各个参与方有可能不按照协议的规则来执行,甚至在计算过程中输入虚假信息,因此各个参与方根据其表现可分为:

(1) 诚实参与方:在协议中完全按照约定的协议完成运算,对自己所有的输入和得到的输出信息保密;

(2) 半诚实参与方:完全按照协议规则执行,但是可能将自己的输入以及得到的输出结果泄露给攻击者;

(3) 恶意参与方:按照攻击者的角度执行协议,不但泄露自己所有的输入/输出信息,并且有可能改变输入或者篡改中间输出信息甚至终止协议.

3.3.2 实现机制

安全多方计算的构造需要使用基本的密码学工具,包括秘密共享、同态加密、零知识证明、不经意传输等等.下面简要介绍几种常用于机器学习隐私保护的实现机制.

(1) 秘密共享[49]:秘密共享解决的问题中,消息分发者 d 将其持有的信息 m 分为 $m_1,\ldots,m_n$ 分别发送给参与者 $P_1,\ldots,P_n$,并只允许 $k(2\leqslant k\leqslant n)$个参与者共同参与,才可以恢复信息 m;

(2) 同态加密[50]:同态加密可以实现在密文上的运算,并且满足运算的同态性质.设明文消息 m 的密文为 $E(m)$,密文上的两种运算记为⊕和⊗,k 为密钥,则满足以下同态性质的为同态方案:

$$E_k(a)\otimes E_k(b)=E_k(a\oplus b) \tag{6}$$

(3) 零知识证明[51]:设有 P 与 V 两方,P 代表证明方,V 代表验证方,P 知道某个信息并向 V 证明自己知道这个信息,但是又不想向 V 泄露这条信息.

3.3.3 关键问题

安全多方计算应可以用于保护集中式机器学习模型的训练过程,例如多方的数据和模型初始参数基于秘密共享协议,以秘密的形式存储在两个不共谋的服务器上,随后,这两个服务器通过两方安全协议完成模型的训练.由于其中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,因此保护了用户隐私.另外,安全多方计算主要用于保护联合式机器学习模型的构建,如在没有服务器参与的情况

下,各参与方通过交换中间计算结果的密文完成训练,其中需要用到加解密技术完成明密文转换,或者用零知识证明技术验证数据的一致性.

安全多方计算是多种密码学基础工具的综合应用,因此密码学理论为其提供了强大的安全保证,但是由于在实现安全多方计算时广泛应用了同态加密技术,因此,第 3.2.3 节中提到的同态加密技术在机器学习上面临的挑战也是安全多方计算保护机器学习隐私的瓶颈.在实际部署中,基于混淆电路技术的安全多方计算方案一般应用于两至三方完成模型训练的场景,基于秘密共享技术的安全多方计算方案可以扩展至数以百计的用户.但是用户数量的增长将会为安全多方协议带来大量额外的通信开销,例如,Bonawitz 等人[52]指出:服务器训练模型的时间随用户数量的幂次增长,因此仅限于数以百计的用户完成多方训练.对于这个问题,目前已有简化了安全多方计算的机器学习隐私保护方案[42,53,54],但另一个关键问题是,在现实应用中无法确保服务器不共谋的假设.

4 机器学习中的隐私防御方案

目前,已有诸多研究应用和扩展了上述 3 种技术,探索了机器学习中隐私攻击的防御方法,本节将现有隐私防御方案分类总结为 5 种策略和若干机制,并在表 3 中列举了典型的研究成果.本节将以表 3 中的防御策略为主线,介绍基于扰动策略、近似策略、泛化策略、对抗策略和本地策略的隐私防御方案,并分析总结.

Table 3 Classic privacy defenses in machine learning

表 3 机器学习中典型隐私防御方法

文献	策略	机制	技术	保护目标	防御阶段	防御场景
Meng, 2018[55]	扰动策略	扰动目标函数	差分隐私	矩阵分解训练集	训练/预测阶段	联合分布式
Chaudhuri, 2011[28]				凸模型训练集	预测阶段	集中式
Abadi, 2016[29]		扰动中间参数		神经网络训练集		
Bassily, 2014[56]				凸模型训练集		
Wu, 2017[34]		扰动输出				
Li, 2014[57]	近似策略	数值近似		原始训练数据	数据收集之后	任意
Acs, 2017[25]						
Bindschaedler, 2017[58]						
Hesamifard, 2017[47]		函数近似	同态加密	神经网络训练集	训练阶段	集中式
Fredikson, 2015[12]	泛化策略	泛化输出	无	任意模型训练集	预测阶段	
Shokri, 2017[1]		泛化模型	L2 正则化	分类模型训练集		集中式/成员推断
Nasr, 2018[38]	对抗策略	正则化	生成对抗			
Huang, 2018[59]		对抗扰动	对抗学习	与训练集关联的原始敏感数据	数据收集之后	集中式
Jia, 2018[60]						分布式
Hamm, 2015[61]	本地策略	标签集成	差分隐私	分类模型训练集	训练/预测阶段	联合分布式
Papernot, 2017[37]			差分隐私知识迁移			
Papernot, 2018[62]						
Mohassel, 2017[42]		安全多方	同态加密安全协议	任意模型训练集	训练阶段	
Zheng, 2019[63]				线性模型训练集		
Geyer, 2017[64]		联合学习	差分隐私	任意模型训练集	预测阶段	
McMahan, 2017[31]				神经网络训练集		
Shokri, 2015[32]					训练阶段	
Bonawits, 2017[52]			安全协议			
Wang, 2019[30]			本地化差分	任意模型训练集	训练/预测阶段	

4.1 扰动策略

扰动策略是指在目标函数、中间参数或者输出结果中添加噪声扰动,以提高模型防御能力的方法.例如:本文图 2 中的攻击者通过反复请求目标模型获得影子模型,并以此辅助训练攻击模型.由于噪声扰动具有随机性,如果添加了扰动,攻击者影子模型的准确率将降低,因此该策略可以提高模型的防御能力.

4.1.1 扰动目标函数机制

扰动目标函数机制是指直接在机器学习模型的目标函数中添加噪声扰动,并最小化此目标函数的方

法[28,35,65].设机器学习模型参数为 f,训练集为 D,原始的目标函数为损失函数 $L(f,D)$,根据差分隐私得到的噪声向量为 b,则在该机制中目标函数为

$$L_{priv}(f,D)=L(f,D)+\frac{1}{n}b^{T}f \tag{7}$$

求解模型参数的过程即是对公式(7)最小化的过程.Chaudhuri 等人[28]根据 Dwork 等人[18]的敏感度理论,通过隐私参数控制添加噪声的量,提出通过在目标函数上添加噪声以训练差分隐私模型的方法,并对模型的隐私性和泛化能力提供了证明.不过,其算法的前提是模型的损失函数是凸函数,且具有可微分性等条件,因此对神经网络等非凸模型并不适用.Bassily 等人[56]对 ε-差分隐私、(ε,δ)-差分隐私两种情况下的经验风险最小化进行了讨论,并且减少了上述对损失函数特性的要求.Meng 等人[55]应用目标函数扰动机制,面向社交推荐系统设计了一种用户、好友、与服务商三者共同训练矩阵分解的模型.为了在共同训练一个模型的同时保护用户隐私,在矩阵分解损失函数上添加了噪声项.

