计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190455
57(2):346-362,2020

机器学习的隐私保护研究综述

刘俊旭　　孟小峰
(中国人民大学信息学院　北京　100872)
(junxu_liu@ruc.edu.cn)

Survey on Privacy-Preserving Machine Learning

Liu Junxu and Meng Xiaofeng
(*College of Information*, *Renmin University of China*, *Beijing* 100872)

Abstract　Large-scale data collection has vastly improved the performance of machine learning, and achieved a win-win situation for both economic and social benefits, while personal privacy preservation is facing new and greater risks and crises. In this paper, we summarize the privacy issues in machine learning and the existing work on privacy-preserving machine learning. We respectively discuss two settings of the model training process—centralized learning and federated learning. The former needs to collect all the user data before training. Although this setting is easy to deploy, it still exists enormous privacy and security hidden troubles. The latter achieves that massive devices can collaborate to train a global model while keeping their data in local. As it is currently in the early stage of the study, it also has many problems to be solved. The existing work on privacy-preserving techniques can be concluded into two main clues—the encryption method including homomorphic encryption and secure multi-party computing and the perturbation method represented by differential privacy, each having its advantages and disadvantages. In this paper, we first focus on the design of differentially-private machine learning algorithm, especially under centralized setting, and discuss the differences between traditional machine learning models and deep learning models. Then, we summarize the problems existing in the current federated learning study. Finally, we propose the main challenges in the future work and point out the connection among privacy protection, model interpretation and data transparency.

Key words　privacy-preserving; differential privacy; machine learning; deep learning; federated learning

摘　要　大规模数据收集大幅提升了机器学习算法的性能,实现了经济效益和社会效益的共赢,但也令个人隐私保护面临更大的风险与挑战.机器学习的训练模式主要分为集中学习和联邦学习2类,前者在模型训练前需统一收集各方数据,尽管易于部署,却存在极大数据隐私与安全隐患;后者实现了将各方数据保留在本地的同时进行模型训练,但该方式目前正处于研究的起步阶段,无论在技术还是部署中仍面临诸多问题与挑战.现有的隐私保护技术研究大致分为2条主线,即以同态加密和安全多方计算为代表的加密方法和以差分隐私为代表的扰动方法,二者各有利弊.为综述当前机器学习的隐私问题,并对

收稿日期:2019-06-21;**修回日期**:2019-09-11
基金项目:国家自然科学基金项目(91646203,61532010,91846204,61532016,61762082);国家重点研发计划项目(2016YFB1000602,2016YFB1000603)
This work was supported by the National Natural Science Foundation of China (91646203, 61532010, 91846204, 61532016, 61762082) and the National Key Research and Development Program of China (2016YFB1000602, 2016YFB1000603).
通信作者:孟小峰(xfmeng@ruc.edu.cn)

现有隐私保护研究工作进行梳理和总结,首先分别针对传统机器学习和深度学习 2 类情况,探讨集中学习下差分隐私保护的算法设计;之后概述联邦学习中存在的隐私问题及保护方法;最后总结目前隐私保护中面临的主要挑战,并着重指出隐私保护与模型可解释性研究、数据透明之间的问题与联系.

关键词 隐私保护;差分隐私;机器学习;深度学习;联邦学习

中图法分类号 TP391

在互联网、大数据和机器学习的助推下,人工智能技术日新月异,刷脸支付、辅助诊断、个性化服务等逐步走入大众视野并深刻改变着人类的生产与生活方式.然而,在这些外表光鲜的智能产品背后,用户的生理特征、医疗记录、社交网络等大量个人敏感数据无时无刻不在被各类企业、机构肆意收集.大规模数据收集能够带动机器学习性能的提升,实现经济效益和社会效益的共赢,但也令个人隐私保护面临更大的风险与挑战.主要表现在 2 方面:首先,由不可靠的数据收集者导致的数据泄露事件频发,不仅对企业造成重大经济和信誉损失,也对社会稳定和国家安全构成极大威胁;其次,大量研究表明,攻击者通过分析机器学习模型的输出结果,能够逆向推理出训练数据中个体的敏感信息.2018 年剑桥分析公司"操纵"美国大选事件,便是通过非法获取 8 700 万 Facebook 用户数据,构建心理分析模型,分析互联网用户人格特征,进而定向投放虚假广告实施的①.

人工智能时代,个人隐私保护愈发受到国内外的重视和关注.2017 年 6 月起施行的《中华人民共和国网络安全法》②第 42 条指出,"网络运营者不得泄露、篡改、毁损其收集的个人信息;未经被收集者同意,不得向他人提供个人信息".2018 年 3 月,欧盟通用数据保护条例(General Data Protection Regulation, GDPR)③正式生效,该条例对企业处理用户数据的行为提出明确要求.可见,企业在用户不知情时进行数据收集、共享与分析已被视为一种违法行为.

实现隐私保护的机器学习,除借助法律法规的约束外,还要求研究者必须以隐私保护为首要前提进行模型的设计、训练与部署,保证数据中的个人敏感信息不会被未授权人员直接或间接获取.

传统的机器学习训练中,各方数据首先被数据收集者集中收集,然后由数据分析者进行模型训练,此模式称为集中学习(centralized learning)[1].其中,数据收集者与数据分析者可以是同一方,如移动应用开发者;也可以是多方,如开发者将数据共享给其他数据分析机构.可见集中学习模式下,用户一旦被收集数据,便很难再拥有对数据的控制权,其数据将被用于何处、如何使用也不得而知.近年来,一部分研究者尝试令各方数据保留在本地的同时训练全局模型,此工作的典型代表为 2017 年 Google 提出的联邦学习(federated learning)[2].尽管联邦学习使用户拥有了个人数据的控制权,但并不能完全防御潜在的隐私攻击.

机器学习的隐私保护研究大致分为 2 条主线:以多方安全计算(secure multi-party computation, SMPC)[3]、同态加密(homomorphic encryption, HE)[4-5]为代表的加密方法和以差分隐私(differential privacy, DP)[6-7]为代表的扰动方法.加密方法既能将数据明文编码为仅特定人员能够解码的密文,保证存储和传输过程中数据的机密性,同时借助安全协议实现直接对密文计算并求得正确结果.然而,数据加密过程往往涉及大量计算,复杂情况下将产生巨大的性能开销,故在实际应用场景中难以落地.差分隐私是一种建立在严格数学理论基础之上的隐私定义,旨在保证攻击者无法根据输出差异推测个体的敏感信息.与加密相比,差分隐私仅通过噪声添加机制[8]便可以实现,故不存在额外的计算开销,但一定程度上会对模型的预测准确性造成影响.该方法面临的主要挑战是设计合理的扰动机制,从而更好地权衡算法隐私与可用性.

迄今为止,已有大量研究工作致力于集中学习模式下的隐私保护,本文重点介绍差分隐私方法,分别讨论传统机器学习和深度学习 2 类模型的隐私算法设计.传统机器学习模型结构简单,其训练本质上是一个凸(convex)优化问题,可以通过在经验风险最小化(empirical risk minimization, ERM)的不同阶段添加扰动的方式实现差分隐私保护.深度学习

① https://en.wikipedia.org/wiki/Cambridge_Analytica

② http://www.cac.gov.cn/2016-11/07/c_1119867116.htm

③ https://eugdpr.org/

模型的训练比传统机器学习更加复杂,其迭代过程需要频繁访问训练数据,故而更难权衡隐私与可用性.解决此问题的方法之一是制定宽松的差分隐私定义,适当降低隐私保护要求,但同时模型受到隐私攻击的概率更大了.

联邦学习模式下的隐私保护同样存在加密和扰动2种方法.区块链(blockchain)技术因其去中心化、安全透明、不可篡改等特点,能够对计算过程进行审计,监控模型训练中的恶意行为,从而加强隐私保护效果[9].不过,联邦学习目前正处于研究的起步阶段,无论在技术还是部署中仍面临诸多问题与挑战,如通信带宽受限、收敛速度慢等.

图1从模型训练模式、隐私保护技术和模型类型3个维度对现有机器学习的隐私保护研究进行划分,颜色深浅代表相应内容在本文中所占的比例.本文详细总结了针对集中学习模式下传统机器学习模型与深度学习模型的隐私保护方法,重点介绍差分隐私保护技术.同时,本文简要概述了联邦学习模式下存在的隐私问题与现有保护技术.最后,本文针对现有研究中存在的主要问题,提出未来的主要研究挑战.

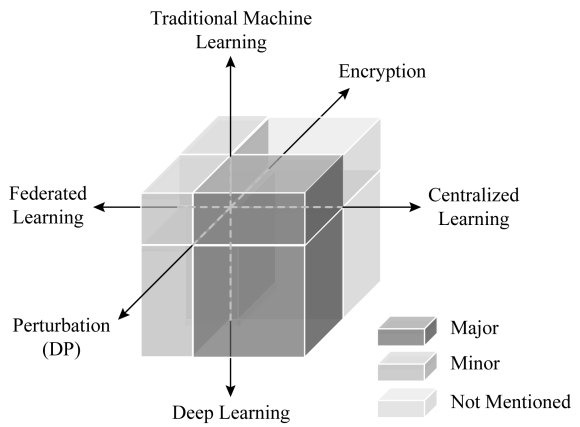


Fig. 1 Overview of privacy-preserving machine learning studies

图1 机器学习的隐私保护研究概况

1 机器学习的隐私问题

数据科学的发展必然伴随着隐私问题.对机器学习而言,隐私问题主要表现在以下2个方面.

1) 由大规模数据收集导致的直接隐私泄露[10].主要表现在不可靠的数据收集者在未经人们许可的情况下擅自收集个人信息、非法进行数据共享和交易等.

