机器学习中的隐私保护综述

赵镇东[1,2], 常晓林[1,2], 王逸翔[1,2]

[1] 智能交通数据安全与隐私保护技术北京市重点实验室 北京 中国 100044
[2] 北京交通大学计算机与信息技术学院 北京 中国 100044

摘要 机器学习被广泛应用于自动推理、自然语言处理、模式识别、计算机视觉、智能机器人等人工智能领域, 成为许多领域研究与技术应用中必不可少的一个工具。然而, 机器学习本身存在隐私安全问题, 已经引起了越来越多的关注。本文专门针对机器学习中的隐私问题进行了分类和较为详细的介绍, 提出了基于攻击对象的隐私威胁分类方式, 并清晰地展示了防御技术的研究思路, 最后给出了亟待解决的问题和发展方向。

关键词 机器学习; 隐私威胁; 隐私保护
中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2019.09.01

A Survey of Privacy Preserving in Machine Learning

ZHAO Zhendong[1,2], CHANG Xiaolin[1,2], WANG Yixiang[1,2]

[1] Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing 100044, China
[2] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Abstract Machine learning is widely used in artificial intelligence, such as automatic reasoning, natural language processing, pattern recognition, computer vision and intelligent robots. It has become an indispensable tool in many fields of research and applications. However, there exist privacy issues in machine learning have attracted more and more attention. This paper specifically classifies and introduces the privacy issues in machine learning, proposes a new classification method of privacy threat according to the attack object, and clearly shows the research ideas of the defense technology. Finally, the problem that needs to be solved and the direction of development is given.

Key words machine learning; privacy threat; privacy protection

1 引言

近些年来, 机器学习理论技术不断成熟, 在很多领域取得了很好的应用成果。在人工智能方面, 机器学习算法是图像处理、模式识别、自然语言处理、自动推理等技术的核心工具, 已经被突破性地应用在文字语音书写识别、无人驾驶汽车、智能检测、智能机器人等实际生活中; 在大数据分析处理方面, 机器学习被用于数据挖掘、用户画像、医学统计中; 同时机器学习算法也为一些安全问题提供了便捷的解决措施, 比如垃圾邮件过滤、恶意代码检测等。综上, 机器学习正逐渐渗透到人们生活的各个领域, 成为方便人们生活和促进社会进步的关键技术。

然而在给人们生活带来便利的同时, 机器学习自身也存在一系列安全问题, 隐私威胁就是其中一个制约人们使用机器学习服务的关键要素。由于机器学习模型的可用性与可用于训练的数据量成正比, 因此大规模收集用户数据的商业公司成为这一行为的主要受益者。在大数据和云环境背景下, 用户担心自身的敏感数据被泄漏, 服务商担心服务模型的相关信息被窃取, 同时还存在攻击者通过一些手段获取数据来进行牟利。如果隐私问题不能得到有效解决, 会导致人们放弃便捷的云服务[1]。研究者们就如何在机器学习的应用过程中保障隐私这一问题, 已经展开了许多研究, 也取得了一系列进展。本文从机器学习中数据和模型两个方面对其中的隐私问题进行了归纳, 包括数据泄漏、模型反演攻击、成员推理攻击、模型窃取攻击; 并对相应的隐私保护技术——同态加密、差分隐私的研究情况进行了介绍和归纳整理。与现有的机器学习安全隐私类综述文献[2-6]相

通讯作者: 常晓林, 博士, 教授, Email: xlchang@bjtu.edu.cn。
本课题得到国家自然科学基金(No.U1836105)资助。
收稿日期: 2019-06-15; 修改日期: 2019-08-13; 定稿日期: 2019-08-20

比, 本文专门针对机器学习中的隐私研究进行了详细的介绍, 提出了新的基于攻击对象的隐私威胁分类方式, 并清晰地展示了防御技术的研究思路。

本文组织结构如下: 第 2 章扼要介绍了机器学习技术及其分类; 第 3 章对机器学习中存在的隐私威胁进行了归纳和分析; 第 4 章总结了隐私保护的思想方法, 并针对第 3 章的隐私问题给出了应对的防御技术的研究进展; 第 5 章进行了总结并展望了隐私保护技术的未来研究方向。

2 相关知识

机器学习涉及多门领域学科, 是对一类算法的总称, 这些算法尝试从大量历史数据中挖掘隐含在其中的规律并使用它们来预测新数据。更具体来说, 机器学习算法可以视为一个函数, 输入是样本数据, 而输出是期望的结果。值得注意的是, 机器学习算法的目标是使学习到的函数能够很好地适用于新样本, 而不仅仅是在训练样本上表现良好。通过学习得到的函数对于新样本的适用能力, 称为泛化能力。机器学习过程包括训练和预测两个阶段, 图 1 显示了其解决问题的过程。

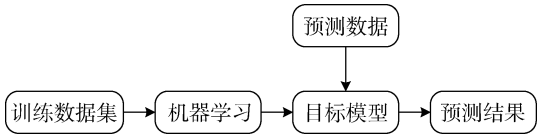

图 1 机器学习解决问题过程

Figure 1 Problem solving prcess of machine learning

机器学习的学习形式可以分为监督学习、无监督学习和强化学习 3 类[7]。监督学习的主要特点是在训练数据时给定类别标签, 多用于回归和分类问题, 常见的算法有支持向量机、决策树、K-近邻、朴素贝叶斯、神经网络、线性回归等。无监督学习的主要特点是缺少足够的先验知识, 训练阶段只提供训练样本, 而没有对应的类别标签信息, 典型的例子是降维和聚类, 常见的算法有主成分分析、k-means 算法。强化学习以试错(Try-and-error)的方式进行学习, 通过与环境进行交互获得的奖赏指导行为、改进行为来发现最优行为策略。

在神经网络的基础上, 机器学习发展出深度学习这一分支, 深度学习是一种对数据进行表征学习、能够模拟出人脑的神经结构的机器学习方法。深度学习形式同样分为监督学习和无监督学习, 卷积神经网络、循环神经网络是监督学习模型, 深度置信网络是无监督学习模型。

3 机器学习中的隐私威胁

机器学习模型的训练用到大量训练数据, 预测结果的准确性直接取决于可用于训练的数据量。而一些数据涉及人们的隐私如个人喜好、身份信息、地理位置、健康数据等, 用户不希望这些敏感信息被泄漏甚至被攻击者使用。如当机器学习应用于医疗系统中, 病人病历等隐私信息存在泄漏风险, 若敌手获取到这些信息并用于修改病人的用药剂量, 会造成极严重的生命危险。同时, 云服务中服务商在为用户提供付费服务时, 也希望保护自己的模型不被用户获取。由于角色不同, 所面临的隐私威胁也不同, 因此本文根据攻击对象的不同, 将机器学习中的隐私威胁分为针对数据的威胁和针对模型的威胁。针对数据的威胁有数据泄漏、模型反演攻击、成员推理攻击; 针对模型的威胁主要指模型窃取攻击。表 1 给出了机器学习隐私问题的分类和各种威胁形式。