4.1.2 扰动中间参数机制

扰动中间参数机制是指在优化目标函数的迭代过程中,在参数的梯度上添加噪声以达到扰动效果的方法.这种扰动机制得益于差分隐私技术的组合理论以及 McSherry 等人[66]提出的累计隐私开销的理论.例如,设 $g_t(x_i)$为梯度下降第 t 次迭代中在数据 x_i 上计算的梯度:

$$g_t(x_i)\leftarrow\nabla_{\theta_t}L(\theta_t,x_i) \tag{8}$$

扰动中间参数时,根据梯度裁剪的大小,设置并添加噪声向量 b,在一个大小为 B 的小批量数据上完成一次迭代:

$$\tilde{g}_t\leftarrow\frac{1}{B}\left(\sum_i g_t(x_i)+b\right) \tag{9}$$

Abadi 等人[29]通过在随机梯度下降过程中对梯度添加高斯噪声来保护训练集数据,并对迭代过程中隐私开销的累计进行了严谨的分析,最终证明,可以在控制隐私开销的同时保证神经网络可用性.

4.1.3 扰动输出机制

扰动输出机制来源于 Dwork 等人[18]提出的敏感度理论,是一种直接在训练好的模型的参数上添加噪声的扰动方法,其首先在干净的数据集上学习无噪模型:

$$f=\text{argmin}\ L(f,d) \tag{10}$$

然后,利用 Laplace 机制或指数机制在参数向量 f 上添加噪声向量 b,再发布 f_{priv}:

$$f_{priv}=f+b \tag{11}$$

例如:文献[67,68]利用拉普拉斯机制添加噪声;文献[69]利用指数机制添加噪声;文献[34]首次指出扰动输出机制在模型部署上的便捷性,并基于扰动输出机制提出一种提高模型准确度的方法,使其无需修改原本机器学习模型,即可直接部署在 RDBMS 系统中.由于差分隐私的扰动直接添加在模型的参数上,也就相当于攻击者不能通过一条请求区分两个不同的参数向量,因此可以避免模型参数提取的攻击,进而为攻击者进一步进行模型训练集的攻击造成了阻碍.

4.2 近似策略

本文将近似策略划分为两种机制:基于差分隐私的数据近似机制、基于同态加密的函数近似机制.通过近似数据,目标模型训练集数据在使用之前已经达到隐私保护的要求,则即使攻击者攻击成功,也无法侵犯用户隐私.通过近似函数,可在密文上计算机器学习模型中加密方案无法计算的函数,从而完成模型训练并保护隐私.

4.2.1 数值近似机制

数据近似机制指使用原始数据的近似数据作为训练集的一类方法.数据提供者通过传统概率统计以及机器学习中的工具生成满足源数据集统计特征的虚假数据,随后交由不可信第三方用于训练模型或者数据分析.

数值近似的方式之一是使用原始数据构建统计模型,并以差分隐私的规则发布数据.

Machanavajjhala 等人[70]首先指出了数据近似可以进行隐私保护的原因来自于构建统计模型时的偏差,以

及从统计模型中采样并发布数据的偏差,并提出了适用于稀疏数据的数据生成方法.Li 等人[57]设计了一种满足差分隐私的 Copula 函数以采样近似数据,在差分隐私的保护下发布直方图,其中,Copula 函数是用来描述多维随机变量之间依赖性的方法,通过这种工具,可以从一维的边缘分布中构建多维联合分布.

数值近似的另一种方式是使用机器学习中的生成对抗模型(generative adversarial network)生成数据,并在训练过程中基于 Abadi 等人[29]的理论在模型中添加噪声以满足差分隐私.Beaulieu 等人[71]在使用原始数据训练生成对抗辅助分类器(AC-GAN[72])时,在判别器的梯度中添加噪声,得到满足差分隐私的生成对抗模型.由于只有判别器的训练过程接触到原始训练数据,因而发布生成器不会对数据隐私造成影响.在此基础上,Acs 等人[25]考虑到多个生成模型同时训练比一个生成模型拟合的速度要快,使用差分隐私核 k-means 先将数据分为 k 类,然后使用单个生成神经网络(如限制玻尔兹曼机或者变分自编码器)学习每一类训练数据分布,并在学习过程中添加噪声,所以发布生成模型的参数受到差分隐私的保护.在衡量生成的数据集是否满足数据使用需要时,他们就基本的计数统计进行了讨论.Xie 等人[26]也提出了一种具有差分隐私保护的生成对抗模型,并利用该模型生成了 MNIST 数据集和电子健康记录数据的人造数据,随后利用这些数据完成分类任务并达到了实际可用的准确率,验证了在差分隐私的保护下,基于生成对抗模型生成的近似数据用于模型训练的可行性.

第 3 种数据近似方式:首先生成数据,测试数据的隐私完备性后,仅发布满足隐私要求的数据.已有研究[58,73]认为,强行让一个生成模型满足差分隐私会极大地损失该模型的可用性,因此度量生成数据隐私性的指标是此类工作的关键.例如,Bindschaedler 等人[58]使用可信否认(plausible deniability)度量生成数据的隐私性,并最终指出满足以下条件的生成模型与差分隐私数据发布的关联:

定义 6(plausible deniability)[27]. k 为一个整数型的隐私参数,且 $k\geqslant 1$,对于任意满足 $|D|\geqslant k$ 的数据集 D,设任意一条由概率生成模型 $\mathcal{M}$ 生成记录 y 的概率为 $\Pr\{y=\mathcal{M}(d)\},d\in D$.若存在至少 $k-1$ 个不同的记录 $d_2,\ldots,d_k\in D\{d_1\}$,且对于任意 $i,j\in\{1,2,\ldots,k\}$ 满足条件:

$$\gamma^{-1}\leqslant\frac{\mathbb{P}r\{y=\mathcal{M}(d_i)\}}{\mathbb{P}r\{y=\mathcal{M}(d_j)\}}\leqslant\gamma \tag{12}$$

则发布记录 y 满足 (k,γ)-plausible deniability.

上述工作中,生成数据的隐私性通过隐私参数度量,可用性根据具体需求主要从以下若干方面进行衡量.

(1) 生成数据的直观质量.例如,图片数据的质量可以直观地根据生成图片的质量判定;

(2) 计数统计的准确率.计数统计是很多数据分析和学习算法的基础,因此可以作为衡量标准之一;

(3) 生成数据分布与原数据分布之间的距离.例如 f-散度、Hellinger 距离、wasserstein 距离等;

(4) 人造数据训练的分类模型的准确率.

4.2.2 函数近似机制

在使用同态加密技术保护机器学习隐私安全的过程中,通常用仅含有加法和乘法的表达式替换一些不便于加密计算的非线性函数,如图 6 所示,右侧为神经网络中常见的 RELU 函数,左侧是其替代函数.本文将此类方法概括为函数近似机制,即寻求原有机器学习算法的近似算法,使其便于完成同态加密运算.

目前有研究在模型预测阶段应用函数近似机制,例如,Dowlin 等人[74]将明文训练好的模型通过函数近似机制转化为可以预测密文数据的模型,利用低阶多项式替代明文神经网络中的非线性激活函数.该方法保护的是预测过程的隐私,其目标在于使提供模型的服务商无法得知用户预测请求的明文,并保证用户不能通过服务商得到任何预测结果之外的信息.

为了保证训练集隐私,需要对原始训练数据加密并在密文数据上训练模型.

Hesamifard 等人[47]利用 Chebyshef 多项式近似模拟激活函数,与前者[74]不同的是,他们在模型训练阶段就将激活函数替换成低阶多项式,最终结果相较前者,在 MNIST 数据集上的精度提高了 0.52%.另外,Wu 等人[75]提出用多项式近似逻辑回归,然后基于同态加密完成模型训练.然而这种方法的复杂度与近似的多项式的次数成指数关系,并且模型的准确率也较非隐私情况下有所下降.综上,此类方法受到训练过程缓慢、同态电路深度有限、客户端和云端通信等诸多限制,目前训练阶段加密的方法不适用于较为复杂的模型.