2) 由模型泛化能力不足导致的间接隐私泄露[11].主要表现在不可靠的数据分析者通过与模型进行交互,从而逆向推理出未知训练数据中的个体敏感属性.此类问题产生的根源在于越复杂的模型在训练中往往具有更强大的数据"记忆"能力,以致模型对训练数据和非训练数据的表现存在较大差异.

本节重点讨论间接隐私泄露问题,具体指针对机器学习模型发起的各类隐私攻击.隐私攻击大多发生在模型应用阶段,由于攻击者无法直接访问训练数据,故只能对相关信息进行推断.攻击者可能对模型和数据一无所知;也可能具有一定的背景知识,如已知模型类型或数据特征.根据攻击者的攻击目标,隐私攻击可分为重构攻击(reconstruction attack)[12]和成员推断攻击(membership inference attack)[13]2类.

除隐私问题外,机器学习同样面临诸多安全威胁.安全问题与隐私问题的主要区别在于:前者尽管造成了训练数据的直接或间接泄露,模型本身却并未受到影响;但后者将会导致模型的内在逻辑被恶意诱导或破坏,从而无法实现预期功能.针对机器学习的安全攻击既有可能发生在模型训练阶段,也可能发生在模型应用阶段,主要包括投毒攻击(poisoning attack)[14-19]和对抗样本攻击(adversarial examples attack)[20-27].安全问题也是现今机器学习所面临的一个挑战性问题,由于非本文重点,此处不再赘述.

已有学者针对机器学习模型的攻击技术进行了梳理和总结[28-32],本节主要对2类隐私攻击进行简要介绍.

1.1 重构攻击

重构攻击指攻击者试图重构出训练数据中特定个体的敏感信息或者目标模型,其中前者称为模型反演攻击(model inversion attack)[33-34],后者称为模型窃取攻击(model extraction attack)[35].

1.1.1 模型反演攻击

对于结构简单的机器学习模型,采用动态分析或计算记录间的相似度等方法便可推测出训练数据中个体的敏感信息,如文献[33]针对个性化用药线性预测模型,在已知特定病人的基本信息和预测结果的情况下,成功推测到该病人的敏感基因型.对于复杂的深度学习模型,文献[34]在样本标签等辅助信息的基础上,利用人脸识别系统的预测置信度对随机生成的"模拟画像"不断修正,成功重构出训练集中个体的真实样貌.然而文献[35]指出,数据重构

仅在训练样本量很小的情况下才能实现,当样本量很大时,攻击效果将大大减弱.

1.1.2 模型窃取攻击

早期的模型窃取攻击主要利用等式求解的方法,仅适用于简单的线性二分类模型[36].文献[35]将该方法应用到非线性支持向量机、深度神经网络等复杂模型中,并利用预测置信度使攻击效果得到了明显提升.除此之外,他们还提出一种针对决策树模型的自适应攻击算法.尽管表面上模型窃取攻击并不以数据为目标,但文献[13]指出,由于模型在训练中可能"记住"某些训练数据,因此基于窃取到的替代模型进行模型反演攻击能够明显提升攻击效果.在实际应用场景下,机器学习模型对企业而言是重要的知识产权,一旦被窃取,也将为企业带来极大的损失.

1.2 成员推断攻击

成员推断攻击指攻击者试图推测1个给定样本是否被用于模型的训练中,即训练数据的"成员"之一.在某些场景下,成员推断攻击可能造成严重的后果,比如对于由艾滋病患者数据构建的诊断模型,若某人的医疗数据被推断是该模型的训练数据,便意味着此人可能患有艾滋病.

文献[13]首次提出成员推断攻击,并假设攻击者只能在"黑盒"模式下访问目标模型,利用模拟数据构建目标模型的影子模型(shadow model),并基于影子模型和目标模型的输出结果训练1个能够判断是否是目标模型训练数据的攻击模型.构建影子模型需满足2个假设条件:1)用来训练影子模型的模拟数据应与真实训练数据具有相似的分布;2)其结构应与目标模型一致.文献[37]放宽了上述约束条件,在保证攻击有效性的情况下提出了一种更通用的攻击模型.文献[11]考虑了"白盒"模式下的攻击,即假设攻击者已知模型在训练集上的平均损失,通过评估模型关于某条数据的损失是否超过训练平均损失以判断该数据是否是训练数据.此外,还有研究工作提出了针对生成模型[38]以及差分隐私保护下的深度学习模型[39]的成员推断攻击.

可见,目前针对机器学习的隐私攻击具有明显的局限性,仅在特定条件和假设下才能成功.但人们依旧不能忽视这些问题,随着研究的逐步深入,这些攻击将会威胁到更多更复杂的模型.解决机器学习的隐私问题,一方面需借助法律和社会道德的制裁和约束,规范对个人数据的收集、处理和传播行为,防止直接隐私泄露;另一方面,研究者还需在模型设计之初便尽可能考虑到训练与应用过程中的潜在隐患,通过优化模型结构和学习算法,或借助数据加密、噪声干扰等隐私保护技术,从而防御一切可能的间接隐私泄露.

2 机器学习的隐私保护

第1节指出,机器学习的隐私问题一方面缘于大规模数据收集,另一方面缘于模型本身会携带训练数据中个体的信息.基于此,机器学习的隐私保护存在2个主要研究思路:第一,改变数据集中收集的训练模式;第二,设计隐私保护方法,使模型训练过程实现隐私保护.

由于机器学习模型包括传统机器学习和深度学习2类,二者在模型结构和复杂程度上具有明显差异,故需分别讨论.本节主要从模型训练模式和隐私保护技术2个维度对机器学习的隐私保护研究整体概况进行梳理和总结.针对上述2类机器学习模型的具体工作将在后续章节加以说明.

2.1 模型训练模式

训练模式可分为集中学习和联邦学习2类,区别在于各方数据在模型训练前是否被集中收集.

2.1.1 集中学习

对执行机器学习任务的互联网服务提供商而言,将训练数据集中存储在单机、集群或云端对于模型训练和部署都方便可控,因此广泛应用于实际场景.但该模式下,各方一旦被收集数据,便很难再拥有对数据的控制权,其数据将被用于何处、如何使用也不得而知.针对集中学习模式下机器学习的隐私保护在过去几十年间得到了广泛研究,本文将在第3,4节分别对传统机器学习和深度学习2种情况加以讨论.

2.1.2 联邦学习

联邦学习与分布式机器学习中的数据并行化训练具有相似的逻辑结构,即拥有不同训练数据的多个节点共同执行一个机器学习任务.其中,各个节点在获得中心模型的副本后独立训练,并将训练后更新的模型参数上传至中心节点;中心节点将所有上传的参数整合至中心模型,并再次将模型分发出去;如此迭代,直至中心模型收敛.联邦学习与数据并行化训练的主要区别在于,前者的主要目的是让各节点的数据保留在本地,以降低隐私泄露的风险;而后者则是加速模型训练,各节点中的数据仍是中心节点先集中收集后再均匀分配的.不过,联邦学习尚

处于研究的起步阶段,在算法设计与实际部署上面临种种问题及挑战,本文将在第 5 节加以讨论.

2.2 隐私保护技术

针对机器学习的隐私保护主要通过加密或扰动 2 种方式.前者主要指密码学技术,常用的有安全多方计算、同态加密等;后者则主要指差分隐私机制.

1) 加密

加密被认为是最基本、最核心的数据安全技术,通过加密算法将数据明文编码为仅特定人员能够解码的密文,旨在保证敏感数据在存储与传输过程中的保密性.对机器学习而言,由于恶意攻击者能够基于模型对数据加以推测,因此同样需保证数据在计算与分析过程中的机密性.

同态加密是一种不需要访问数据本身就可以处理数据的密码学技术[4].文献[5]进一步提出的全同态加密则实现了能够在加密数据上进行任意计算,目前已被广泛应用于云计算场景的隐私保护研究中.除此之外,安全多方计算作为一种让互不信任的参与方进行协同计算的协议,允许在不公开各方真实数据的同时保证计算结果的正确性[40-42].故该方法非常适合由多方参与、并共同训练机器学习模型的情况,如联邦学习.安全多方计算常与同态加密方法结合使用,以应对多种分析任务.

加密方法的优点在于能够保证计算结果的正确性,缺点是该方法十分依赖于函数的复杂度.对于存在大量的非线性计算的深度学习模型,算法的计算开销十分高昂,这也是加密方法至今在有效性和实用性方面饱受争议、无法在实际应用中落地的主要原因.

2) 扰动

扰动技术指在模型训练过程中引入随机性,即添加一定的随机噪声,使输出结果与真实结果具有一定程度的偏差,以防止攻击者恶意推理,差分隐私机制是目前扰动技术的代表性方法.差分隐私是 Dwork 等人[6]提出的一种具有严格的数学理论支撑的隐私定义,最早用以解决统计数据库在数据发布过程中的隐私泄露问题.满足差分隐私的算法,其输出结果的概率分布不会因增加、删除或修改数据集中的一条记录而产生明显的差异.这一定程度上避免了攻击者通过捕捉输出差异进而推测个体记录的敏感属性值.形式上,差分隐私的定义如下.