其中, 数据泄漏可发生在训练和预测阶段, 模型反演攻击、成员推理攻击、模型窃取攻击均发生在预测阶段(如图 2)。

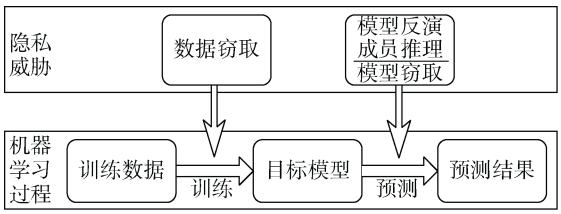

图 2 机器学习隐私威胁发生阶段

Figure 2 Stages of privacy threats

表 1 机器学习中的隐私威胁

Table 1 Privacy threats in machine learning

分类	威胁形式	发生阶段	敌手能力	敌手知识
针对数据	训练、预测数据泄漏	训练/预测阶段	获取数据信息	有限知识
	模型反演攻击	预测阶段	已知模型信息	黑盒/白盒
	成员推理攻击	预测阶段	访问目标模型	黑盒
针对模型	模型窃取攻击	预测阶段	访问目标模型	黑盒

3.1 针对数据的威胁

大数据时代的数据隐私极为重要, 机器学习模型的训练需要提供大量训练数据, 其中一些数据用户不愿对外公开, 攻击者可能会利用获取的数据进行非法活动, 造成严重后果。

3.1.1 训练、预测数据泄漏

机器学习对数据的训练可以分为集中式训练和联合分布式训练。在集中式训练中, 这些训练数据大多来源于人们日常, 其中不乏一些隐私数据。人们出于隐私考虑, 不愿这些数据被泄漏给无关方。而尤其在医疗数据的处理方面, 病患数据的高度敏感性, 更增加了数据泄漏的风险。另外, 有时数据需要发布给服务商以请求服务, 这些服务需要执行索引、查询等相关操作, 数据加密会导致这些操作出现问题, 因此云端的应用程序使用的静态数据经常不会加密。除了在云中, 传统的 IT 环境对于一些要处理的程序, 数据也几乎不经过加密, 这严重威胁了数据的安全性[8]。

在云环境下, 由于数据量过于庞大, 有时单一平台计算能力不足, 需要跨平台进行联合分布式计算。由各个参与方在独自的数据集上训练模型, 通过共享训练结果完成最终的服务模型。在这种环境下如果存在不诚实的参与者, 就有可能导致其他参与方的数据被窃取(如图 3)。Hitaj 等[9]用生成对抗网络(Generative adversarial networks)对在联合分布式环境下的训练模型发起攻击, 结果表明任何参与者都有可能成为攻击者, 通过在联合环境下生成与其他参与者近似的假样本来窃取数据隐私。

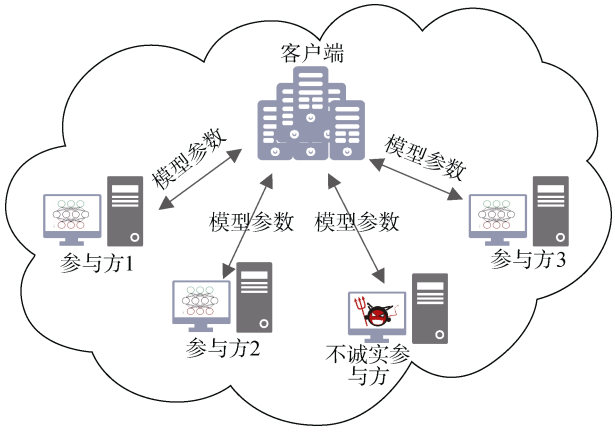

图 3 联合分布计算中的数据窃取

Figure 3 Data theft in collaborative distributed compting

3.1.2 模型反演攻击

模型反演攻击(Model inversion attack)是利用机器学习模型提供的一些 API(Application program interface)来获取系统模型的一些信息, 通过这些信息逆向反演(如图 4), 进而获取用于训练数据集中的隐私信息(病人诊断数据、用户记录信息、生物特征数据等)。根据敌手背景知识的多少, 可将其分为白盒攻击(White-box attack)和黑盒攻击(Black-box attack)[10]。白盒攻击指攻击者知道机器学习算法的细节以及相关参数, 黑盒攻击是指攻击者对模型没有了解, 只能与机器学习系统提供的接口进行交互。

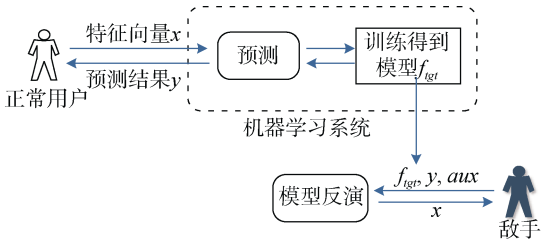

图 4 模型反演攻击

Figure 4 Model inversion attack

2014 年, Fredrikson 等[11]重点研究了模型本身会在多大程度上泄漏数据隐私, 并在药物剂量预测实验中, 用一种通用方法实现了对基于线性回归算法定制化个人药物系统的模型反演攻击, 利用病人的一些非敏感属性恢复出患者的基因组信息。2015 年, Fredrikson 等[10]在之前的基础上, 展开了对决策树算法的模型反演攻击研究, 实现了对决策树算法的白盒攻击和黑盒攻击, 并进一步使用梯度下降算法对相应的人脸图像进行重建, 不断调用模型获取输出, 最终获得了用于训练的人脸图像数据。Ateniese 等[12]表明可以从机器学习模型中推断出训练数据的统计特性。Wu 等人[13]分别描述了针对黑盒和针对白盒的模型反演攻击, 并发现了高度可逆模型通过增加少量噪声可以变得不可逆的现象。

与模型反演有关的其他工作表明[14], 对训练算法的微小改变可以产生几乎无法区分的相似模型, 可以利用这些模型泄漏大量私人数据, 该工作评估了机器学习技术中用于图像分类的 CIFAR-10, 人脸识别的 LFW 和 FaceScrub 以及文本分析的 20 个新闻组和 IMDB, 在这几种数据集上, 展示出他们的算法如何创建具有高预测能力但能够准确提取其训练数据子集的模型。