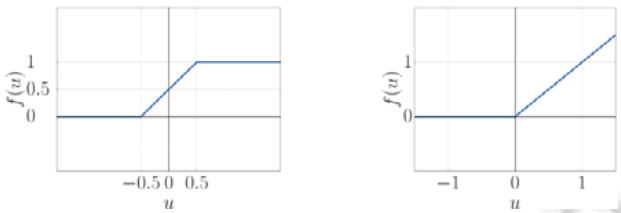

Fig.6 Replacement of activation function[42]

图 6 激活函数的替代表示[42]

4.3 泛化策略

泛化意为由具体的、个别的扩展为一般的,本文将泛化策略分为泛化输出机制以及泛化模型机制两类:模型输出的泛化是指降低模型输出结果的精度;模型的泛化是指通过正则化等手段消除模型在训练集和非训练集的表现差异.

4.3.1 泛化输出机制

泛化输出机制是一种简单初级的防御手段,例如分类模型的输出是一个以分类个数为维度的置信度向量,每个分量代表将输入数据预测为对应类的概率,其中概率最高的一类为预测结果,泛化输出是指降低输出置信度的精度.

Fredrikson 等人[12]指出:由于对人脸识别系统的黑盒攻击模型是通过反复请求目标模型训练的,因此如果降低输出置信度的精度,则相当于降低攻击模型训练集的精度,进而降低模型倒推攻击的准确率.他们的实验表明:当输出分类置信度的精度设置为 0.05,攻击者就不能够构建出可供识别的图片.与此同时,0.05 的精度仍可以满足用户对人脸识别模型的需求.但是由于白盒攻击的攻击者已经掌握了模型的参数,所以这种方法无法防御白盒攻击.与上述限制输出精度的方法不同,Shokri 等人[1]讨论了限制输出置信度维度的情况,即:对于多分类模型只输出置信度最大的若干类,极端情况下只发布置信度最大的类别作为预测标签.表 4 汇总的实验结果表明:如果限制了置信度精度,成员攻击[1]和模型参数提取[12]的准确率将有不同程度的下降,但是对于决策树模型,这种方法不起作用[13];如果限制了置信度的维度,就算在只发布模型输出的标签的极端情况下,成员攻击[1]仍可以达到 66%的准确率.

Table 4 Summary of effects of output generalization mechanism

表 4 泛化输出机制效果小结

限制前	限制后	攻击	攻击类型	目标模型	限制效果
精度 0.1	精度 0.001	Shokri, 2017[1]	成员推断攻击	MLaaS	攻击准确率下降 0.03%
维度 3	维度 1	Shokri, 2017[1]	成员推断攻击	MLaaS	攻击准确率下降 0.03%
精度 10-6	精度 0.05	Fredikson, 2014[12]	模型倒推攻击(黑盒)	神经网络	不能还原出训练集人脸图片
精度 10-6	精度 10-3	Tramer, 2016[14]	参数提取攻击	逻辑回归	攻击错误率增加 10 倍
精度 10-6	精度 10-2	Tramer, 2016[14]	参数提取攻击	决策树	没有影响

以上基于泛化输出机制的方法在实际使用中都具有很大的局限性:一方面,用户对精度有要求;另一方面,就算用户对输出结果的精度和维度要求不高,模型在这种情况下仍有可能被攻击,攻击者只是需要更多的请求次数.实际应用中,不同 MLaaS 平台面向的用户群体对输出精度和维度的要求不一样,因此各平台在为不同用户提供服务时,应该针对用户群的需求提供相应的服务,并对可能造成的隐私泄露风险做出提示.

4.3.2 泛化模型机制

一个模型泛化效果不好是指其在训练集上的表现和在非训练集上的表现存在较大差异,成员推断攻击[1]正是利用了这种差异.如果输出的模型对于任何单个用户的数据都很好地泛化,则能保护单一用户的个人隐私.泛化模型机制是指从改善模型泛化能力的角度防御隐私攻击的方法.

诸多研究都在关注过拟合和攻击的关系,Shokri 等人[1]提出:通过正则化目标函数,减小模型在训练集和测

试集上的差别以防御成员推断攻击.他们在实验中简单地利用 L2 范式作为损失函数的罚项以达到正则化的效果.其中,罚项系数越大,正则化的力度越大.在实验中,随着不断加大正则项系数,目标模型的预测准确率甚至有所增长,因此从提升模型泛化能力的角度出发,模型的隐私性和可用性不一定是矛盾的.但是过大的正则力度,仍会减小目标模型上预测的准确率.

Yoem 等人[17]和上述研究观点一致,认为过拟合会使模型受到成员推断攻击和模型倒推攻击.他们假设成员推断的攻击者知道模型训练集的数据分布和数据集大小,以一个泛化效果好的稳定模型为攻击目标,设计了一种黑盒成员推断攻击,并得到和 Shokri 等人[1]类似的成员推断攻击准确率.并且他们认为:模型的过拟合是模型易受成员推断攻击的充分条件,但不是必要条件.Nasr 等人[3]针对泛化得很好的模型(神经网络模型)在不同场景设定下都设计了具有一定攻击力的成员推断攻击,佐证了上述观点.

4.4 对抗策略

模型设计者面对的问题是如何防御任何可能存在的隐私攻击,但是穷尽任何可能的攻击形式是不现实的.因此,利用游戏理论以及对抗思想构建和优化数据隐私问题受到很多研究者的关注.这类研究[76−79]的目标是在最强攻击下减少隐私损失,并用此时的最小隐私损失度量防御方案对于任何威胁隐私的攻击的风险.对抗策略中,正则化机制和扰动机制也同时属于前文提到的泛化策略和近似策略,介于对抗策略是一种特殊的防御思路,本文将其独立归为一类.

4.4.1 正则化机制

对抗策略下的正则化机制是指以对抗的方式正则化模型,以防御模型发布之后预测阶段的隐私攻击.Nasr 等人[38]受到泛化模型机制的启发并结合生成对抗思想,将攻击者的最大攻击增益作为分类器损失函数的正则项,试图在最小化损失函数的同时最小化最强攻击的攻击能力,以此达到降低攻击风险的目的.其中,攻击者的增益被形式化为攻击者正确分类训练集数据和非训练集数据的对数期望,如图 7 所示.他们的实验表明,这种方法在降低模型过拟合以保护隐私的同时,还在一定的正则项系数范围内提升了总体的分类准确率.这种先计算最大增益再对其求最小值的过程,可以看作防御者与攻击者之间的博弈.不过,该研究中关注的是在黑盒设定下的成员推断攻击的防御,在白盒攻击下无法起到相同的防御效果.

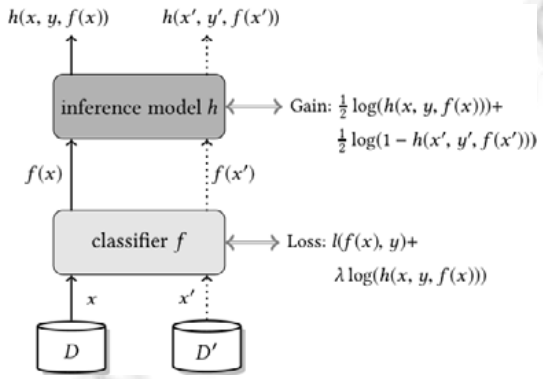

Fig.7 Adversarial generalization mechanism[38]

图 7 对抗策略-正则化机制[38]