定义 1. ε-差分隐私.对任意的邻接数据集 D 和 D',$D,D'\in\mathcal{D}$.给定随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$和任意的输出结果 $S\subseteq\mathbb{R}$,若不等式

$$\max_S\left[\ln\frac{Pr[f(D)\in S]}{Pr[f(D')\in S]}\right]\leqslant\varepsilon \tag{1}$$

成立,则称算法 f 满足 ε-差分隐私.其中,邻接数据集(neighbor datasets)[7]指有且仅有 1 条记录不同的 2 个数据集;不等式左边可视为算法访问数据集后造成的隐私损失(privacy loss);ε 用于控制算法的隐私保护程度,称为隐私预算(privacy budget).差分隐私机制将算法的隐私损失控制在一个有限的范围内,ε 越小,则算法隐私保护效果越好.常用的有拉普拉斯机制(Laplace mech-anism)[6]、指数机制(exponential mechanism)[43]和高斯机制(Gaussian mechanism)[44].这些机制中,噪声大小取决于算法的敏感度(sensitivity)[44].

定义 2. 敏感度.对于函数 $f:\mathcal{X}\to\mathcal{Y}$,$\boldsymbol{x}_i\in\mathcal{X}$ 和 $\boldsymbol{r}\in\mathcal{X}$ 为特征向量.当且仅当输入数据集中任意 1 条数据改变时,其输出结果变化的最大值称为该函数的敏感度,形式化定义为

$$S(f)=\max_{\boldsymbol{r},\boldsymbol{r}'}|f(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1},\boldsymbol{r})-f(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1},\boldsymbol{r}')|. \tag{2}$$

差分隐私机制是目前机器学习的隐私保护研究中最常采用的方法之一.由于模型训练过程往往需要多次访问敏感数据集,如数据预处理、计算损失函数、梯度下降求解最优参数等,故必须将整个训练过程的全局隐私损失控制在尽可能小的范围内.对于简单模型,此要求较容易实现.然而,对结构复杂、参数量大的深度学习模型而言,将难以平衡模型可用性与隐私保护效果,这是该技术面临的最大问题与挑战.

2.3 差分隐私保护的机器学习

与加密方法相比,差分隐私机制更易于在实际场景中部署和应用,故本文重点讨论差分隐私保护下的机器学习算法设计.

根据数据处理与分析能力的不同,机器学习模型可分为以线性回归(linear regression)、逻辑回归(logistic regression)、支持向量机(support vector machine, SVM)等基于统计学习理论的传统机器学习方法,和以各类神经网络(neural network, NN)模型为代表的深度学习方法.对大多数模型而言,经验风险最小化是最常用的模型学习策略,其基本思想是在整个参数域中搜索使经验风险最小的最优模型参数.其形式化定义如下.

定义 3. 经验风险最小化.对模型 $f:\mathcal{W}\times\mathcal{X}\to\mathcal{Y}$,给定训练集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$,$D\subseteq\mathcal{D}$ 和损失函数 ℓ:$\mathcal{W}\times\mathcal{D}\to\mathbb{R}$,其中 $\boldsymbol{x}_i\in\mathcal{X}$ 为特征向量,$y\in\mathcal{Y}$ 为类别

标签,$w\in\mathcal{W}$ 为模型参数向量,$\Omega(w)$ 为用来防止模型过拟合的正则化项.则模型在数据集 D 上的经验风险[①]为

$$J(\boldsymbol{w};D)=\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))+\Omega(\boldsymbol{w}). \quad (3)$$

依据经验风险最小化策略,最优模型参数为

$$\boldsymbol{w}^*=\arg\min_{\boldsymbol{w}} J(\boldsymbol{w};D). \quad (4)$$

经验风险最小化求解最优模型参数的常用算法是基于迭代计算的梯度下降法(gradient descent, GD).传统机器学习模型由于结构简单,故在设计目标函数 $J(\boldsymbol{w};D)$ 时会尽可能令其为一个凸函数,以便求得一个确定的最优解.深度学习模型由于引入了大量非线性因素,目标函数常常是非凸(non-convex)函数,故求解时极易陷入局部最优解.此外,深度学习还存在参数量大、迭代次数多、算法收敛慢等问题.因此,上述 2 类模型的隐私保护方法具有较大的差异,在设计时需分别加以考虑.

经验风险最小化不满足差分隐私.对于传统机器学习,根据经验风险最小化得到的最优模型往往与决策边界附近的某些训练样本密切相关(如 SVM 中的支持向量).若这些样本的集合被增加、删除或修改,将会导致模型完全改变,使得式(1)中的比值将趋近无穷大.在这种情况下,训练样本的信息将很容易被推测出来,如图 2 所示.对深度学习而言,由于模型大多为非线性的,该问题将更为明显.

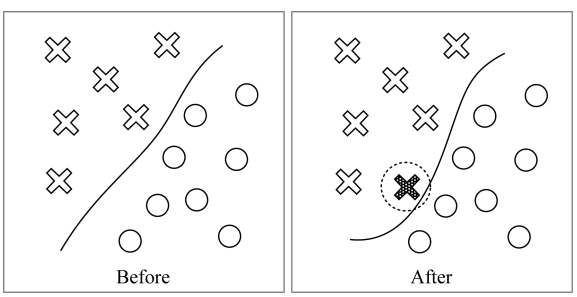


Fig. 2 ERM does not satisfy differential privacy[②]

图 2 经验风险最小化不满足差分隐私

综上,对绝大多数机器学习任务而言,若令经验风险最小化过程满足差分隐私,则模型一定程度上便实现了隐私保护[45].

3 传统机器学习的隐私保护

在深度学习出现之前,基于统计学习理论的传统机器学习模型是用来解决各类数据挖掘任务和简单学习任务主要方法.在机器学习的隐私保护早期研究中,学术界也对此进行了大量的探索.本节将以有监督学习任务为例,讨论差分隐私保护下的传统机器学习的方法及其存在的问题.

3.1 差分隐私保护的经验风险最小化

如图 3 所示,根据随机噪声在经验风险最小化过程添加的位置,本文将差分隐私保护的经验风险最小化方法总结为输入扰动、目标扰动、梯度扰动和输出扰动 4 种类型.

3.1.1 输入扰动

输入扰动(input perturbation)是指个人数据在交由模型学习或分析前,先对其进行一定程度的随机扰动,以避免模型获取真实数据.考虑 2 种情况:1)全局隐私(global privacy),即个人数据首先被集中收集,收集者发布数据时先要对敏感数据集进行扰动;2)本地隐私(local privacy),即个人首先在本地端对数据进行扰动,再将其发送给收集者[46].前者在早期研究中已证明存在较大的局限性[47];后者由于用户之间并不知道彼此的数据,故基于全局敏感度的扰动机制已不再适用.人们进一步提出了本地化差分隐私(local differential privacy, LDP)的定义,并针对不同数据类型[48]、各类数据挖掘任务[49-52]以及线性回归、逻辑回归等简单机器学习模型[46,53]进行了大量的尝试,且成为现今隐私保护技术研究的主流方法之一.文献[54-55]针对本地化差分隐私的研究现状进行了较为全面的总结,此处不再赘述.

3.1.2 输出扰动

输出扰动(output perturbation)是指直接对经验风险最小化得到的最优参数添加噪声,如式(5)所示:

$$\hat{\boldsymbol{w}}=\arg\min_{\boldsymbol{w}} J(\boldsymbol{w};D)+\boldsymbol{z}. \quad (5)$$

文献[45]指出,当 z 服从概率密度函数如式(6)所示的拉普拉斯分布时,经验风险最小化过程满足差分隐私.

① 很多文献在定义经验风险时并不考虑正则化项,并称引入正则化项的学习策略为结构风险最小化(structural risk minimization, SRM),或正则化的经验风险最小化(regularized empirical risk minimization).由于正则化项并不对最优参数求解过程造成影响,为表述简单,本文将上述 2 种形式统称为经验风险最小化,且除非特别说明,一律指正则化的经验风险最小化.

② Chaudhuri K, Sarwate A D. Differentially Private Machine Learning: Theory, Algorithms, and Applications(tutorial). https://www.ece.rutgers.edu/~asarw ate/nips2017/

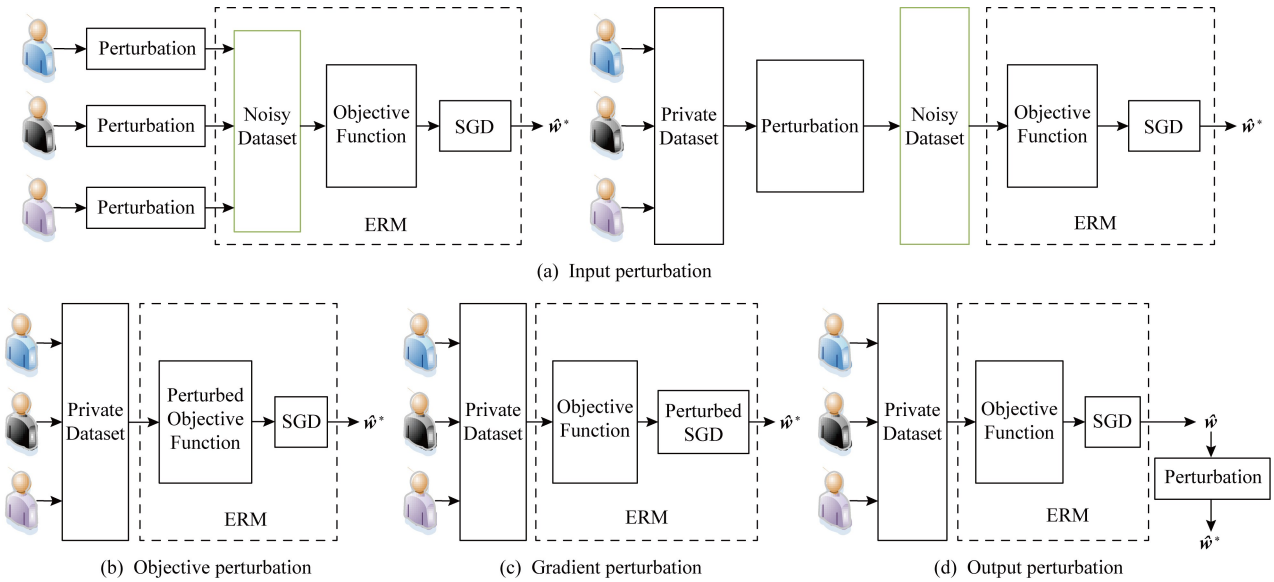


Fig. 3 Differentially private empirical risk minimization

图 3 差分隐私保护的经验风险最小化

$$\rho(\boldsymbol{z})\propto \mathrm{e}^{-\beta\|\boldsymbol{z}\|}, \tag{6}$$

其中,β 为与隐私预算 ε 和函数 $\arg\min_{\boldsymbol{w}} J(\boldsymbol{w};D)$ 的敏感度有关的表达式.