在对模型反演攻击的最近研究中, Carlini 等[15]提供了一种方法来检验模型是否可能记住并泄漏部分敏感训练数据, 并为此开发了一种定量测试程序, 该程序可以衡量模型记忆的敏感数据量。Hidano 等[16]

提出了一个通用的模型反演框架, 该框架模拟了对手可用的辅助信息量, 他们将恶意数据注入机器学习系统, 以便对模型进行修改并生成一种新的特殊类型的模型, 允许对手在不了解非敏感属性的情况下执行模型反演攻击。

3.1.3 成员推理攻击

2017 年, Shokri 等[17]提出成员推理攻击(Membership inference attack): 对于给定的数据, 可以确定其是否是训练集中的原数据。他们表明, 使用机器学习即服务(Machine learning as a service, MLaaS)平台创建的模型在一定程度上, 都会存在泄漏原始数据信息的风险。攻击方法是根据目标模型的输入数据及预测标签训练一个和目标模型相似的影子模型, 然后将给定的数据分别输入目标模型和影子模型, 通过观察影子模型与目标模型所输出的预测向量之间的差别来判断所给定的数据是否是用来训练目标模型的训练数据。一些文献[18-19]论述了在协作环境中如何对模型进行成员推理攻击, 说明了任何处在协作环境下的参与者, 都有可能从其他参与方的设备中推理得到敏感信息。

在最近的研究中, 文献[20]对成员推理攻击进行了比较全面的研究, 提出了一个黑盒成员推理攻击模型的通用公式(如图5), 并给出了评估模型在什么条件下容易受到这种黑盒成员推理攻击的方法。Hayes 等[21]利用生成对抗网络结合判别式模型和生成式模型来识别输入, 针对几种最先进的生成模型, 对面部图像、医学图像、特定对象等数据集进行了攻击, 还讨论了不同训练参数对该攻击的影响。Sablayrolles 等[22]通过对参数分布的一些假设推导出成员推理的最优策略, 并证明了最优攻击仅依赖于损失函数;这说明了白盒攻击并不能比黑盒攻击提供更多信息。

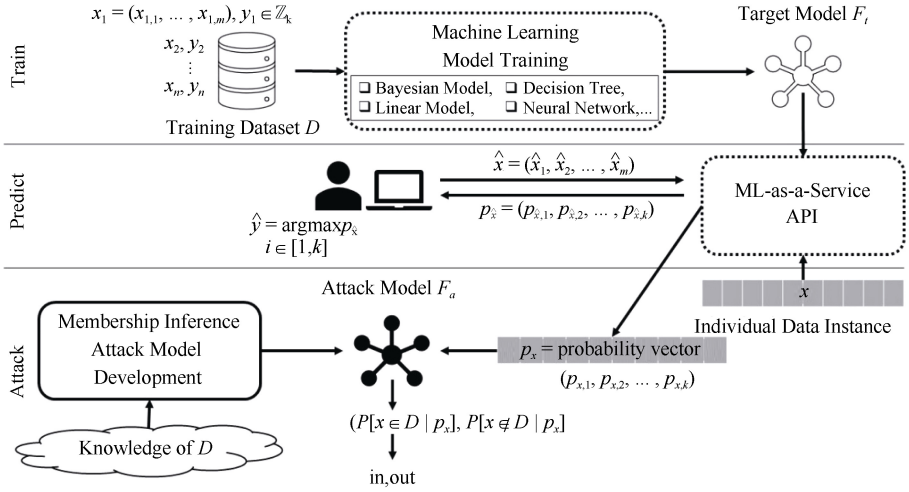

图 5 成员推理攻击工作流程[20]

Figure 5 Process of membership inference attack

3.2 针对模型的威胁

针对模型的威胁主要指对模型的窃取, 即通过一些手段窃取机器学习模型结构或模型内部参数。机器学习云服务商拥有预先训练好的处理模型, 用户向服务方提供数据, 并借助服务商的计算能力对数据进行处理, 服务商会对用户的访问收取一定费用。在这种场景下, 服务平台模型的内部信息成为隐私。服务使用者只能通过调用服务商提供的接口来使用机器学习算法处理自己的数据, 对算法和模型的内部信息一无所知。因此, 用户的访问被定义为黑盒访问, 服务商需要保护自身的模型不被使用者获取。

窃取模型的目的可以分为 3 类: 1)免费使用模型。Tramer 等人[23]利用一些机器学习服务供应商提供的黑盒模型接口输出的置信值, 对其进行等式求解攻击(Equation solving attacks), 并在模型只有输出标签信息的情况下实现了对决策树、逻辑回归和神经网络等机器学习算法的简单攻击, 获取了目标模型的信息。2)窃取训练数据。攻击者获取了模型, 继而可以使用模型反演攻击反过来对训练数据的信息进行恢复。3)逃逸攻击。当前许多场景应用机器学习进行恶意检测, 比如垃圾邮件分类、恶意代码检测, 攻击者获取目标模型之后, 可以根据模型信息构造对抗样本来逃避安全检测[24-25]。

对模型窃取的方法显示, 可以通过观察模型的输入和输出来提取模型信息。如果模型是线性的, 理

论上一个 n 维的模型通过 n+1 次查询就可以得到其全部参数。这可以简单地归结为已知 x 和 $h_\theta(x)$ 求解 θ。同时可以通过对模型的黑盒访问, 利用生成的数据训练出一个原模型的近似模型, 借此来获取原模型的相关信息来展开攻击。

后来一些研究人员将模型窃取攻击扩展到超参数窃取攻击, 文献[26]中提出的超参数窃取攻击可适用于逻辑回归、岭回归、支持向量机以及神经网络等流行的机器学习算法, 并评估了对亚马逊机器学习平台的攻击。Seong 等人[27]的相关工作也提出了类似论点, 即窃取超参数可能有助于在模型上进行更强大的隐私攻击。

在最近的研究里, Hu 等[28]用一种语音识别的思想, 分析了 DNN 设计中层内架构特征和层间关联情况的可能性, 实现了深度神经网络模型的提取。Orekondy 等[29]通过黑盒访问模型, 将查询得到的预测数据用于训练一个“仿冒”模型, 通过这种知识迁移的办法将受害者模型的功能转移到仿冒模型中, 具有强大的攻击有效性。