4.4.2 扰动机制

对抗策略下的扰动机制是指通过对抗的思想在原始数据中添加噪声,以保证发布数据可用性和隐私之间的平衡,属于数据收集阶段的隐私保护机制.有的研究利用对抗学习模拟博弈过程,Huang 等人[59]基于数据驱动提取数据中的关键统计信息,并利用对抗学习的方法设计了一种上下文感知的隐私架构.此方法通过选择性地在关键数据上添加噪声,实现了可用性和隐私的动态平衡.对于同时包含不敏感属性 X 和敏感属性 Y 的数据集 D,假设每条数据(X,Y)独立且服从同一分布,他们将隐私机制定义为随机映射:

$$\hat{X}=g(X,Y) \tag{13}$$

同时,攻击者基于决策规则 h,对敏感属性 Y 的推断定义为

$$\hat{Y}=h(g(X,Y)) \tag{14}$$

为了量化攻击者的威胁,定义如下损失函数:

$$L(h,g)\triangleq\mathbb{E}[l(h(g(X,Y)),Y)] \tag{15}$$

数据拥有者希望得到一个隐私机制 g,同时满足可用性和隐私性,因此可以形式化为以下优化问题:

$$\left.\begin{aligned}&\min_{g(\cdot)}\max_{h(\cdot)}-L(h,g)\\&\text{s.t. }\mathbb{E}[d(g(X,Y),X)]\leqslant D\end{aligned}\right\} \tag{16}$$

但是,基于博弈理论防御属性推断攻击的隐私保护机制通常难以优化.为了解决这个问题,一些研究[80−83]提出了博弈理论优化问题的近似问题,比如基于关联的启发式方法,但是这类方法会带来很大的可用性损失,并且需要防御者能直接得到用户隐私数据.Jia 等人[60]基于对抗学习中的规避攻击,为每个属性找到一个最小噪声值,最小噪声是指添加到用户的公开数据之后,使得攻击者的分类器以较高概率推断出某敏感属性的最小噪声.随后用解决凸约束优化问题的方法找到一个满足条件的分布,并从该分布中采样出一个属性值,最终在用户的公开数据中添加对应该属性的最小噪声并发布数据.

4.5 本地策略

将数据拥有者的数据保持在本地是一种最简单直观的隐私保护方法,本文将采用此思路的方法归类为本地策略.为了利用存储在用户本地的数据完成训练机器学习模型的任务,数据拥有者与模型训练者需要保持一定的交互,根据交互方式的不同,以下分为标签集成机制、安全多方机制以及联合学习机制.

4.5.1 标签集成机制

标签集成机制借鉴了集成学习的思想,把用户拥有的数据作为总体数据的子集并在各子集上分别训练模型,随后,中心服务器使用公开的无标签数据作为输入,请求本地模型获取标签并对输出结果进行汇总,得到带标签的训练集,最终在该训练集上完成总体模型的训练.以上流程中,不同的研究添加隐私保护机制的环节各不相同.

Hamm 等人[61]在训练总体模型之后,使用前文第 4.1.3 节中提及的模型输出扰动机制,在总体模型的参数上添加噪声并发布.但是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模型集成的过程中通过多次请求,攻击本地训练集的隐私.并且此种方式和集中式学习之后再扰动目标函数的方法相比,实现相同隐私保护程度时模型的准确率更低[28].

为了进一步保护隐私,另一类研究在集成输出结果的环节添加隐私保护机制.Papernot 等人[37]基于知识集成和转移的思想[84],在集成方请求本地模型获取的标签上添加噪声.首先,他们在由不相交的敏感数据子集上分别训练一个教师模型,如图 8 所示;随后,他们在每个教师分类器的投票结果上加入满足差分隐私的拉普拉斯噪声,汇总输入无标签数据时所有教师分类器的预测结果;最终,由公开的无标签数据集和部分具有差分隐私标签的数据组成学生模型的训练集,并利用生成对抗网络(GAN)基于半监督学习完成训练.从直觉上来说,单一的教师模型和数据集中的某条数据不能直接影响它的输出;从理论上来说,它满足差分隐私的形式,因此可以将其视作一种通用的机器学习隐私训练的解决方案.

在上述工作的基础之上,研究者[62]对加噪方法和总体模型的训练方式进行了改进,将投票阶段添加的拉普拉斯噪声换成高斯噪声,基于更易组合的 RDP[85]隐私机制计算总体隐私开销.同时,提出了基于置信度和基于师生交互的两种机制以最大化利用隐私开销,并将虚拟对抗训练(VAT)用于学生模型的半监督学习过程,以增强最终模型的可用性.

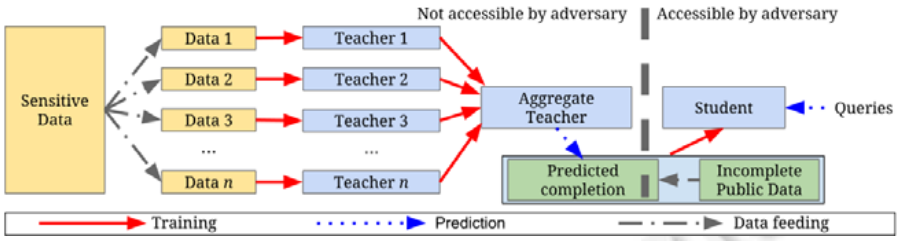

Fig.8 Local label aggregation mechanism[37]

图 8 本地策略-标签集成机制[37]

标签集成机制的优势在于该方法并不需要损失函数以及优化方法假设,然而为了降低差分隐私中的噪声敏感度,基于单一模型的扰动输出机制(见第 4.1.3 节)和扰动目标函数机制(见第 4.1.1 节)都需要相关假设.并且这种机制中将单一数据拥有者隐藏在多方之中,攻击者只能获得对攻击目标的相对粗糙的估计,因此可以抵御模型发布之后的隐私威胁.

4.5.2 安全多方机制

安全多方训练机制意为借助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方法,其中关键在于:(1) 需要选用合适的基础密码学工具以保证安全性;(2) 重构非线性函数,对机器学习模型中的非线性函数设计高效的替代表达式.

早期的研究已将安全多方计算用于决策树[86]、k-means 聚类[87]、SVM[88]、线性回归[53,89]、贝叶斯分类器[90]、逻辑回归[91]等模型,当前,基于安全多方训练机制的机器学习隐私保护方法大致分为两类:一种是将多方训练简化为两方训练的 two-server 架构;另一种是多方的 n-party 架构.

基于 two-server 架构是大多数[53,92,93]安全多方计算训练机器学习模型的思路,最新研究是 SecureML[42]模型,他们假设存在两个不可信且不共谋的服务器 S_1 和 S_2,并保证该模型抵御半诚实攻击者的攻击,即:当攻击者 A 与服务器 S_1 以及用户 C_1,C_2 共谋,则 A 不能得到 C_1,C_2 以外用户的数据.其中,训练过程被分为两个阶段.

(1) 离线阶段:所有用户拥有的训练数据以秘密共享的形式一次性存放在两个服务器中,并按照小批量划分训练数据,服务器计算在线阶段所需要的乘法三元组;

(2) 在线阶段:两个服务器以小批量的形式分别在训练数据的秘密共享上完成随机梯度下降.训练完毕之后,通过秘密共享协议的解密操作得到最终的模型.

他们使用了秘密共享、同态加密、不经意传输等基础密码学工具,分别对线性回归、逻辑回归、神经网络这 3 种模型提出了高效的两方安全计算协议,并且用截断的方式解决了之前研究中两方安全计算中小数计算的瓶颈;同时,在计算乘法运算三元组时通过向量化的技术提升了算法效率.最终,比之前的工作[53]在效率上提升了若干个数量级.

基于 n-party 架构的机器学习隐私保护方案无需借助任何辅助服务器,完全是在没有可信第三方的情况下进行的.Zheng 等人[63]提出了一种在 n-party 安全多方计算下基于交替方向乘子法(ADMM)训练线性模型的安全协议 Helen,使用了部分阈值的同态加密、零知识证明、恶意多方安全计算协议等基础密码学工具.在协商阶段,各方共同约定用一种优化方式训练一个特定模型;在初始化阶段,各个参与方使用一般的安全计算协议得到公开的公钥以及私钥的每一份秘密共享;在输入准备阶段,每一方预先加密本地输入的数据,并以广播的方式把加密的结果发送给各方,并用零知识证明以证明自己知道该密文对应的明文;在模型计算阶段,运行 ADMM 优化方法的迭代过程,各方更新一次本地参数,同时用零知识证明验证自己完成正确运算;经过多轮迭代之后,在模型发布阶段,在使用零知识证明验证各方执行的诚实性之后,即可发布最终模型.