输出扰动直接对算法的真实输出结果添加噪声,故是最直观的一种扰动方式.由于噪声大小与函数 $\arg\min_{\boldsymbol{w}} J(\boldsymbol{w};D)$ 的敏感度密切相关,文献[45]指出,当经验风险最小化的目标函数 $J(\boldsymbol{w};D)$ 满足连续、可微且为凸函数时,能够求得函数 $\arg\min_{\boldsymbol{w}} J(\boldsymbol{w};D)$ 的敏感度,进而证明算法满足差分隐私.上述条件使得该方法存在较大的局限性,即当正则化项或损失函数为非凸函数时,该方法便不再适用.除此之外,由于所加噪声服从一定的概率分布,若攻击者重复执行相同的查询,仍然有可能根据噪声结果的分布情况推测算法输出的真实结果.

3.1.3 目标扰动

目标扰动(objective perturbation)是指向经验风险最小化的目标函数表达式中引入随机项,并保证求解过程满足差分隐私.扰动后的目标函数为

$$\hat{J}(\boldsymbol{w};D)=J(\boldsymbol{w};D)+\frac{1}{n}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{w}, \tag{7}$$

其中,随机变量 $\boldsymbol{z}$ 同样服从概率密度函数如式(6)所示的分布.注意,此时 β 为一仅与隐私预算 ε 有关的表达式,与目标函数的敏感度无关.

目标扰动同样要求目标函数 $J(\boldsymbol{w};D)$ 连续、可微且为凸函数,以证明其满足差分隐私[45],故而该方法同样具有极大的局限性.文献[56]提出一种多项式近似的方法,即利用泰勒展开式求解目标函数的近似多项式表达,并对各系数添加拉普拉斯噪声.尽管该方法被成功应用于逻辑回归模型中,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.

3.1.4 梯度扰动

梯度扰动(gradient perturbation)是指在利用梯度下降法求解最优模型参数的过程中引入随机噪声,并保证整个过程满足差分隐私.为保证算法的计算效率,实际应用中常采用随机梯度下降(stochastic gradient descent, SGD)或小批量梯度下降(mini-batch gradient descent, MBGD)方法,即每次迭代仅对单个或少量样本计算梯度.以 SGD 和 MBGD 为例,梯度扰动方法为[57]

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\eta_t(\nabla\Omega(\boldsymbol{w}_t)+\nabla\ell(\boldsymbol{w}_t,(\boldsymbol{x}_i,y_i))+\boldsymbol{z}_t), \tag{8}$$

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\eta_t\left(\nabla\Omega(\boldsymbol{w}_t)+\frac{1}{b}\sum_{(\boldsymbol{x}_i,y_i)\in B_t}\nabla\ell(\boldsymbol{w}_t,(\boldsymbol{x}_i,y_i))+\frac{1}{b}\boldsymbol{z}_t\right), \tag{9}$$

其中,η_t 为第 t 次迭代的学习率,B_t 为第 t 次迭代随机选取小批量样本,$\boldsymbol{z}_t$ 表示第 t 次迭代时添加的随机噪声且服从式(6)所示的概率分布.由于 SGD 和 MBGD 并不能保证算法有很好的收敛性,引入随机噪声后,此问题将更加严重.

下面以逻辑回归模型为例说明输出扰动、目标扰动和梯度扰动 3 种方式下添加噪声的差异.其中,经验风险最小化的目标函数为

$$J(\boldsymbol{w})=\frac{1}{n}\sum_{i=1}^{n}\ln(1+\mathrm{e}^{-y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i})+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2. \tag{10}$$

为保证优化问题的目标函数为凸函数,正则化项采用 L_2 正则项 (L_2-norm),λ 为正则化系数.

当 $\boldsymbol{x}\in\mathbb{R}^d$ 且 $\|\boldsymbol{x}\|_2\leqslant 1$ 时,若随机噪声 $\boldsymbol{z}$ 服从参数为 $\beta=n\varepsilon/2$ 时的概率分布(见式(6)),则目标扰动方法满足差分隐私;若随机噪声 $\boldsymbol{z}$ 服从 $\beta=n\lambda\varepsilon/2$ 时的概率分布,则输出扰动方法满足差分隐私;若随机梯度下降过程每次参数更新时所添加的随机噪声 $\boldsymbol{z}_t$ 服从 $\beta=\varepsilon/2$ 时的概率分布,则梯度扰动方法满足差分隐私[45,57].在文献[58]的启发下,表 1 从模型种类、隐私定义、扰动方式、实验设置及效果等角度对差分隐私保护的传统机器学习代表性研究工作进行了总结,实验涉及到的相关数据集信息见表 2.

Table 1 Comparison of Studies on Differentially Private Traditional Machine Learning

表 1 差分隐私保护的传统机器学习代表性研究工作比较

Related Work	Model	DP & Variants	Perturbation Method	Dataset	Number of Classes	ε
Chaudhuri et al. (2011)[59]	Logistic Regression	DP	Output/Objective Perturbation	Adult	2	0.2
	SVM			KDDCup99	2	0.2
Fukuchi et al.(2017)[53]	ERM	DP/LDP	Input Perturbation			
Zhang et al. (2012)[56]	Linear & Logistic Regression	DP	Objective Perturbation	US	2	0.8
				Brazil	2	0.8
Song et al. (2013)[57]	Logistic Regression	DP	Gradient Perturbation	KDDCup99	2	1
				MNIST	10	1
Wu et al. (2017)[60]	Logistic Regression	DP	Output Perturbation	CoverType	2	0.05
	Huber SVM			MNIST①	10	2
Jain et al. (2012)[61]	Linear & Logistic Regression (Online Learning)	DP	Objective Perturbation	CoverType	2	10

Table 2 Related Datasets

表 2 相关数据集

Dataset	Abstract	Number of Instances	Number of Attributes	Data Type
Adult	Census Data	45 220	15	Structured Data
US	Census Data	370 000	13	
Brazil	Census Data	190 000	13	
CoverType	Forest Cover Type	498 010	54	
KDDCup99	Network Connection Data	70 000	119	
MNIST	Handwritten Digits Images	60 000	784	Unstructured Data
CIFAR-10	Color Images	60 000	3 072	
SVHN	Color House-Number Images	630 420	3 072	

3.2 问题与不足

差分隐私机制已被广泛应用于逻辑回归[53,56-57,60-62]、支持向量机[59-60]等简单二分类模型,并能够实现较好的隐私与可用性的平衡.对于相对复杂的多分类任务,常用的一种学习策略是将其拆分为若干二分类任务,每个任务训练一个二分类模型,最终集成这些二分类模型的结果得到最终结果.不过,每训练一个二分类模型,都需多次访问训练数据.若每次访问数据的操作满足 ε_i-差分隐私,共访问 n 次,根据基本组合定理[43],整个训练过程满足 $\sum_{i=1}^{n}\varepsilon_i$-差分隐私,当 $\sum_{i=1}^{n}\varepsilon_i$ 很大时,势必将削弱隐私保护的效果[58];反之,若将 $\sum_{i=1}^{n}\varepsilon_i$ 维持在一个较小的范围内,相应需减小 ε_i 的值,必须对模型训练过程添加更多的噪声,从而导致模型可用性变差.对于结构更加复杂、训练过程需更多次迭代的深度学习模型,上述问题将更为严重.同时,深度学习的经验风险最小化目标函数是一个非凸函数,故基于函数敏感度分析的输出扰动和在目标函数后添加扰动项的目标扰动方法已不再适用.

① 作者在该实验中将 MNIST 数据集的数据特征(784 维)通过随机映射方法降至 50 维.

4 深度学习的隐私保护

求解深层网络模型的最优参数是一个非凸优化问题,不仅训练过程收敛慢,且极易陷入局部最优;同时,一个超大规模的深度学习模型可能涉及亿万级别的参数,故需进行大量的参数优化.上述问题致使在设计深度学习的隐私保护方法时面临更大的挑战.

基于函数敏感度分析的输出扰动方法不再适用,通过在目标函数后添加随机扰动项的目标扰动方法也无法应用于深度学习.Phan 等人[63-64]针对自编码器(auto-encoder, AE)和卷积深度置信网络(convolutional deep belief network, CDBN)提出先将非线性目标函数近似表示为参数的多项式形式,进而通过目标扰动,使训练过程满足差分隐私,不足之处是不易拓展到其他类型的深度神经网络.

模型训练过程需要更大的隐私预算:利用梯度下降法求解深度学习模型参数时,由于目标函数是非凸函数,且参数量大、结构复杂,故算法需经过更多次的迭代才可能收敛至最优解,且常常是局部最优解.若每次参数更新都满足差分隐私,整个训练过程的全局隐私成本将很大,从而难以合理地权衡数据隐私与模型可用性.