4 机器学习中的隐私保护技术

4.1 隐私保护思想

常用隐私保护策略有泛化、匿名、添加随机扰动、加密等。

泛化(Generalization)是指发布数据时, 隐藏一些具体细节, 但发布的数据在语义上和原数据保持一致。采用的方法可以是将原数据进行适当的变形, 变形后得到的新数据具有更少的信息含量。这样在防止推理攻击、保护特定敏感属性的同时, 也保持了原始数据的统计特性。泛化的主要方法有如下几种: 二元搜索、完全搜索和先验动态规划等, 它们在保证隐私的条件下, 尽量做到减少信息损失, 但是仍然带来一些不可避免的信息损失。

匿名(Anonymous)是最早提出的隐私保护技术, 方法是隐藏发布数据中的标识属性。其中应用较多的 k-匿名(k-anonymization)是一种有效的保护私有信息的数据发布方法。k-匿名技术于 1998 年由 Samarati 等人[30]提出, 它通过参数 k 指定用户可承受的最大信息泄露风险, 要求发布的数据在准标识符上存在至少 k 条不可区分的记录, 使攻击者不能判断特定信息所属的具体个体。k-匿名在一定程度可以保护数据隐私,但也降低了数据的可用性。之后, l-diversity[31]、t-closeness[32]等技术的提出不断完善着针对不同攻击者背景知识的匿名保护理论。对匿名化的研究工作主要集中在保护私有信息的同时提高数据的可用性。

随机扰动(Perturbation)是指在原始数据中引入噪声, 使得新数据与原始数据产生差异, 具有一定随机性, 而不暴露原信息, 从而减少了隐私攻击的可能性, 这可以在保持原始数据相关性和统计特性不变的前提下, 通过降低某一具体条目的信息准确性来抵抗隐私推理攻击, 一般噪声越大隐私保护度越高, 但数据的实用性越小。差分隐私技术[33]应用了添加随机扰动的思想。

加密(Encryption)是保护数据隐私最常用的方法, 指一个信息通过加密密钥和加密函数转换, 变成无意义的密文, 使得未拥有解密密钥的用户即便获取了数据, 因为不知晓解密方法, 而无法获得数据中的有用信息。接收方可以通过解密函数和解密密钥将密文还原成明文。加密类型分为对称加密和非对称加密。

表 2 隐私保护主要思想归纳

Table 2 The main methods of privacy preserving

隐私保护思想	原理	实例
泛化	减少细节使其含有较少的信息含量	二元搜索
匿名	隐藏发布数据中的标识属性	k-匿名
添加随机扰动	对数据添加噪声, 具有随机性	差分隐私
加密	将信息转化成无意义的密文	同态加密

4.2 隐私保护技术

为了保护机器学习环境下数据和模型的隐私, 针对上一部分提到的数据泄漏、模型反演攻击、成员推理攻击, 在应用中, 常常根据实际情况, 通过设计安全多方计算[34]和隐私保护协议来保护隐私性, 这些协议中往往会结合同态加密技术和差分隐私技术; 而对于模型窃取攻击, 可以通过在神经网络中加入一些假神经元或对机器学习算法的模型参数添加随机扰动来保护模型不被窃取。表 3 给出了针对各种攻击的防范措施, 后面几个小节对这些技术的研究进展进行了介绍和分析。

表 3 针对各种隐私威胁的防范措施

Table 3 Privacy preserving methods to privacy threat

威胁形式	防范措施	防范对象
数据泄漏	同态加密、安全多方计算协议	服务商、不良参与方
模型反演攻击	同态加密	不良用户
成员推理攻击	差分隐私	不良用户、不良参与方
模型窃取攻击	假神经元[54]、对模型参数添加随机扰动[69]	不良用户

4.2.1 同态加密技术

同态加密(Homomorphic encryption)支持用户直接在密文上进行运算, 得到的结果解密后与直接在明文下运算一致, 是直接有效保护数据隐私的一项技术。1978 年, Rivest 等[35]在银行应用背景下提出了同态加密的概念。同态加密满足的公式是

$$Enc(f(m_1,m_2)) = f(Enc(m_1),Enc(m_2))$$

同态加密分为部分同态加密和全同态加密。部分同态加密分为加法同态和乘法同态, 若一个方案同时满足加法同态和乘法同态, 则该方案为全同态加密方案。经典的 RSA 加密算法和 Elgamal 加密算法[36]具有乘法同态性。1999 年, Paillier[37]提出了一种加法同态的加密方案, 而且是可证明安全的加密方案。2009 年, Gentry[38]提出了一种基于理想格的全同态加密方案, 提供了一种构造全同态的蓝图。2015 年, 微软开源了一个全同态加密算法库 SEAL①。图 6 显示了同态加密下的云服务过程。

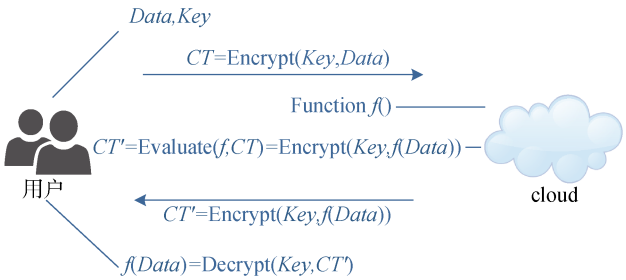

图 6 同态加密下的云服务

Figure 6 Cloud service under homomorphic encryption

虽然理论上同态加密可以对密文进行运算与预测, 但是在实际的应用过程中还存在许多局限性, 其中包括:

1) 加密后的同态运算只支持加法和乘法的基本操作, 而机器学习算法中包含比较、求最大值、幂运算等复杂操作;

2) 同态加密机制会引入噪声, 噪声水平会在同态运算过程中不断增长, 尤其是乘法运算。当一个运算的乘法深度过高, 会导致最终的噪声水平超过阈值, 而无法正确解密。

在如何将同态加密更好地应用于机器学习算法中这一问题上, 研究工作者已经取得了研究成果。解决办法总结起来可以分为两种策略: 一种是通过安全多方计算方法构建适合同态加密的隐私保护协议; 另一种是寻找机器学习算法的近似算法来适应同态加密, 即"构建协议"的办法和"算法近似"的办法。

(一) 传统机器学习算法的同态加密研究

为了实现决策树算法中的数据隐私保护, 早期 Du 等[39]提出了一种私有标量积协议, 实现合作决策树的分类; 之后, Zhan[40]应用 Paillier 加密算法, 构建了同态加密下的决策树分类模型。2015 年, Aslett 等人[41]提出了基于全同态加密数据的完全随机森林算法。Wu 等人[42]提供了隐私决策树和随机森林的评估协议, 用到了加法同态加密算法和不经意传输协议, 使用深度高达 20、超过 10,000 个决策节点的决策树来评估他们协议的可扩展性, 结果表明该协议在保证算法效率的前提下也保护了算法的数据隐私。最近的研究中, Alabdulkarim 等[43]在医疗服务设计了一种隐私保护临床决策支持系统, 该系统基于决策树算法, 并运用同态加密技术来保护用户数据, 满足了医院数据集的隐私要求。