上述两个研究的区别在于:

(1) 架构不同;

(2) 解决的优化问题不同:SecureML 是集中式机器学习中梯度下降优化算法的安全两方计算版本,而

Helen 是分布式机器学习中交替方向乘子法的安全多方计算版本.

4.5.3 联合学习机制

联合学习一词由 Google 正式提出[94],即每轮迭代过程中,各参与方先从中心服务器下载参数,本地更新参数之后上传,参数服务器平均收集到的各方梯度并更新总体模型的参数.联合学习机制和标签集成机制的区别在于,标签集成中需要预先训练本地模型;联合学习机制和安全多方机制的区别在于,联合学习中有参数服务器且主要针对移动设备上数据的联合训练.最初的联合学习认为这种分布式的方式已经能在一定程度上保护隐私,但实际上没有隐私机制保护的联合学习仍会面临隐私威胁[2−4].

Geyer 等人[64]利用时刻累计技术(moments accountant)[29]在平均参数时添加满足差分隐私的噪声,防御了模型训练阶段的隐私攻击.McMahan 等人[31]将这种隐私保护机制应用于 LSTM 模型.但是这类方法中用户传递给服务器的参数仍是根据本地数据计算出的准确梯度,因此无法抵御不可信服务器的隐私攻击.

对于服务器不可信的问题,Shokri 等人[32]在联合学习提出之前就提出了分布式训练保护隐私的工作,其中,本地用户随机选择部分梯度上传,且在上传之前添加满足差分隐私的拉普拉斯噪声.但 Abadi[29]和 Papernot[37]等人认为,这种方法会产生很高的可用性损失.Hitaj 等人[4]利用生成对抗模型针对这种联合分布式训练的神经网络判别模型设计了一种攻击,证明了即使只上传一部分小梯度信息,也可能在模型倒推攻击中泄露用户隐私.

另外,近年来,本地化差分研究逐步成为研究的热点,为服务器不可信的场景提供了解决方案.Wang 等人[30]使用本地化差分隐私技术,在参数服务器采集参数之前,以满足本地化差分的方式扰动各方更新的参数保护了训练过程的隐私.Bonawitz 等人[52]运用秘密共享等技术,为用户向参数服务器上传参数的过程提供了隐私保证.另外,Le 等人[95]改进了 Shokri 等人[32]的研究,在上传梯度时采用加法同态加密算法隐藏梯度信息.

4.6 各类隐私防御的比较分析

表 3 中各种隐私防御方案的防御思路各有不同,但是所有方案都基于第 3 节提到的三大主流技术,并且各种策略之间联系紧密.其中,不同机制的关联与对比关系如图 9 所示,虚线表示两种不同但是相似的攻击机制,箭头表示前一种防御机制可以用于所指的防御机制.

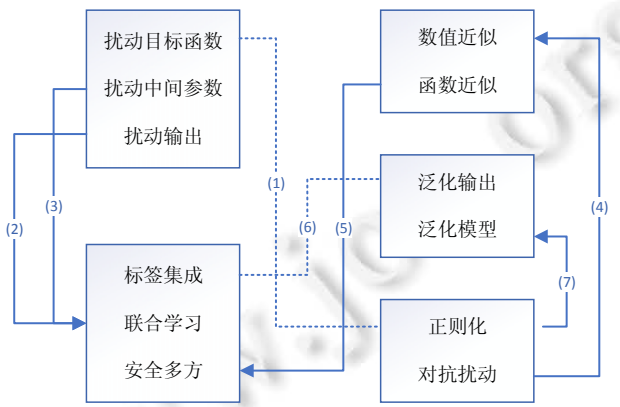

Fig.9 Connections of different privacy defense strategies

图 9 隐私防御策略之间的关联关系

其中,主要的 7 大结论如下.

(1) 扰动目标函数机制与正则化机制类似,都在目标模型训练的目标函数中添加了一项.其中的区别在于:扰动目标函数机制中添加的一项是基于差分隐私需要添加的噪声量确定的;而在对抗策略的正则化机制中,添加的是根据攻击者增益得到的正则项,与差分隐私的定义无关;

(2) 扰动输出机制可用于联合学习中服务器发布模型的过程中,并以此防御观察全局模型参数并试图攻击其他用户的攻击者;

(3) 扰动中间参数机制也可用于联合学习服务器迭代全局模型的过程中,以此防御来自用户的攻击;
(4) 对抗策略下的扰动机制属于数值近似的一种方式,对抗的思想可使近似数据能更好地权衡隐私性和可用性;
(5) 函数近似机制在安全多方机制中有广泛的用途,常用于在密文上进行非线性函数运算的过程中;
(6) 泛化输出策略和标签集成策略的核心思想都是对模型的预测输出进行处理,区别在于泛化输出机制是降低输出置信度的精度,标签集成机制则是对预测标签进行模糊处理;
(7) 对抗策略中的正则化机制实质上是一种泛化模型的方法,对抗的过程是为了控制泛化强度,权衡隐私与可用性.

总体而言,扰动策略作为差分隐私技术的典型应用,可用于其他基于差分隐私的防御方案中.对抗策略作为一种特殊的防御思路,可以为泛化策略和近似策略提供更好的隐私性与可用性之间的权衡.本地策略作为一种新兴的方式,可以广泛地借鉴其他防御机制.

5 研究展望

不同于传统数据管理中的隐私问题,机器学习中新的攻击机制与严苛多样的需求对隐私保护领域提出了新的挑战.目前,相关研究尚处于起步阶段,下面将结合现有研究阐述机器学习中的隐私保护存在的问题和未来的研究方向.

5.1 平衡隐私性、高效性、可用性相互制约的矛盾

隐私保护技术对于机器学习而言是一把双刃剑,一方面保护了模型训练集的隐私,另一方面降低了模型效率和可用性.面对机器学习中丰富的隐私信息以及变化多端的场景,突破隐私性、高效性、可用性相互矛盾制约的瓶颈成为了无本之木.

隐私保护机制对效率的影响体现在计算开销和通信代价两个方面.

(1) 计算开销来源于隐私算法或协议中的额外计算:差分隐私拥有较低的计算开销;而同态加密技术和安全多方计算技术由于引入了密文运算,计算开销不可忽视.目前为止,这两种技术不能应用于复杂的机器学习模型;
(2) 通信代价来自于联合分布式训练过程中用户与服务器之间的交互,与传统分布式训练过程相比,使用差分隐私进行隐私保护并不会带来额外的通信代价,使用本地化差分保护用户梯度甚至会减小通信代价;而安全多方计算技术则存在参与方信息不对称、通信次数较多、通信开销庞大的特点.

现有的大多数隐私保护技术都会损失一部分模型可用性:

(1) 基于差分隐私的防御方法在原始数据或模型上添加了扰动,因此大多数情况下,隐私性越强,可用性越差.不过从模型泛化的角度看,模型的准确性和隐私性之间的取舍并不绝对,提高模型对单条数据的泛化能力不仅能解决模型的过拟合问题,提升模型在测试集上的表现,还能降低受到成员推断攻击和模型倒推攻击的风险;
(2) 基于同态加密的防御法方法中,近似或者扰动也会降低可用性,比如为了便于密文运算,用多项式近似非线性函数;为了便于构造同态特性,基于理想格的全同态加密方案在加密时添加了噪声;
(3) 安全多方计算中同样存在密文上的近似问题,为了完成小数计算,可能会采用数值截断的方法.

为了平衡三者之间相互制约的矛盾,可以从以下 4 个方面开展工作.