4.1 针对深度学习的隐私保护方法

为解决上述问题,近年来,基于宽松差分隐私(relaxed differential privacy)[65]定义的保护方法陆续被提出,并已应用到多种机器学习模型的隐私保护研究中.除此之外,利用集成模型将底层数据与用户访问接口隔离,一定程度上也能实现对训练数据的保护.

4.1.1 宽松差分隐私

最原始的差分隐私定义[6]要求算法在最大背景攻击——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.但实际应用中上述攻击往往很难实现.若一味基于这种过于保守的假设来设计隐私算法,其后果便是数据隐私与模型可用性的天平极大地偏向了隐私这一端,从而导致模型不可用.例如,文献[33]在针对个性化用药预测模型的实验中发现,若强制让模型满足 ε-差分隐私,其预测结果将导致病人治疗效果大大降低,甚至会增加患者的死亡风险.

解决该问题的一种方法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定义,这意味着算法存在一定隐私泄露的概率,尽管这个概率被限制在合理范围内.

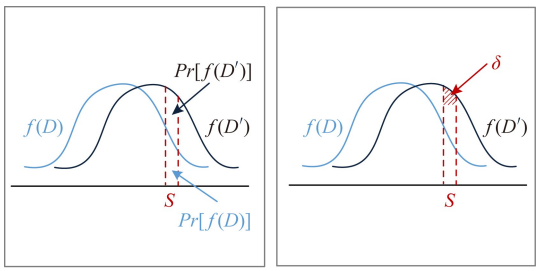
(ε,δ)-差分隐私((ε,δ)-differential privacy, (ε,δ)-DP)[66]是最早提出的一种宽松差分隐私定义,其形式化定义如下.

定义 4. (ε,δ)-差分隐私[66].对于任意的邻接数据集 D 和 D',且 $D,D'\in\mathcal{D}$.给定随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$和任意的输出结果 $S\subseteq\mathbb{R}$,若不等式

$$\max_S\left[\ln\frac{Pr[f(D)\in S]-\delta}{Pr[f(D')\in S]}\right]\leqslant\varepsilon \tag{11}$$

成立,则称算法 f 满足(ε,δ)-差分隐私.其中,δ 为一非零实数,且常常是一个很小的值.

如图 4 所示,$f(D)$与 $f(D')$输出结果在 S 之间的概率分别表示为对应曲线下 2 条垂直虚线间的面积,由于 δ 的存在,(ε,δ)-差分隐私(图 4(b))的隐私损失比 ε-差分隐私(图 4(a))小,表明更易满足定义的要求.



(a) ε-DP (b) (ε,δ)-DP

Fig. 4 ε-DP versus (ε,δ)-DP

图 4 ε-差分隐私与(ε,δ)-差分隐私

文献[67]在 ε-差分隐私的基础上提出仅让隐私损失的期望值,而不是最大值,控制在一定范围之内,从而进一步放宽了隐私的要求,其形式化定义如下.

定义 5. KL 差分隐私①[67].对于任意的邻接数据集 D 和 D',且 $D,D'\in\mathcal{D}$,给定随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$和任意的输出结果 $S\in\mathbb{R}$,若不等式

$$\underset{S\in\mathcal{R}}{E}\left[\ln\frac{Pr\,[f(D)\in S]}{Pr\,[f(D')\in S]}\right]\leqslant\varepsilon$$

成立,则称算法 f 满足ε-KL 差分隐私.其中,不等式左边等价于 $f(D)$和 $f(D')$的 KL 散度(KL-divergence).KL 散度也称相对熵(relative entropy),可用来度量 2 个概率分布之间的差异,故上述不等式可简化为

$$D_{\mathrm{KL}}(f(D)\,\|\,f(D'))\leqslant\varepsilon.$$

① 文献[67]中并未对这种隐私定义给出明确的名称,考虑到该定义与 KL 散度密切相关,故本文称其为 KL 差分隐私.

除此之外,基于类似定义的集中差分隐私(concentrated differential privacy, CDP)[68]、零式集中差分隐私(zero concentrated differential privacy, zCDP)[69]和雷尼差分隐私(Rényi differential privacy, RDP)[70]相继被提出.文献[58]对上述3种宽松差分隐私定义进行了较为全面的总结与对比,此处仅给出定义,不再详述.

定义6. 集中差分隐私[68].给定任意的邻接数据集 D 和 D' 以及随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$,$D_{\text{subG}}$ 为亚高斯散度(sub-Gaussian divergence).若不等式

$$D_{\text{subG}}(f(D)\parallel f(D'))\leqslant(\mu,\tau) \tag{12}$$

成立,则称算法 f 满足 (μ,τ)-集中式差分隐私.

CDP将隐私损失定义为一个服从亚高斯分布① 的随机变量,μ 和 τ 分别控制着该随机变量的均值和集中程度.若算法满足 ε-DP,则满足 $(\varepsilon(e^{\varepsilon}-1)/2,\varepsilon)$-CDP,然而反过来却不成立.

定义7. 零式集中差分隐私[69].给定任意的邻接数据集 D 和 D' 以及随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$,$D_{\alpha}$ 为 α-Rényi 散度(α-Rényi divergence)且 $\alpha\in(1,\infty)$.若不等式

$$D_{\alpha}(f(D)\parallel f(D'))\leqslant\xi+\rho\alpha \tag{13}$$

成立,则称算法 f 满足 (ξ,ρ)-零式集中差分隐私.

zCDP是CDP的变种,该定义下隐私损失将紧紧围绕在零均值周围.同样,若算法满足 ε-DP,则满足 $\varepsilon^2/2$-zCDP.Rényi 散度允许从 zCDP 直接映射到 DP,即若算法满足 ρ-zCDP,则满足

$$(\rho+2\sqrt{\rho\ln(1/\delta)},\delta)\text{-DP},\ \delta>0.$$

定义8. 雷尼差分隐私[70].给定任意的邻接数据集 D 和 D' 以及随机算法 $f:\mathcal{D}\mapsto\mathbb{R}$,$D_{\alpha}$ 为 α-Rényi 散度(α-Rényi divergence)且 $\alpha\in(1,\infty)$.若不等式

$$D_{\alpha}(f(D)\parallel f(D'))\leqslant\varepsilon \tag{14}$$

成立,则称算法 f 满足 (α,ε)-雷尼差分隐私.

相比于CDP和zCDP,RDP能够更准确进行隐私损失的相关计算.若算法满足 ε-DP,则满足 (α,ε)-RDP;相反,若算法满足 (α,ε)-RDP,则满足 $(\varepsilon+\ln(1/\delta)/(\alpha-1),\delta)$-DP,$0<\delta<1$.

为了控制深度学习模型时训练过程的全局隐私损失,算法中有必要引入一个能够对每次访问训练数据时所产生的隐私损失进行核算的模块,从而对整个分析活动的全过程加以控制和引导.该模块与现实生活中会计的职能十分相近,文献[65]形象地称之为"隐私会计(privacy accountant)",同时提出基于RDP的MA(moments accountant)机制.目前开发者已公开了MA及相关算法②,且用户可以方便地在Tensorflow深度学习框架中调用.

4.1.2 集成模型

文献[71-72]提出了一种基于知识迁移的深度学习隐私保护框架PATE,通过引入"学生"模型和多个"教师"模型,实现了将底层数据与用户访问接口隔离.不过,在某些极端情况下,如绝大多数"教师"模型的预测结果一致时,个体仍存在隐私泄露的风险.

表3总结了差分隐私保护下的深度学习代表性研究工作,涉及到的相关数据集信息见表2.

Table 3 Comparison of Studies on Differentially Private Deep Learning

表3 深度学习的差分隐私保护典型工作比较

Related Work	Model	DP & Variants	Perturbation Method	Dataset	Number of Classes	ε
Phan et al. (2016)[63]	AE	DP	Objective Perturbation	Behavior	2	1
Phan et al.(2017)[64]	CDBN	DP	Objective Perturbation	MNIST	10	1
Abadi et al.(2016)[65]	PCA+MLP	MA	Gradient Perturbation	MNIST	10	2
				SVHN	10	8
Papernot et al.(2016)[71]	GAN	MA	Output Perturbation	MNIST	10	2
				SVHN	10	8
Shokri et al.(2015)[73]	MLP/CNN	DP	Output Perturbation	MNIST	10	10
				SVHN	10	10
Jayaraman et al.(2017)[74]	ERM	zCDP	Output/ Gradient Perturbation	Adult	2	0.5
				KDDCup99	2	0.5
Geumlek et al.(2017)[75]	GLMs	RDP	Objective Perturbation	Adult	2	0.05
				MNIST	2	0.14
Geyer et al.(2018)[76]	CNN/LSTMs (Federated Learning)	MA	Gradient Perturbation	MNIST	10	8
Yu et al.(2019)[77]	PCA+DNN	zCDP MA	Gradient Perturbation	MNIST	10	0.781 25 (constant)

① https://en.wikipedia.org/wiki/Sub-Gaussian_distribution

② https://github.com/tensorflow/privacy/tree/master/privacy/analysis

4.2 问题与不足

基于宽松差分隐私定义的保护方法的代价便是当模型受到成员推理攻击或模型反演攻击时,造成泄露隐私的可能性更大了[58].文献[78]指出差分隐私仅能实现单点的隐私保护,若不同记录之间存在关联,攻击者仍可以对满足差分隐私保护的算法实施推理攻击.例如在社交网络中,某用户与其他用户节点之间存在多条社交关系,这些关系在数据集中以多条记录的形式保存.差分隐私只能孤立地为每一条记录提供保护,而不能同时保护该用户的所有记录,达到完全隐藏其存在于社交网络之中的目的.而在实际场景下,只有当保证攻击者无法推测出个体是否参与了数据生成过程时,才真正意味着实现了个体隐私保护.