刘晓红[44]提出了一种加密支持向量机算法, Rahulamathavan 等[45]实现了加密支持向量机的协议; 文献[41]中给出了同态加密的朴素贝叶斯算法; Bost 等[46]用同态加密构建了超平面决策, 朴素贝叶斯和决策树三种算法的隐私保护协议, 并提供了从一种加密方案切换到另一种加密方案的机制, 开发了一个构建块库, 可用于构建其他分类器。Wood 等[47]提出了在私钥体制全同态加密方案下的朴素贝叶斯分类协议的实现。

在逻辑回归中, Zhu 等[48]考虑建立外包数据的交互式协议, 其中客户必须始终与云服务器进行多轮通信。协议中的通信成本取决于数据集大小和维度, 同样取决于客户的计算成本。Aono 等人[49]将似然函数转换为低次多项式, 并使用加法同态加密算法来对中间统计进行加密, 系统具有一定的可扩展性。但

他们的方案为了在本地最小化参数, 依赖于客户端来解密这些中间统计信息。Kim 等[50]采用针对实数计算优化的新型同态加密方案, 设计了逻辑回归的最小二乘近似算法, 获得了较好的准确率和效率。

在聚类算法里, 2017 年姚禹丞等人[51]提出了一种同态加密的分布式 K 均值聚类算法, 新的加密体制解决了分布式环境下多方执行 K-means 聚类挖掘任务的安全问题。最近的研究中 Meng 等[52]在层次聚类算法中, 使用同态加密并构造了加密数据计算的隐私保护协议, 既给出了正式的安全定义和证明, 也提供了实现方法, 为无监督学习的隐私保护研究提供了新的途径。

(二) 神经网络及深度学习算法的同态加密研究

针对机器学习神经网络应用中的隐私保护问题, 2006 年, Barni 等[53]人提出了一种隐私保护计算协议, 使用 Paillier 加密算法对神经网络的输入数据进行加密, 并提供了 3 种安全等级: 保护数据、保护激活函数、保护拓扑结构。之后, Orlandi 等人[54]加强了协议, 首先, 通过 Damgard 等人[55]的扩展, 替换了 Paillier 密码体制。其次, 在将激活函数的评估委托给客户端之前, 对标量积结果进行隐藏, 并在神经网络的结构中添加了一些假神经元, 起到模型不被窃取的作用。该协议的缺点是不符合效率要求。为了克服这个缺点, Gilad-Bachrach 等[56]提出了一种叫做 Cryptonets 的隐私保护神经网络模型, 他们使用全同态加密算法对预测数据进行加密, 在训练好的卷积神经网络上, 用具有乘法深度 1 的平方函数代替激活函数, 最终在 MNIST 数据集上获得了 99%的精确度。但其存在的不足是, 该方案仅适用于小型神经网络, 对于超过 2 个非线性层的神经网络, 准确度非常低。之后, Chabanne 等人[57]为解决这一问题, 建议训练阶段的神经网络仍使用具有 ReLu 激活功能的神经网络, 而在测试阶段每个低度非线性多项式激活层前加入一层批量标准化层, 以便在激活函数输入上具有稳定的正态分布, 提高了分类的准确率。最近, Chou 等[58]提出了更快的 CryptoNets, 他们推导出常用的激活函数的最佳近似, 实现了最大稀疏编码, 最小化逼近误差, 并通过整个神经网络的稀疏表示加速了对加密数据的深度学习模型的同态计算。

表 4 机器学习算法与同态加密结合的研究思路

Table 4 Research ideas of machine learning algorithm combined with homomorphic encryption

算法	与同态结合办法	相关文献	研究思路
决策树和完全随机森林	构建协议	[39, 59]	标量积协议、基于可信计算平台的协议
	算法近似	[41]	设计适用于同态加密的近似算法来解决数学运算中的比较、分割等操作
支持向量机	构建协议	[45, 59]	建立用于保护隐私的多方学习系统
	算法近似	[60]	研究合适的核方法和分类策略
朴素贝叶斯	构建协议	[46-47]	引入密码学的相关理论和技术辅助构建协议
	算法近似	[41]	对不能直接实现的分布中的一些运算进行近似、模拟决策边界
逻辑回归	构建协议	[48]	构建多次通信的交互式协议
	算法近似	[49-50]	最大似然或最小二乘法的近似算法来解决其中的复杂运算
聚类	构建协议	[51-52]	设计合适的加密机制和安全多方计算协议
神经网络	构建协议	[53-54, 61]	早期通过标量积协议、交互式协议与部分同态加密结合, 现在通过安全多方计算协议
	算法近似	[56, 62, 58]	找到激活函数的多项式近似、神经网络结构的优化、批量标准化

表 4 对同态加密在机器学习算法中的应用方法和相关研究思路进行了归纳。上述文献的方法大多在预测阶段应用同态加密, 理论上训练阶段也可以同态加密数据。Xie 等[62]利用 Stone-Weierstrass 理论[63], 提出应用 crypto-nets 框架在密文上做预测。作者用加密后的数据作为训练集进行神经网络的训练, 并讨论了使用加密数据训练的适用性, 得出: 1)只有在样本很少的情况下, 或者网络层数较少才是可行的; 2)多个样本不能使用不同的密钥加密, 这种情况只能通过安全多方计算解决; 3)对于已训练好的目标模型, 可通过微调来适应特定数据, 只要数据量不大且网络可以通过低度多项式来近似就可以实现。Zhang 等[64]也给出了在密文上训练神经网络的解决方案, 他们用泰勒公式对神经网络中的激活函数进行近似, 用户将全同态加密后的数据提供给云端训练, 每经过一次反向传播过程, 云端给用户返回神经网络参数进行解密, 然后重新加密后再上传到云端进行迭代运算, 以此方式来避免网络深度加深导致的无法正确解密, 繁琐的训练过程导致很低效的问题。Wiesberg[65]在加密数据上运行 K-Means 聚类算法, 调整了算法使其能够在全同态加密的数据中有效执行, 具有较好的性能, 缩短了运行时间。