(1) 建立隐私机制评估体系,从隐私性、高效性、可用性对机器学习中的隐私保护机制进行多维度评价,为三者之间的权衡提供客观全面的度量;
(2) 根据不同应用场景的需求进行自适应的动态调整,以参数形式刻画影响某一方面的变量,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;
(3) 根据需求改进过于严格的隐私理论,提出更贴近实际可用性的可缩放的隐私理论;
(4) 加深学习理论的研究,寻找三方面之间的受益共同点(比如避免模型过拟合),共同提升隐私算法在多

方面的特性.

5.2 实现个性化度量与按需保护

隐私保护必将损失部分模型可用性和效率,对于用户个体而言这意味着牺牲部分服务质量以换取隐私保护.然而机器学习训练集中隐私保护需求存在巨大差异,对所有用户的所有数据统一进行隐私保护是不合理的,实现个性化度量和按需保护是十分关键的问题.

这种隐私需求的差异表现在横向和纵向两个方面,假设一个有 n 条数据的训练集 D,其中每条数据都有 d 维特征,第 i 条数据$(x_{i,1},\ldots,x_{1,d},y_i)$对应数据拥有者 u_i.

(1) 横向隐私需求差异性:训练集中不同数据拥有者具有不同的隐私保护需求,例如用户 u_i 愿意公开自己的信息,用户 u_j 不愿意公开自己的信息;

(2) 纵向隐私需求差异性:用户 u_i 可能想要保护的是数据$(x_{i,1},\ldots,x_{1,d},y_i)$在 D 中的成员属性,或者想要保护的是其中某一维的敏感属性 s 对应的值 $x_{i,s}$,或者$(x_{i,1},\ldots,x_{1,d})$是用户 u_i 个人照片的像素向量,u_i 不希望攻击者得到对整个图片信息的近似.

然而这种差异为机器学习中的隐私保护带来了巨大挑战:横向隐私需求差异要求机器学习在保护不同用户个性化隐私的同时提供具有差异性的服务质量,纵向隐私需求差异要求机器学习在保护不同属性个性化隐私的同时保证模型的可用性.

未来的研究可以从如下方面开展.

(1) 隐私预算的个性化分配.基于差分隐私的隐私保护机制中,可以用隐私预算度量隐私保护强度,大多数现有方法面对隐私预算分配问题时仅仅采用平均分摊的方式,因而横向或纵向地个性化分配差分隐私中的隐私预算是未来的方向;

(2) 不同隐私保护数据之间的相互补充.对于存在横向隐私需求差异的机器学习场景,可以使用隐私保护强度小的数据对隐私保护强度大的数据进行扩充,从而在保护个性化隐私的同时最大化模型可用性的极限;

(3) 隐私需求的动态感知.用户隐私需求随着机器学习模型的应用场景等因素发生动态变化,而用户不可能定期设置的自己的隐私需求,因此可以使用机器学习作为工具挖掘用户隐私行为和数据之间的关联,动态地预测用户的隐私需求.

5.3 控制跨数据源训练的共享与交换

隐私保护研究的出发点是使用更可靠的隐私保护技术为用户数据保驾护航,进而消除数据孤岛,为机器学习技术真正的落地和发展做充分的准备.本文第 4 节中提及的基于本地策略的隐私保护方案就是一种既能保护隐私又能满足多方数据共享这一现实需求的方法,因此,多方数据共享、协同训练是未来机器学习隐私保护的必然趋势,控制跨数据源训练的共享与交换,成为了机器学习隐私保护的重要挑战.

但是目前的相关研究都基于一些非常理想的假设:标签集成机制中,本地模型的结构可以不同,但是本地训练数据的形式必须是一致的;普通的联合学习机制中要求本地模型和全局模型的结构和输入数据形式都是一致的.然而实际应用中,参与方的数据结构不一定和总体模型要求的一致,如果从特征、标签、用户 id 这 3 个维度构建机器学习训练数据,数据的多样性存在于以下几种情况:

(1) 水平分割.比如不同城市的两家银行,他们的业务类型类似,也就是两方数据的特征和标签一致,但是用户 id 不一致;

(2) 垂直分割.比如同一个城市的一家医院和一家保险公司,两方数据的特征仅有少部分重合,标签也不一致,但是用户 id 基本一致,因为该城市的居民通常既去过这家医院也在这家保险公司购买服务;

(3) 混合分割,即不同城市的不同类型组织之间的数据,特征、标签和用户 id 均不一致的情况.

另外,多个参与方本地的模型也存在多样性,比如银行 A 基于本地数据训练的是信贷风险预测模型,银行 B 基于本地数据训练的是投资受益模型.

因此,如何解决多方跨数据源训练中数据形式多样性和模型多样性成为了关键问题.未来的工作可以从以下方面开展:(1) 使用迁移学习[96]解决数据源多样性问题,在保证隐私安全的前提下,将基于一方数据训练的模型A迁移到训练数据与其部分重合的模型B上;(2) 使用多任务学习解决模型多样性问题,多个参与方各自在本地数据上训练得到本地模型,中心服务器负责利用多个单任务模型得到多任务模型,为多个模型提供安全隐私的共享平台.

5.4 解决图结构数据训练集的隐私保护

目前,很多数据都可以用图结构的数据存储和分析,例如由人和人之间的关系构成社交网络、生物领域蛋白质之间的互相作用网络、用户与商品构成的推荐系统网络等.近年来,图神经网络的提出加速了图结构数据上的机器学习进展,并在文本分类、序列标注、关系抽取、事件抽取、视觉推理等方面得以广泛应用.已有工作[97]对社交网络这种图结构数据在发布中的隐私保护进行了总结,然而对图神经网络训练集的隐私泄露和隐私保护还在探索中.

现有机器学习中的隐私保护机制主要是围绕以关系型数据为训练集的模型,然而此类保护关系型数据训练集的隐私保护机制不能为图结构训练集提供隐私保护,原因如下:

(1) 新的背景知识.现有保护关系型数据训练集的隐私防御仅考虑了成员信息、属性信息作为攻击者的背景知识,而图结构数据中隐私攻击者可能掌握图结构、节点信息、边信息等背景知识,因此敌手模型更为复杂;

(2) 新的保护对象.图结构数据中,新的隐私保护对象包括节点隐私、边隐私和图性质隐私.其中,节点隐私还可以细分为存在性、再识别性、属性值、图结构,边隐私还可以细分为存在性、再识别性、权重、属性值.现有成员推断攻击将会攻击其存在性,模型倒推攻击将会攻击其属性值等信息;

(3) 新的可用性挑战.图数据中,节点并不是独立存在的而是相互关联的,如果直接简单应用已有的机器学习隐私保护方法,例如基于差分隐私添加随机噪声,将会破坏复杂性和节点之间的关系,从而极大地损失图数据的可用性.

为了实现以图结构数据为训练集的机器学习模型中的隐私保护,未来的工作可以从以下方面开展:(1) 深入探究面向图结构数据隐私分析基础理论,构建适合图结构数据的动态隐私度量模型和形式化描述方法;(2) 改进已有防御策略和机制,本文提及的扰动策略、对抗策略、本地策略、泛化策略、近似策略都是当前机器学习中隐私保护方法的高度抽象,因此未来可以将此类策略迁移至图结构数据训练集的隐私保护中,并基于图数据隐私的特殊性质改进优化.

6 总 结

海量的数据、丰富的场景催生了机器学习技术的蓬勃发展,然而成员推断攻击、模型倒推攻击、参数提取攻击揭露了机器学习模型的隐私漏洞,为数据管理和人工智能标准的制定带来了巨大挑战.