5 联邦学习下的隐私保护

随着移动互联网与移动智能设备(如手机、平板电脑等)的高速发展,未来移动设备将成为技术创新和个人隐私保护的主战场.由于数据中包含了越来越多的个人敏感信息,早期将数据集中存储在单一服务器上进行机器学习的方式已不再可行,这一方面在于海量数据的存储与计算对服务器要求极高,另一方面在于一旦个人数据被集中收集,人们便失去了对其的知情权与控制权.一种解决方法是让存储与计算过程均在云端进行,如机器学习服务平台(ML-as-a-service, MLaaS)①,虽极大提高了计算效率,却并未改善隐私问题.为此,Google 提出了联邦学习[1-2],试图实现将各个设备的数据保留在本地的同时得到全局模型.目前联邦学习已在 GBoard 输入法中针对联想词[79]和智能提示[80]等功能进行了应用实践.

联邦学习与分布式机器学习中的数据并行化训练具有相似的逻辑结构.在联邦学习中,各方首先从服务端下载一个基本的共享模型,基于本地数据训练后将更新的模型参数上传至服务端;服务端将来自各方的参数整合至全局模型后再次共享出去,如此反复,直至全局模型收敛或达到停止条件(如图 5 所示).如同联邦制度一般,该训练模式下每个节点彼此独立且享有本地数据控制权,服务端无法直接访问各节点中的本地数据,仅能在参数层面进行模型的整合与发布.与数据并行化训练相比,联邦学习主要具有以下 4 个特点[81]:

1) 非独立同分布的数据样本.传统数据并行化训练由各个节点处理数据集的不同部分,且数据往往是独立同分布的.对联邦学习而言,由于数据是各参与方在其本地生成的,故很难具有相同的分布.

2) 各节点的数据量不平衡.以移动设备为例,用户数据大多来源于设备中安装的应用程序.由于各用户使用频率不同,其数据量往往存在较大差异.与人为地将数据集拆分不同,这种差异是由参与者的多样性决定的,是不可控的.

3) 超大规模分布式网络.随着移动设备覆盖率持续增长,诸如 Facebook、微信等热门应用程序的月活跃用户已超 10 亿②,此类应用场景中分布式网络的节点数量甚至远多于节点中存储的数据量,这种规模对传统分布式机器学习而言是难以实现的.

4) 通信受限.联邦学习同样具有传统分布式机器学习存在的通信问题,另外,受到硬件的限制和制约,移动场景面临更高的通信要求,如设备必须在接入无线网络以及充电状态下才能参与模型训练.

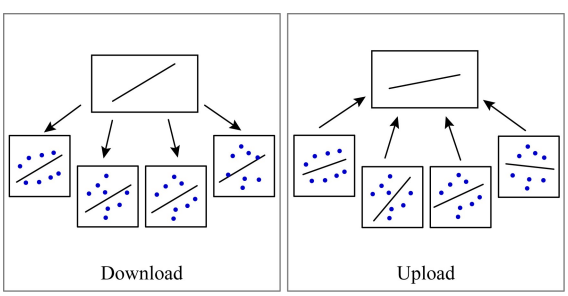


Fig. 5 Federated learning

图 5 联邦学习③

5.1 隐私攻击

相比于数据集中训练,联邦学习在隐私保护上具有更大的应用价值,但这并不代表它能完全防御外部隐私攻击.文献[12]对联邦学习面临的隐私问题进行了较为全面的分析与总结.对联邦学习而言,攻击既可能来自服务端,也可能来自其他恶意设备[82-83].服务端由于能够获得来自各个设备的模型

① ML-as-a-Service,指互联网服务商利用自己的数据和计算资源的优势,向用户有偿提供预先训练好的模型,或允许用户自己构建模型的一种服务平台.绝大多数互联网服务商仅向用户提供“黑盒”的模型访问方式,即用户只能通过 API 与模型进行交互;极少数服务商提供“白盒”的模型访问方法,即允许用户下载训练好的模型并部署到本地,本文不考虑该类型.

② https://www.appannie.com/en/go/state-of-mobile-2019/

③ 图片参考网络.http://vision.cloudera.com/an-introduction-to-federated-learning/

更新参数,故既能通过分析每轮更新的模型参数进行被动攻击,也可以通过将目标设备隔离,并向其传输设计好的参数以推测本地数据信息.其他设备由于只能获取来自服务端整合后的全局参数信息,故难以通过观察参数进行有效的推理,但可以利用梯度上升算法,观察全局参数每轮训练的变化趋势,进而实施攻击[12].

5.2 隐私保护技术

由5.1节可知,联邦学习中各参与方得本地数据可能在训练过程中被逆向推理从而造成隐私泄露.针对上述威胁,可通过以下3种技术予以保护.

5.2.1 加密技术

本文2.2节提到,传统加密技术的一大瓶颈是计算代价过于高昂从而在实际应用中可用性极差.文献[82]提出一种基于秘密共享的安全多方计算协议——安全聚合(secure aggregation),旨在保证设备与服务端之间通信及服务端参数聚合过程的安全性.与传统密码学方法相比,该协议的优点在于其计算代价并不高,但由于通信过程涉及大量安全密钥及其他参数,导致通信代价甚至会高于计算代价.另外,该方法假设服务端得到的全局参数不会泄露设备信息,然而,文献[84]基于聚合后的位置信息成功实施了成员推理攻击,由此证明该假设并不成立.

5.2.2 差分隐私机制

利用差分隐私,可以在本地模型训练及全局模型整合过程中对相关参数进行扰动,从而令攻击者无法获取真实模型参数.文献[77]提出对上传至服务端的参数更新值添加扰动的方法,使联邦学习过程满足差分隐私保护.文献[85]将类似的方法应用到联想词预测模型中,并在真实数据上进行评估,表现出较好的可行性.然而,与加密技术相比,差分隐私无法保证参数传递过程中的机密性,从而增加了模型遭受隐私攻击的可能性.另外,隐私与可用性的权衡问题在联邦学习下依旧存在.

5.2.3 区块链技术

区块链技术因其去中心化、安全可信、不可篡改等特性,能够监测服务端或设备在联邦学习中存在的恶意行为,保证训练过程的透明,从而为隐私保护提供一种新的解决思路.基于此,文献[9]提出DeepChain框架,该框架将区块链与5.2.1节提到的安全聚合协议相结合,既能保证本地参数在通信过程中的保密性与正确性,还能对联邦学习的全过程跟踪审计,并引入价值驱动机制,促进各方公平地参与协作训练.尽管如此,区块链技术本身仍存在吞吐量有限、可扩展性差等问题,故此类方法在实际场景中难以支撑大规模的应用,其有效性仍有待商榷.

5.3 问题与不足

与集中学习相比,联邦学习更强调个人对数据的控制权,故该方法对于医疗、金融、交通等领域下的机器学习任务尤为适用:一方面,此类场景下的数据往往包含大量个人敏感信息,且受政策与法律的制约不可传播与共享;另一方面,有限的数据使模型性能提升面临瓶颈.直觉上,联邦学习能够有效解决上述问题,并最终达到一个多方共赢的局面.不过,目前联邦学习仍处于起步阶段,无论是技术还是硬件条件,距离真正实现上述目标仍面临诸多问题与挑战,具体表现在以下3个方面[81].

1) 通信带宽受限.深度学习模型参数量大、结构复杂,故联邦学习下,其训练过程对设备内存、计算能力、带宽等有着极高的要求.尽管近年来复杂模型压缩研究取得了极大的进展[86],使得压缩后的模型能够在内存和计算资源有限的移动设备上高效运行,有限的带宽却使得设备与服务端之间参数的通信代价甚至高于将数据发送给服务端.

2) 模型收敛性.联邦学习是一个多轮训练过程,当全局模型收敛或满足停止条件时终止训练.由于联邦学习全局模型是由来自多个设备的参数聚合而成的,故如何保证算法能够逐渐稳定地收敛到最优解,提高算法的收敛速度,也是联邦学习的面临的挑战之一.

3) 联邦学习与云服务.联邦学习中,各个设备基于本地存储的数据训练模型,这些数据既包括应用程序客户端的行为与异常日志,也包括设备中存储的图片、语音等各类数据资源.不过,由于移动设备本身的物理资源十分有限,将所有数据都存储于设备中是很不现实的.文献[81]采取定期删除历史数据的方式解决上述问题,但此方法在实际应用中并不可行.随着云服务发展逐渐成熟,越来越多的人应用云备份和云存储来管理个人数据,如iCloud、百度云盘等,这些方式仍为集中式数据存储,一旦云服务提供方不可靠,数据隐私将面临极大挑战.如何协调联邦学习与云服务之间的关系也是目前亟待解决的问题之一.

6 未来挑战

纵观如今的机器学习的隐私保护研究,主要呈现出3个特点:一是存在被大多数人忽视的研究盲区,这些领域由于目前应用面较窄,情况更为复杂,

故人们在研究中很少考虑,或尚未提出有效的解决方法;二是隐私保护方法较为单一,基本围绕同态加密、安全多方计算和差分隐私机制 3 种方法,尽管这些方法表现出一定的有效性,但本身也存在固有的、难以解决的缺陷,缺乏本质上的创新;三是随着研究的不断深入,涌现出越来越多新的研究目标和研究任务,对保护算法的设计和应用提出了更高的要求.针对上述特点,本文提出未来机器学习的隐私保护研究中存在的五大研究挑战.