从以上使用加密数据进行训练的方法结果中可以看出, 在训练过程中采用同态加密来保护数据隐私虽然理论上可行, 但是由于深度神经网络的训练本身已耗费大量资源, 同态加密在应用上也存在效率低的缺陷, 训练神经网络的速度成为巨大的问题。而同态运算乘法深度加深导致的噪声水平增长, 一旦超过阈值, 结果将无法正确解密。因此对于训练数据的加密在实际中还不能得到很好的应用, 利用加密技术来保护机器学习用户敏感数据多用于预测阶段。同态加密的效率是应用中的一个较大问题, 对于同态加密的研究还在继续。

4.2.2 差分隐私技术

差分隐私(Differential privacy)是一种具有强大数学理论支撑的密码学手段。2006 年, Dwork 等[66]提出差分隐私概念——通过引入噪声使至多相差 1 个数据的 2 个数据集查询结果概率不可分。具体公式如下:

$$pr\left[A(T)\in S\right]\leqslant e^{\varepsilon}pr\left[A(T')\in S\right]+\delta$$

其中, T、T'为至多相差 1 个数据的两个数据集; ε 为隐私预算, ε 越小, 方案具有越强的隐私保护能力; δ 代表隐私预算不成立的容忍度。在机器学习的训练和预测阶段中, 可以通过对数据集、模型参数、预测结果中加入噪声的防御手段, 使特定的数据失去现实意义, 而统计信息仍具有应用价值。常用的添加噪声的机制为 Laplace 机制和指数机制。Laplace 机制用于数值型结果, 指数机制用于非数值型结果。噪声的引入会降低模型预测的精度, 随着噪声的增加会使得模型的可用性降低。如何在保证隐私的情况下, 提高目标模型的精度是研究人员的重点研究方向。

(一) 传统机器学习算法与差分隐私的结合

在分类技术中, 决策树(Decision tree)是一种典型的树形分类模型, SuLQ-based ID3[66]、DiffP-C4.5[67]以及 DiffGen[68]是决策树分类器与差分隐私技术结合的代表方法, 它们递归地构建决策树, 并采用了信息增益的方式分割数据属性。支持向量机也是常用的分类方法, Smith 等[69]提出了差分隐私的支持向量机算法, 利用 Laplace 噪声对模型权重进行扰动使得算法满足 ε-差分隐私, 但如果模型权重敏感性过高会导致噪音量过大, 降低分类准确度。Jing 等[70]对代价函数添加了 Laplace 分布的噪声, 在分类结果中具有较高的准确度, 但是缺陷是其必须要特定的代价函数。

回归是从一组数据出发, 确定某些变量之间的定量关系, 即建立数学模型并估计未知参数, 常用的回归模型是线性回归。回归模型包含一组权重, 如果权重被泄漏则模型和数据都会受到威胁。对于线性回归, Zhang 等人[71]提出了一种函数机制, 通过向代价函数的系数添加噪声, 再用梯度下降算法求解参数, 来执行 ε-差分隐私。通过理论分析和实验, 该函数被证明确保了隐私并具有较高效率, 它的局限性是只能应用于线性表示的目标函数中。

在聚类中, Nissim 等[72]结合采样与聚集技术, 提出了一种满足差分隐私的 k-means 聚簇中心发布方法 Pk-means, 该方法给出了聚类敏感性的度量方法以及聚类误差的下界。在 k-means 聚类过程中, 隐私预算 ε 的设置也非常关键, Dwork[73]提出了两种分配方法: 1)迭代次数 n 已知的情况下, 每一轮聚类隐私预算为 ε/n; 2)迭代次数未知的情况下, 每轮隐私预算为上轮预算的一半。以上提到的两种聚类方法虽然均满足 ε-差分隐私, 但具有较差的实际应用性。当数据集很大时, n 的选择是一个 NP-hard 问题, 有可能泄漏真实的数据点, 并且对 n 的每次选择都要消耗隐私预算。2018 年, Zhang 等[74]通过向聚类中心点添加符合 Laplace 分布的扰动来实现隐私保护。为了解决导致中心点偏离的拉普拉斯噪声随机性问题, 他们使用轮廓系数来定量评估每次迭代的聚类效果, 并为不同的聚类种类添加不同的噪声, 提高了算法聚类结果的可用性。

(二) 深度学习差分隐私的研究进展

在集中式学习中, 早期工作尝试通过仅处理训练过程产生的最终参数来保护训练数据隐私, 忽略掉了模型参数对训练数据的依赖性。针对这个问题, 2016 年 Abadi 等[75]提出满足差分隐私的深度学习训练算法, 在随机梯度下降的迭代过程中对梯度添加扰动, 并设计了一种计算隐私成本的方法, 渐进地对整体隐私成本加以衡量, 经实验证明, 可以在隐私成本可控的情况下完成深层神经网络的训练。2017 年, Papernot 等[76]提出了一个 PATE(Private aggregation of teacher ensemble)框架, 将训练数据集不相交地划分为 N, 在每个数据集上独立训练得到 N 个称为教师的机器学习模型, 每个教师模型的预测视为一次投票, 在投票的过程中对投票结果添加噪声, 算出得票最高的预测结果。然后再用教师标注的数据集训练学生模型, 最终使用学生模型进行预测服务。这样可以在不使用敏感数据的情况下间接训练公开模型, 并能够防止模型反演攻击对原始数据的窃取。最近的研究里, Xu 等人[77]提出了一种满足差分隐私的生成对抗网络(GAN), 可以通过在学习过程中为梯度添加精心设计的噪声来实现 GAN 下的

差分隐私, 并证明了其在实际隐私预算下可以产生高质量的生成数据。Yu 等人[78]分析了文献[75]中数据批处理方法造成的隐私成本的低估, 使用称为集中差分隐私(CDP)的差分隐私方案, 实施了几种动态隐私预算分配技术, 提高了现有统一预算分配方案的模型准确性。Wu 等[79]在病理图像分类中, 引入了一种新的随机梯度下降方案, 将精心设计的噪声注入每一步的更新迭代中, 配备了策略可以自适应地控制注入噪声的规模, 并对差分隐私下的隐私成本进行了严格的分析。

在联合分布式学习中, 不同参与方在各自数据集上独立训练模型, 共享训练结果。2015 年, Shokri 等人[80]将隐私保护概念首次引入深度学习中, 他们设计了一个能够使多方协作完成训练的神经网络模型, 采用梯度下降算法, 在训练期间选择性地共享参数, 而无需共享其输入数据集, 完成了联合分布平台下的隐私保护。2016 年, Liu 等[81]在此基础上, 构建了 XMPP 服务器和多个移动设备上实现的隐私保护移动分布式环境。2017 年, Phong 等[82]针对文献[80]进行分析, 认为即便一小部分梯度信息也会被敌手通过模型反演的方式间接泄漏用户数据隐私, 他们为了对梯度信息进行保护, 在上传参数时, 使用了加法同态进行加密。