本文在充分调研和深入分析的机制之上,描述了攻击场景、敌手模型等背景知识,细致地归纳分析了隐私攻击和隐私保护的最新研究,对机器学习中的隐私保护方法进行凝练和抽象,并指出了当前机器学习中的隐私保护存在的问题,探讨了未来的研究方向.总之,机器学习中的隐私保护要随着机器学习技术的进步而发展,现有的诸多关键问题和挑战仍需进一步研究.

References:

[1] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the Security & Privacy. 2017.

[2] Wang Z, Song M, Zhang Z, Song, Y, Wang Q, Qi H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In: Proc. of the IEEE INFOCOM 2019—IEEE Conf. on Computer Communications. IEEE, 2019. 2512–2520.

[3] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: Proc. of the Security & Privacy. 2019.

[4] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2017.

[5] https://gdpr-info.eu/

[6] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proc. of the 2014 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2014.

[7] https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html

[8] Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2017. 587−601.

[9] Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD. Can machine learning be secure? In: Proc. of the 2006 ACM Symp. on Information, Computer and Communications Security. ACM, 2006. 16−25.

[10] Hayes J, Melis L, Danezis G, De Cristofaro E. LOGAN: Evaluating privacy leakage of generative models using generative adversarial networks. arXiv preprint arXiv:1705.07663, 2017.

[11] 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 UNIX Security Symp. 2014..

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

[13] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing machine learning models via prediction APIs. In: Proc. of the USENIX Security Symp. 2016. 601−618.

[14] Ateniese 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. Int'l Journal of Security and Networks, 2015,10(3):137−150.

[15] Robert C. Machine Learning, A Probabilistic Perspective. 2014.

[16] Long Y, Bindschaedler V, Wang L, Bu D, Wang X, Tang H, Chen K. Understanding membership inferences on well-generalized learning models. arXiv preprint arXiv:1802.04889.

[17] Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. In: Proc. of the 2018 IEEE 31st Computer Security Foundations Symp. 2018. 268−282.

[18] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Proc. of the Theory of Cryptography Conf. Springer, Berlin, Heidelberg, 2006. 265−284.

[19] Duchi JC, Jordan MI, Wainwright MJ. Local privacy and statistical minimax rates. Annual IEEE Symp. on Foundations of Computer Science, 2013. 429−438.

[20] Nikolov A, Talwar K, Zhang L. The geometry of differential privacy: The sparse and approximate cases. In: Proc. of the 45th Annual ACM Symp. on Theory of Computing. ACM, 2013. 351−360.

[21] McSherry F, Talwar K. Mechanism design via differential privacy. In: Proc. of the IEEE Symp. on Foundations of Computer Science. 2007. 94−103.

[22] Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965,60(309):63−69.

[23] 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 with English abstract). http://www.jos.org.cn/1000-9825/5364.htm [doi: 10.13328/j.cnki.jos.005364]

[24] Dwork C, Smith A, Steinke T, Ullman J. Exposed! A survey of attacks on private data. Annual Review of Statistics and Its Application, 2017,4:61−84.

[25] Acs G, Melis L, Castelluccia C, De Cristofaro E. Differentially private mixture of generative neural networks. IEEE Trans. on Knowledge and Data Engineering, 2018,31(6):1109−1121.

[26] Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739, 2018.

[27] Bindschaedler V, Shokri R. Synthesizing plausible privacy-preserving location traces. In: Proc. of the 2016 IEEE Symp. on Security and Privacy (SP). IEEE, 2016. 546−563.

[28] Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. Journal of Machine Learning Research, 2011,12:1069−1109.

[29] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2016. 308−318.

[30] Wang N, Xiao X, Yang Y, Zhao J, Hui SC, Shin H, Yu G. Collecting and analyzing multidimensional data with local differential privacy. In: Proc. of the 2019 IEEE 35th Int'l Conf. on Data Engineering (ICDE). IEEE, 2019. 638−649.

[31] McMahan HB, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. arXiv preprint arXiv: 1710.06963, 2017.

[32] Shokri R, Shmatikov V. Privacy-Preserving deep learning. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2015. 1310−1321.

[33] Wang Y, Si C, Wu X. Regression model fitting under differential privacy and model inversion attack. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015.

[34] 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 2017 ACM Int'l Conf. on Management of Data. ACM, 2017. 1307−1322.

[35] Zhang J, Zhang Z, Xiao X, Yang Y, Winslett M. Functional mechanism: Regression analysis under differential privacy. Proc. of the VLDB Endowment, 2012,5(11):1364−1375.

[36] Jagannathan G, Pillaipakkamnatt K, Wright RN. A practical differentially private random decision tree classifier. In: Proc. of the 2009 IEEE Int'l Conf. on Data Mining Workshops. IEEE, 2009. 114−121.

[37] Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-Supervised knowledge transfer for deep learning from private training data. In: Proc. of the 6th Int'l Conf. on Learning Representations. 2017.

[38] Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2018. 634−646.

[39] Rivest RL, Adleman L, Dertouzos ML. On data banks and privacy homomorphisms. Foundations of Secure Computation, 1978, 4(11):169−180.

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

[41] Paillier P. Public-Key cryptosystems based on composite degree residuosity classes. In: Proc. of the Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Berlin, Heidelberg: Springer-Verlag, 1999. 223−238.

[42] Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 19−38.

[43] Orlandi C, Piva A, Barni M. Oblivious neural network computing via homomorphic encryption. EURASIP Journal on Information Security, 2007(1):037343.

[44] Van Dijk M, Gentry C, Halevi S, Vaikuntanathan V. Fully homomorphic encryption over the integers. In: Proc. of the Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Berlin, Heidelberg: Springer-Verlag, 2010. 24−43.

[45] Brakerski Z, Vaikuntanathan V. Fully homomorphic encryption from ring-LWE and security for key dependent messages. In: Proc. of the Annual Cryptology Conf. Berlin, Heidelberg: Springer-Verlag, 2011. 505−524.

[46] Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping. ACM Trans. on Computation Theory, 2014,6(3):13.

[47] Hesamifard E, Takabi H, Ghasemi M, Jones C. Privacy-Preserving machine learning in cloud. In: Proc. of the 2017 on Cloud Computing Security Workshop. ACM, 2017. 39−43.

[48] Goldreich O, Warning A. Secure multi-party computation. In: Proc. of the Information Security & Communications Privacy. 2014.

[49] Shamir A. How to share a secret. Communications of the ACM, 1979,22(11):612−613.

[50] Fousse L, Lafourcade P, Alnuaimi M. Benaloh's dense probabilistic encryption revisited. In: Proc. of the Int'l Conf. on Cryptology in Africa. Berlin, Heidelberg: Springer-Verlag, 2011. 348−362.

[51] Goldwasser S, Micali S, Rackoff C. The knowledge complexity of interactive proof systems. SIAM Journal on Computing, 1989, 18(1):186−208.

[52] Bonawitz K, Ivanov V, Kreuter B, *et al*. Practical secure aggregation for privacy-preserving machine learning. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2017. 1175−1191.

[53] 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 2013 IEEE Symp. on Security and Privacy. IEEE, 2013. 334−348.

[54] Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Privacy-Preserving distributed linear regression on high-dimensional data. Proc. on Privacy Enhancing Technologies, 2017,2017(4):345−364.

[55] Meng X, Wang S, Shu K, Li J, Chen B, Liu H, Zhang Y. Personalized privacy-preserving social recommendation. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018.

[56] Bassily R, Smith A, Thakurta A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In: Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science. IEEE, 2014. 464−473.

[57] Li H, Xiong L, Jiang X. Differentially private synthesization of multi-dimensional data using copula function. In: Proc. of the Advances in Database Technology—Int'l Conf. on Extending Database Technology. NIH Public Access, 2014. 475.

[58] Bindschaedler V, Shokri R, Gunter CA. Plausible deniability for privacy-preserving data synthesis. Proc. of the VLDB Endowment, 2017,10(5):481−492.