1) 推进无监督学习下的隐私保护研究

有监督学习是实际场景中最常见的一类机器学习任务,纵观现今的机器学习隐私保护研究工作,大多都是针对此类任务设计或改进保护方法的.反观无监督学习任务的研究却并没有有监督学习成熟.众所周知,人工数据标记费时费力,随着数据量的增长和数据维度的增加,未来无监督学习的研究价值也将愈加凸显.更重要的是,无监督学习下的隐私问题同样严峻,如针对聚类算法常见的背景知识攻击和一致性攻击.设想,若算法将所有病人的电子病历分为艾滋病患者、疑似艾滋病患者和正常患者 3 类人群,且攻击者已知与某病患同类的多数人均患有艾滋病,便能够推测该病患也患有艾滋病.匿名技术是解决上述隐私问题的一种常用手段,然而该技术的健壮性饱受质疑.此外,差分隐私也曾应用于聚类分析的隐私保护研究中,主要缺点是实现较难、误差较大,故未来仍需进一步深入研究[8].

2) 权衡差分隐私保护的模型可用性与隐私性

权衡模型的可用性与隐私是差分隐私机制的核心问题与未来发展的最大阻碍.尽管如今的一大发展趋势是抛弃传统严格的差分隐私定义,试图让算法满足一种相对宽松的隐私定义(见 4.1 节)以缓解复杂机器学习中存在的可用性与隐私难以平衡的问题,但模型受到隐私攻击的风险也增大了.寻找二者的平衡需综合考虑多种因素,包括数据对个体的敏感程度、服务提供商对模型性能的预期、不同个体对个人隐私的敏感程度等.在一些极度依赖模型可用性的应用场景下,人们甚至应严格控制模型的隐私性.例如基于病人的基因型及历史用药记录构造的个性化用药模型若过度强调病人的隐私,可能会使输出结果偏差过大,影响病人治疗进度甚至令其死亡.可见,对差分隐私机制而言,合理地权衡模型的可用性与隐私是一个十分复杂的问题,必须具体情况具体分析,甚至在特定情况下,差分隐私并不适合作为机器学习模型的隐私保护方法.

3) 探索多种技术结合的保护方法

差分隐私机制的优点在于添加随机噪声不会造成过高的性能代价,缺点在于扰动机制将可能使模型精度变差、输出结果的可用性降低;加密方法能够保证数据在存储、传输与计算过程中的机密性和正确性,但由于中间过程涉及大量计算和密钥传输,在应对复杂模型时其计算和通信开销都不容乐观;对区块链技术而言,因其具有的去中心化、安全可信、不可篡改等特性,能够为模型训练过程提供审计功能,识别恶意干扰的攻击者,然而区块链自身的性能瓶颈和不可拓展性使其难以支撑大规模的应用.直观来看,若能将上述 3 类方法加以结合,一定程度上能够实现功能的互补,提高隐私保护的效果.如文献[9]中提出将安全多方计算协议与区块链的结合,实现联邦学习下参数的安全聚合.差分隐私同样可以与区块链结合,从而在保护个体隐私同时保证训练过程的可审计.但目前而言,上述 3 类方法在实际部署或应用中均存在着不容忽视的局限性,这要求研究者在设计方法时必须充分考虑算法有效性和现实可行性,这也为算法创新带来了更大的挑战.

4) 支持单点和全局隐私保护

大数据时代,越来越多的应用场景对数据隐私的要求已不单单局限在对单个记录的保护,例如在社交网络中,一个用户往往与其他多个用户存在多条社交关系,而仅仅孤立地保护其中 1 条关系并不能掩盖用户在网络中存在的事实[78];医疗场景中,通过连续的心电图数据能够观察到病人是否患有心脏病,而保护单个数据点并没有实际意义.上述例子与位置隐私保护中的单点位置隐私和连续轨迹隐私[87]有异曲同工之妙,本文将此概念加以拓展,称为单点隐私与全局隐私.实现全局隐私保护并不是一个新问题,不过此前的研究工作大多针对计数、求和等简单的统计查询,很少考虑复杂的机器学习任务.改进已有的隐私保护方法,使其同时支持复杂机器学习过程中的单点和全局隐私保护,也是未来研究中的一大主要挑战.

5) 开发机器学习隐私保护框架

开发机器学习模型的隐私保护框架是近年来的一大研究热点.由本文第 1 节可知,现今机器学习隐私保护的研究延续了信息安全领域中的攻防机制,即针对特定的隐私攻击提出相应的防御方法,这使得隐私保护非常被动.设计一个通用、高效且健壮的隐私保护框架,是保证机器学习安全与隐私的另一大挑战.文献[88]提出一种在联邦学习方式下训练

深度学习模型的隐私保护框架 PySyft,该框架集成了安全多方计算和差分隐私机制 2 种隐私保护技术,并向用户提供深度学习应用程序接口.尽管该框架并没有解决 2 种技术各自存在的计算效率和预测精度问题,但仍是一次大胆的尝试.

7 结束语

机器学习的隐私问题是当前人工智能伦理研究的子问题,除此之外还包括数据伦理、算法偏见等.人们的最终目标是实现以人为本的人工智能,只有这样,社会才能真正信任技术,从而使人工智能长久地造福于人类.为此,2019 年 4 月,欧盟委员会(European Commission)发布人工智能道德准则 7 项要求,内容包括隐私和数据管理(privacy and data governance),透明性(transparency),多样性、非歧视和公平性(diversity, non-discrimination and fairness)等[①]实现上述准则离不开对机器学习可解释性的探索.理论上,可解释使人们有能力验证模型是否与自身需求一致,能够提供决策结果的审计和溯源,保证了决策公平,从而为解决伦理问题提供重要依据;同时,一些可解释性研究方法也可用作隐私保护算法设计的工具.但实现上,可解释的模型与其隐私保护之间却存在难以调和的矛盾,主要表现在 2 个方面:第一,实现可解释的前提是保证数据和模型的正确性,但基于扰动的隐私保护方法往往会导致隐私模型与真实模型存在偏差;第二,模型的可解释性越好,意味着人们能够对模型了解得更透彻,这也为攻击者提供了更多实施隐私攻击的机会.

上述问题一方面要求研究人员合理设计隐私保护方法和可解释分析框架,另一方面还需建立数据透明治理体系,保证数据在采集、存储、共享和决策过程中的透明,同时结合适当的法律法规与政策引导,此为解决人工智能伦理问题之关键.

参 考 文 献

[1] McMahan H B, Ramage D. Federated learning: Collaborative machine learning without centralized training data [EB/OL]. (2017-04-06) [2019-03-27] https://ai.googleblog.com/2017/04/federated-learning-collaborative. html

[2] Konecny J, McMahan H B, Yu F L. Federated learning: Strategies for improving communication efficiency [J]. arXiv preprint, arXiv: 1610.05492, 2016

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

[4] Rivest R, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms [J]. Foundations of Secure Computation, 1978, 4(11): 169-180

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

[6] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf. Berlin: Springer, 2006: 265-284

[7] Dwork C. Differential privacy [C] //Proc of the 33rd Int Colloquium on Automata, Languages and Programming. Berlin: Springer, 2006: 1-12

[8] Zhang Xiaojian, Meng Xiaofeng. Differential privacy in data publication and analysis [J]. Chinese Journal of Computers, 2014, 37(4): 927-949 (in Chinese)
(张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014, 37(4): 927-949)

[9] Weng Jiasi, Weng Jian, Zhang Jilian, et al. DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive [EB/OL]. IACR Cryptology ePrint Archive. 2018 [2019-03-27]. https://eprint.iacr.org/2018/679

[10] Meng Xiaofeng, Zhang Xiaojian. Big data management [J]. Journal of Computer Research and Development, 2015, 52(2): 265-281 (in Chinese)
(孟小峰, 张啸剑. 大数据隐私管理[J]. 计算机研究与发展, 2015, 52(2): 265-281)

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

[12] 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 [C/OL] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2019 [2019-06-09]. http://www.ieee-security.org/TC/SP2019/program-papers.html

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

[14] Newsome J, Karp B, Song D X. Paragraph: Thwarting signature learning by training maliciously [C] //Proc of the 9th Int Symp on Recent Advances in Intrusion Detection. Berlin: Springer, 2006: 81-105

① https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[15] Nelson B, Barreno M, Chi F J, et al. Exploiting machine learning to subvert your spam filter [C] //Proc of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. Berkeley, CA: USENIX Association, 2008 [2019-03-27]. https://www.usenix.org/legacy/events/leet08/tech/full_papers/nelson/nelson.pdf

[16] Rubinstein B I, Nelson B, Huang Ling, et al. Antidote: understanding and defending against poisoning of anomaly detectors [C] //Proc of the 9th ACM SIGCOMM Internet Measurement Conf. New York: ACM, 2009: 1-14

[17] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines [C] //Proc of the 29th Int Conf on Machine Learning. Madison, WI: Omnipress, 2012: 1807-1814

[18] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization [C] //Proc of the 11th ACM Workshop on Security and Artificial Intelligence. New York: ACM, 2018: 27-38

[19] Garcia-Ulloa D A, Xiong Li, Sunderam V S. Truth discovery for spatiotemporal events from crowdsourced data [J]. Proceedings of the VLDB Endowment, 2017, 10(11): 1562-1573

[20] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014 [2019-03-27]. https://iclr.cc/archive/2014/conference-proceedings/