表 5 列出了以上一些文献的研究中使用差分隐私技术对机器学习算法添加扰动的方式和优缺点。理论上任何算法都可以通过添加足够多的噪声来实现差分隐私, 在极端的例子下, 模型只输出噪声也是满足差分隐私的。所以对差分隐私的研究方向依然在于, 如何使一个满足差分隐私的算法更具有可用性。

表 5 机器学习添加扰动方式

Table 5 Methods of adding perturbation to machine learning

添加扰动方式	优点	缺点
对数据[68]	容易实现, 有较高的准确度	不易控制隐私预算参数分配
对模型权重[69]	保护模型的隐私	分类精度降低
对代价函数[70]	提高了精度	必须要特定的代价函数
对代价函数系数[71]	具有较高效率和可用性	只能应用于线性表示的目标函数
对梯度[75]	可控隐私成本完成深度网络训练	隐私损失低估
对 PATE 框架的投票标签[76]	模型对隐私和学习有良好的兼顾	teacher 的共识度影响结果
对聚类中心点[74]	提高了聚类结果的可用性	没有考虑隔离点对初始中心点的影响, 可能导致聚类结果不稳定

4.2.3 其他隐私保护技术

同态加密和差分隐私是机器学习隐私保护中应用最广泛的技术。除此之外, 在有些场景的安全多方计算协议构建中, 一些密码学的技术如不经意传输、乱码电路、秘密共享也得到了应用。

Nikolaenko 等[83]使用乱码电路构建了一个具有隐私保护的岭回归系统, 并在具有数百万个样本的数据集上对其进行评估, 系统以明文输出最佳拟合曲线, 但不显示有关输入数据的其他信息。Rouhani 等[84]提出了一个在云服务中可以保护各方隐私的框架, 框架的安全多方计算使用乱码电路协议执行, 并引入了低开销的预处理技术, 降低了系统的整体运行时间。Makri 等[85]提出了一种基于支持向量机学习的高效隐私保护图像分类系统, 使用附加秘密共享技术, 每个 MPC 服务器都不了解双方的输入, 服务器通过一起执行 MPC 协议以产生最终的分类结果。

这些研究中, 研究者们在构建安全多方计算协议时应用到了乱码电路或秘密共享的密码学技术, 这些技术同样存在一定的局限性, 乱码电路的引入会导致模型训练的高开销, 同时许多安全多方计算协议的设计要考虑特定的场景, 对数据集和模型有一定的要求。在实际应用中, 同态加密、差分隐私以及其它一些隐私保护技术常常根据特定情况结合使用, 以更好的构建更适合的隐私保护协议。

5 结束语

本文对机器学习中的隐私问题及相关隐私保护技术进行了分类和介绍, 也清晰地展示了研究人员的研究进展和研究思路。当前机器学习应用广泛, 其隐私问题不容忽视。对于使用服务的用户而言, 他们担心自身的敏感数据被泄漏; 对于一些不法的攻击者而言, 攻击者会主动采取模型反演攻击、成员推理攻击企图获取用户的数据; 于服务商而言, 他们担心自身的模型被窃取。在解决机器学习隐私问题中, 同态加密技术和差分隐私技术可以起到很好的防范

效果。

同态加密是一个快速发展的领域, 但是它的实际限制意味着现有技术不能总是直接转换成相应的安全算法。目前针对同态加密与机器学习模型相结合的过程中, 研究目标是以下三点:

1) 解决其噪声增长导致的无法解密, 使之成为真正意义上的全同态。

2) 解决除加法和乘法以外的复杂运算问题。

3) 解决在实际应用中的低效率问题。

未来研究角度分为以下三个方面:

1) 从同态加密算法本身研究, 设计出在保留同态加密属性的前提下更好地与机器学习模型相结合的同态加密算法, 使其具有更强大的同态运算能力。

2) 从机器学习的算法及模型展开研究, 调整模型的结构或相关参数, 使其更好地适用同态加密算法, 提高模型在实际应用中的准确度。

3) 从协议角度入手, 对于特定场景, 在使用同态加密算法时, 设计执行起来具有更高效率的协议。

差分隐私是一种灵活、通用的隐私保护技术, 具有强大的数学理论支撑[33]。许多机器学习算法与差分隐私结合的研究都得到了实现。但在满足差分隐私的条件下, 数据的可用性与整个系统的效率是目前应用中存在的问题, 研究差分隐私的目标在于使得隐私保护数据具有更好的可用性以及更高的运算效率。所以未来对于差分隐私的研究角度为以下两个方面:

1) 从数学理论角度入手, 设计具有更合适分布的噪声来契合机器学习模型的参数和数据。

2) 从与机器学习的结合入手, 改变添加随机扰动的对象来取得更好的可用性与更高的效率。

除同态加密与差分隐私外, 一些密码学的技术如秘密共享、不经意传输、乱码电路也被应用于保护机器学习隐私中, 由于受不同场景的限制, 这些技术往往不会单独应用, 而是与同态加密和差分隐私结合起来共同构建符合应用场景的安全多方计算协议, 以实现隐私、准确度、效率这三个目标。

参考文献

[1] A. Ghorbel, M. Ghorbel, and M. Jmaiel, "Privacy in cloud computing environments: a survey and research challenges," *The Journal of Supercomputing*, vol. 73, no. 6, pp. 2763-2800, Jun. 2017.

[2] L. Song, C.G. Ma, and G.H. Duan, "Machine learning security and privacy: a survey," *Chinese Journal of Network and Information Security*, vol. 4, no. 8, pp. 5-15, Aug. 2018.
(宋蕾, 马春光, 段广晗, "机器学习安全及隐私保护研究进展", *网络与信息安全学报*, 2018, 4(8):5-15。)

[3] J.J. Cui, J. Long, E.X. Min, Y. Yu, and J.P. Yin, "Survey on Application of Homomorphic Encryption in Encrypted Machine Learning," *COMPUTER SCIENCE*, vol. 45, no. 4, Apr. 2018.
(崔建京, 龙军, 闵尔学, 于洋, 殷建平, "同态加密在加密机器学习中的应用研究综述", *计算机科学*, 2018, 45(4): 46-52。)

[4] Y.C. Yu, L. Ding, and Z.N. Chen, "Research on Attacks and Defenses towards Machine Learning Systems," *Netinfo Security*, vol. 18, no. 9, pp. 16-24, 2018.
(于颖超, 丁琳, 陈左宁, "机器学习系统面临的安全攻击及其防御技术研究", *信息网络安全*, 2018, 213(9): 16-24。)

[5] N. Papernot, P.D. McDaniel, A. Sinha, and M.P. Wellman, "Towards the Science of Security and Privacy in Machine Learning," arXiv preprint arXiv, 1611.03814, 2016.