[59] Huang C, Kairouz P, Chen X, Sankar L, Rajagopal R. Generative adversarial privacy. In: Proc. of the ACM, ICML Privacy in Machine Learning and Artificial Intelligence Workshop. 2018.

[60] Jia J, Gong NZ. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In: Proc. of the 27th USENIX Security Symp. 2018. 513−529.

[61] Hamm J, Cao Y, Belkin M. Learning privately from multiparty data. In: Proc. of the Int'l Conf. on Machine Learning. 2016. 555−563.

[62] Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson U. Scalable private learning with PATE. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver, 2018.

[63] Zheng W, Popa RA, Gonzalez JE, Stoica I. Helen: Maliciously secure coopetitive learning for linear models. In: Proc. of the 2019 IEEE Symp. on Security and Privacy (SP). IEEE, 2019.

[64] Geyer RC, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557, 2017.

[65] Chaudhuri K, Monteleoni C. Privacy-Preserving logistic regression. In: Proc. of the Advances in Neural Information Processing Systems. 2009. 289−296.

[66] Mcsherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2009. 19−30.

[67] Mir DJ, Wright RN. A differentially private graph estimator. In: Proc. of the 2009 IEEE Int'l Conf. on Data Mining Workshops. IEEE, 2009. 122−129.

[68] Sala A, Zhao X, Wilson C, Zheng H, Zhao BY. Sharing graphs using differentially private graph models. In: Proc. of the 2011 ACM SIGCOMM Conf. on Internet Measurement Conf. ACM, 2011. 81−98.

[69] Chaudhuri K, Sarwate AD, Sinha K. A near-optimal algorithm for differentially-private principal components. The Journal of Machine Learning Research, 2013,14(1):2905−2943.

[70] Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: Theory meets practice on the map. In: Proc. of the 2008 IEEE 24th Int'l Conf. on Data Engineering. 2008. 277−286.

[71] Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, Greene CS. Privacy-Preserving generative deep neural networks support clinical data sharing. Cardiovascular Quality and Outcomes, 2019,12(7):e005122.

[72] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: Proc. of the 34th Int'l Conf. on Machine Learning—Vol.70. JMLR.org, 2017. 2642−2651.

[73] Reiter JP, Mitra R. Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality, 2009,1(1).

[74] Dowlin N, Gilad-Bachrach R, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: Proc. of the Int'l Conf. on Machine Learning. 2016. 201−210.

[75] Wu S, Teruya T, Kawamoto J. Privacy-Preservation for stochastic gradient descent application to secure logistic regression. In: Proc. of the 27th Annual Conf. of the Japanese Society for Artificial Intelligence. 3L1-OS-06a-3. 2013.

[76] AlvimM S, Chatzikokolakis K, Kawamoto Y, Palamidessi C. Information leakage games. In: Proc. of the Int'l Conf. on Decision and Game Theory for Security. Cham: Springer-Valag, 2017. 437−457.

[77] Manshaei MH, Zhu Q, Alpcan T, Bacşar T, Hubaux JP. Game theory meets network security and privacy. ACM Computing Surveys (CSUR), 2013,45(3):25.

[78] Shokri R. Privacy games: Optimal user-centric data obfuscation. Proc. of the Privacy Enhancing Technologies, 2015,2015(2): 299−315.

[79] Shokri R, Theodorakopoulos G, Troncoso C, Hubaux JP, Le Boudec JY. Protecting location privacy: Optimal strategy against localization attacks. In: Proc. of the 2012 ACM Conf. on Computer and Communications Security. ACM, 2012. 617−627.

[80] Weinsberg U, Bhagat S, Ioannidis S, Taft N. BlurMe: Inferring and obfuscating user gender based on ratings. In: Proc. of the 6th ACM Conf. on Recommender Systems. ACM, 2012. 195−202.

[81] Heatherly R, Kantarcioglu M, Thuraisingham B. Preventing private information inference attacks on social networks. IEEE Trans. on Knowledge and Data Engineering, 2012,25(8):1849−1862.

[82] Chen T, Boreli R, Kaafar MA, Friedman A. On the effectiveness of obfuscation techniques in online social networks. In: Proc. of the Int'l Symp. on Privacy Enhancing Technologies Symp. Cham: Springer-Verlag, 2014. 42−62.

[83] Salamatian S, Zhang A, du Pin Calmon F, Bhamidipati S, Fawaz N, Kveton B, Taft N. Managing your private and public data: Bringing down inference attacks against your privacy. IEEE Journal of Selected Topics in Signal Processing, 2015,9(7):1240−1255.

[84] Breiman L. Bagging predictors. Machine Learning, 1996,24(2):123−140.

[85] Mironov I. Rényi differential privacy. In: Proc. of the 2017 IEEE 30th Computer Security Foundations Symp. IEEE, 2017. 263−275.

[86] Lindell Y, Pinkas B. Privacy preserving data mining. In: Bellare M, ed. Proc. of the Advances in Cryptology—CRYPTO 2000. LNCS 1880, Berlin, Heidelberg: Springer-Verlag, 2000.

[87] Bunn P, Ostrovsky R. Secure two-party *k*-means clustering. In: Proc. of the 14th ACM Conf. on Computer and Communications Security. ACM, 2007. 486−497.

[88] Vaidya J, Yu H, Jiang X. Privacy-Preserving SVM classification. Knowledge and Information Systems, 2008,14(2):161−178.

[89] Sanil AP, Karr AF, Lin X, Reiter JP. Privacy preserving regression modelling via distributed computation. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2004. 677−682.

[90] Huai M, Huang L, Wei Y, Lu L, Qi M. Privacy-Preserving naive bayes classification. In: Proc. of the Int'l Conf. on Knowledge Science. 2015.

[91] Slavkovic AB, Nardi Y, Tibbits MM. Secure logistic regression of horizontally and vertically partitioned distributed databases. In: Proc. of the 7th IEEE Int'l Conf. on Data Mining Workshops (ICDMW 2007). IEEE, 2007. 723−728.

[92] Nikolaenko V, Ioannidis S, Weinsberg U, Joye M, Taft N, Boneh D. Privacy-Preserving matrix factorization. In: Proc. of the 2013 ACM SIGSAC Conf. on Computer & Communications Security. ACM, 2013. 801−812.

[93] Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Secure linear regression on vertically partitioned datasets. IACR Cryptology ePrint Archive, 2016. 892.

[94] Mcmahan HB, Moore E, Ramage D, Arcas BAY. Federated learning of deep networks using model averaging. arXiv:1602.05629 2016.

[95] Phong LT, Aono Y, Hayashi T, Wang L, Moriai S. Privacy-Preserving deep learning via additively homomorphic encryption. IEEE Trans. on Information Forensics and Security, 2018,13(5):1333−1345.

[96] Liu Y, Chen T, Yang Q. Secure federated transfer learning. arXiv preprint arXiv:1812.03337, 2018.

[97] Liu XY, Wang B, Yang XC. Survey on privacy preserving techniques for publishing social network data. Ruan Jian Xue Bao/ Journal of Software, 2014,25(3):576−590 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

附中文参考文献:

[23] 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]

[97] 刘向宇,王斌,杨晓春.社会网络数据发布隐私保护技术综述.软件学报,2014,25(3):576−590. http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

刘睿瑄(1995－),女,主要研究领域为隐私保护,机器学习,数据挖掘.

陈红(1965－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库技术,新硬件平台下的高性能计算.

郭若杨(1994－),女,CCF 学生会员,主要研究领域为应用性数据安全协议包括安全计算,加密查询.

赵丹(1988－),男,硕士,主要研究领域为本地化差分隐私,数据发布.

梁文娟(1980－),女,硕士,主要研究领域为数据隐私保护,数据挖掘,物联网数据管理.

李翠平(1971－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为社交网络分析,社会推荐,大数据分析及挖掘.