[21] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world [C/OL] //Proc of the 5th Int Conf on Learning Representations Workshop Track. 2017 [2019-03-27]. https://openreview.net/forum?id=HJGU3Rodl

[22] Gu Tianyu, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain [J]. arXiv preprint, arXiv: 1708.06733, 2017

[23] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv: 1412.6572, 2014

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

[25] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57

[26] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 2154-2156

[27] Kurakin A, Goodfellow I J, Bengio S, et al. Adversarial attacks and defences competition [J]. arXiv preprint, arXiv: 1804.00097, 2018

[28] Song Lei, Ma Chunguang, Duan Guanghan. Machine learning security and privacy: A survey [J]. Chinese Journal of Network and Information Security, 2018, 4(8): 5-15 (in Chinese)
(宋蕾, 马春光, 段广晗. 机器学习安全及隐私保护研究进展 [J]. 网络与信息安全学报, 2018, 4(8): 5-15)

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

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

[31] Huang Ling, Joseph A, Nelson B, et al. Adversarial machine learning [C] //Proc of the 4th ACM Workshop on Security and Artificial Intelligence. New York: ACM, 2011: 43-58

[32] Biggio B, Fumera G, Roli F. Security evaluation of pattern classifiers under attack [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(4): 984-996

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

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

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

[36] Lowd D, Meek C. Adversarial learning [C] //Proc of the 11th ACM SIGKDD Int Conf on Knowledge Discovery in Data Mining. New York: ACM, 2005: 641-647

[37] Salem A, Zhang Yang, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models [J]. arXiv preprint, arXiv:1806.01246, 2018

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

[39] Rahman M A, Rahman T, Laganière R, et al. Membership inference attack against differentially private deep learning model [J]. Transaction on Data Privacy, 2018, 11(1): 61-79

[40] Du Wenliang, Han Y S, Chen Shigang. Privacy-preserving multivariate statistical analysis: Linear regression and classification [C] //Proc of the 4th SIAM Int Conf on Data Mining, Philadelphia, PA: SIAM, 2004: 222-233

[41] Jagannathan G, Wright R. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data [C] //Proc of the 11th ACM SIGKDD Int Conf on Knowledge Discovery in Data Mining. New York: ACM, 2005: 593-599

[42] Vaidya J, Kantarcıoglu M, Clifton C. Privacy-preserving naive Bayes classification [J]. International Journal on Very Large Data Bases, 2008, 17(4): 879-898

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

[44] Dwork C, Roth A. The algorithmic foundations of differential privacy [J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407

[45] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression [C] //Proc of the 21st Conf on Neural Information Processing Systems. New York: Curran Associates, 2008: 289-296

[46] Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates [C] //Proc of the 54th IEEE Annual Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2013: 429-438

[47] Agrawal R, Srikant R. Privacy preserving data mining [C] //Proc of the 19th ACM SIGMOD Conf on Management of Data. New York: ACM, 2000: 439-450

[48] Ye Qinging, Hu Haiho, Meng Xiaofeng, et al. PrivKV: Key-value data collection with local differential privacy [C/OL] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2019 [2019-06-09]. http://www.ieee-security.org/TC/SP2019/program-papers.html

[49] Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately?[C] //Proc of the 49th IEEE Annual Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2008: 531-540

[50] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms [C] //Proc of the 20th ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2001: 247-255

[51] Evfimievski A, Gehrke J, Srikant R. Limiting privacy breaches in privacy preserving data mining [C] //Proc of the 22nd ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2003: 211-222

[52] Agrawal S, Haritsa J R. A framework for high-accuracy privacy-preserving mining [C] //Proc of the 21st Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2005: 193-204

[53] Fukuchi K, Tran Q K, Sakuma J. Differentially private empirical risk minimization with input perturbation [C] //Proc of the 20th Int Conf on Discovery Science. Berlin: Springer, 2017: 82-90

[54] Ye Qingqing, Meng Xiaofeng, Zhu Minjie, et al. Survey on local differential privacy [J]. Journal of Software, 2018, 29(7): 159-183 (in Chinese)
(叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述 [J]. 软件学报, 2018, 29(7): 159-183)

[55] Li Ninghui, Ye Qingqing. Mobile data collection and analysis with local differential privacy [C] //Proc of the 20th IEEE Int Conf on Mobile Data Management. Piscataway, NJ: IEEE, 2019: 4-7

[56] Zhang Jun, Zhang Zhenjie, Xiao Xiaokui, et al. Functional mechanism: regression analysis under differential privacy [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375

[57] Song Shuang, Chaudhuri K, Sarwate A D. Stochastic gradient descent with differentially private updates [C] //Proc of the 2013 IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2013: 245-248

[58] Jayaraman B, Evans D. When relaxations go bad: Differentially-private machine learning [J]. arXiv preprint, arXiv: 1902.08874, 2019

[59] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization [J]. The Journal of Machine Learning Research, 2011, 12(2): 1069-1109

[60] Wu Xi, Li Fengan, Kumar A, et al. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics [C] //Proc of the 2017 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2017: 1307-1322

[61] Jain P, Kothari P, Thakurta A. Differentially private online learning [C] //Proc of the 25th Annual Conf on Learning Theory. Berlin: Springer, 2012: 24.1-24.34

[62] Kifer D, Smith A D, Thakurta A. Private convex empirical risk minimization and high-dimensional regression [C] //Proc of the 25th Annual Conf on Learning Theory. Berlin: Springer, 2012: 25.1-25.40

[63] Phan N H, Wang Yue, Wu Xintao, et al. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1309-1316

[64] Phan N H, Wu Xintao, Dou Dejing. Preserving differential privacy in convolutional deep belief networks [J]. Machine Learning, 2017, 106(9/10): 1681-1704

[65] Abadi M, Chu A, Goodfellow I J, et al. Deep learning with differential privacy [C] //Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308-318

[66] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation [C] //Proc of the 25th Annual Int Conf on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2006: 486-503

[67] Dwork C, Rothblum G N, Vadhan S P. Boosting and differential privacy [C] //Proc of the 51st Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2010: 51-60

[68] Dwork C, Rothblum G N. Concentrated differential privacy [J]. arXiv preprint, arXiv: 1603.01887, 2016

[69] Bun M, Steinke T. Concentrated differential privacy: Simplifications, extensions, and lower bounds [C] //Proc of the 14th Int Conf on Theory of Cryptography. New York: Springer, 2016: 635-658

[70] Mironov I. Rényi differential privacy [C] //Proc of the 30th IEEE Computer Security Foundations Symp. Piscataway, NJ: IEEE, 2017: 263-275

[71] Papernot N, Abadi M, Erlingsson Ú, et al. Semi-supervised knowledge transfer for deep learning from private training data [C/OL] //Proc of the 5th Int Conf on Learning Representations. 2017 [2019-06-12]. https://openreview.net/forum?id=HkwoSDPgg

[72] Papernot N, Song Shuang, Mironov I, et al. Scalable private learning with PATE [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018 [2019-06-12]. https://openreview.net/forum?id=rkZB1XbRZ

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

[74] Jayaraman B, Wang Lingxiao, Evans D, et al. Distributed learning without distress: Privacy-preserving empirical risk minimization [C] //Proc of the 31st Conf on Neural Information Processing Systems. New York: Curran Associates, 2018: 1439-1447

[75] Geumlek J, Song Shuang, Chaudhuri K. Renyi differential privacy mechanisms for posterior sampling [C] //Proc of the 30th Conf on Neural Information Processing Systems. New York: Curran Associates, 2017: 5289-5298

[76] Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective [J]. arXiv preprint, arXiv:1712.07557, 2017

[77] Yu Lei, Liu Ling, Pu C, et al. Differentially private model publishing for deep learning [C/OL] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2019 [2019-06-09]. http://www.ieee-security.org/TC/SP2019/program-papers.html

[78] Kifer D, Machanavajjhala A. No free lunch in data privacy [C] //Proc of the 2011 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2011: 193-204

[79] Hard A, Rao K, Mathews R, et al. Federated learning for mobile keyboard prediction [J]. arXiv preprint, arXiv: 1811.03604, 2019

[80] Yang T, Andrew G, Eichner H, et al. Applied federated learning: Improving Google keyboard query suggestions [J]. arXiv preprint, arXiv: 1812.02903, 2018

[81] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design [C/OL] //Proc of the 2nd Conf on Systems and Machine Learning. 2019 [2019-06-10]. https://www.sysml.cc/#schedule

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

[83] Briland H, Giuseppe A, Fernando P. Deep models under the GAN: Information leakage from collaborative deep learning [C] //Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 603-618

[84] Pyrgelis A, Troncoso C, Cristofaro E D. Knock knock, who's there? membership inference on aggregate location data [J]. arXiv preprint, arXiv:1708.06145, 2017

[85] McMahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018 [2019-05-10]. https://openreview.net/forum?id=BJ0hF1Z0b

[86] Ravi S. Custom on-device ML models with Learn2Compress [EB/OL]. (2018-05-09) [2019-05-10]. https://ai.googleblog.com/2018/05/custom-on-device-ml-models.html

[87] Pan Xiao, Huo Zheng, Meng Xiaofeng. Location Privacy Management in Big Data Era [M]. Beijing: China Machine Press, 2017 (in Chinese)
(潘晓, 霍铮, 孟小峰. 位置大数据隐私管理[M]. 北京: 机械工业出版社, 2017)

[88] Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning [J]. arXiv preprint, arXiv: 1811.04017, 2018

Liu Junxu, born in 1995. PhD candidate at Renmin University of China. Student member of CCF. Her main research interests include privacy preservation and machine learning.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Fellow of CCF. His main research interests include cloud data management, Web data management, privacy preservation, etc.