[6] H. Bae, J. Jang, D. Jung, H. Jang, H. Ha, and S. Yoon, "Security and Privacy Issues in Deep Learning," arXiv preprint arXiv, 1807.11655, 2018.

[7] Y.B. Yan, and Y.Y. Chen, "A survey on machine learning and its main strategy," *Application Research of Computers*, vol. 21, no. 7, pp. 4-10, 2004.

[8] D.Y. Chen, and H. Zhao, "Data Security and Privacy Protection Issues in Cloud Computing," *International Conference on Computer Science and Electronics Engineering* (*ICCSEE*), pp. 647-651, 2012.

[9] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning," *ACM Conference on Computer and Communications Security* (*CCS*), pp. 603-618, 2017.

[10] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Informationand Basic Countermeasures," *ACM Conference on Computer and Communications Security* (*CCS*), pp. 1322-1333, 2015.

[11] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," *USENIX Security Symposium*, pp. 17-32, 2014.

[12] G. Ateniese, L.V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifers, " *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137-150, 2015.

[13] X. Wu, M. Fredrikson, S. Jha, and J.F. Naughton, "A Methodology for Formalizing Model-Inversion Attacks," *2016 IEEE 29th Computer Security Foundations Symposium* (*CSF*), pp. 355-370, 2016.

[14] C.Z. Song, T. Ristenpart, and V. Shmatikov, "Machine Learning Models that Remember Too Much," *ACM Conference on Com-*

puter and Communications Security (*CCS*), pp. 587-601, 2017.

[15] N. Carlini, C. Liu, J. Kos, U. Erlingsson, and D. Song, "The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets," *Computer Research Repository* abs/1802.08232, 2018.

[16] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model Inversion Attacks for Prediction Systems: Without Knowledge of Non-Sensitive Attributes," *2017 15th Annual Conference on Privacy, Security and Trust* (*PST*), pp. 115-126, 2017.

[17] R. Shokri, M. Stronati, C.Z. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," *IEEE Symposium on Security and Privacy* (*SP'17*), pp. 3-18, 2017.

[18] B. Hitaj, G. Ateniese, and Fe. Pérez-Cruz, "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning," *ACM Conference on Computer and Communications Security* (*CCS*), pp. 603-618, 2017.

[19] L. Melis, C.Z. Song, E.D. Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning". IEEE 2019.

[20] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W.Q. Wei, "Towards Demystifying Membership Inference Attacks," *Computer Research Repository* abs/1807.09173, 2018.

[21] J. Hayes, L. Melis, G. Danezis, and E.D. Cristofaro, "LOGAN: Membership Inference Attacks Against Generative Models," *Proceedings on Privacy Enhancing Technologies*, no. 1, pp. 133-152, 2019.

[22] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Je´gou, "White-box vs Black-box: Bayes Optimal Strategies for Membership Inference," *International Conference on Machine Learning* (*ICML*), pp. 5558-5567, 2019.

[23] F. Tramèr, F. Zhang, A. Juels, M.K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," *USENIX Security Symposium*, pp. 601-618, 2016.

[24] H. Dang, Y. Huang, and E.C. Chang, "Evading Classifiers by Morphing in the Dark," *ACM Conference on Computer and Communications Security* (*CCS*), pp. 119-133, 2017.

[25] N. Papernot, P.D. McDaniel, I.J. Goodfellow, S. Jha, Z.B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (*AsiaCCS*), pp. 506-519, 2017.

[26] B. Wang, and N.Z. Gong, "Stealing Hyperparameters in Machine Learning," *IEEE Symposium on Security and Privacy* (*SP'18*), pp. 36-52, 2018.

[27] S.J. Oh, M. Augustin, M. Fritz, and B. Schiele, "Towards Reverse-Engineering Black-Box Neural Networks," arXiv preprint arXiv, 1711.01768, 2017.

[28] X. Hu, L. Liang, L. Deng, S.C. Li, X.F. Xie, Y. Ji, Y.F. Ding, C. Liu, T. Sherwood, and Y. Xie, "Neural Network Model Extraction Attacks in Edge Devices by Hearing Architectural Hints," arXiv preprint arXiv, 1903.03916, 2019.

[29] T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 4954-4963, 2019.

[30] P. Samarati, and L. Sweeney, "Generalizing Data to Provide Anonymity when Disclosing Information," *PODS*, vol. 98, p. 188, 1998.

[31] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," *22nd International Conference on Data Engineering* (*ICDE'06*), pp. 24-24, 2006.

[32] N.H. Li, T.C. Li, and S. Venkatasubramanian, "Privacy Beyond k-Anonymity and l-Diversity," *22nd International Conference on Data Engineering* (*ICDE'07*), pp. 106-115, 2007.

[33] C. Dwork, F. McSherry, K. Nissim, and A.D. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," *Theory of cryptography conference* (*TCC*), PP. 265-284, 2006.

[34] A.C.C Yao, "Protocols for Secure Computations," *FOCS*, vol. 82, pp. 160-164, 1982.

[35] R. Rivest, L. Adleman, and M.L. Dertouzos, "On data banks and privacy homomorphisms," *Foundations of secure computation*, vol. 4, no. 11, pp. 169-180, 1978.

[36] T.E. Gamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469-472, 1985.

[37] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 223-238, 1999.

[38] C. Gentry, "Fully homomorphic encryption using ideal lattices," *Stoc*. vol. 9. no. 2009. pp. 169-178, 2009.

[39] W.L. Du, and Z.J. Zhan, "Building decision tree classifier on private data," *IEEE International Conference on Privacy Australian Computer Society*, vol. 14, pp. 1-8, 2002.

[40] J.Z. Zhan, "Using Homomorphic Encryption for Privacy-Preserving Collaborative Decision Tree Classification," *2007 IEEE Symposium on Computational Intelligence and Data Mining* (*CIDM*) pp. 637-645, 2007.

[41] L.J.M. Aslett, P.M. Esperança, and C.C. Holmes, "Encrypted statistical machine learning: new privacy preserving methods," *computer science*, 2015.

[42] D.J. Wu, T. Feng, M. Naehrig, and K.E. Lauter, "Privately Evaluating Decision Trees and Random Forests," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 4, pp. 335-355, 2016.

[43] A. Alabdulkarim, M. Al-Rodhaan, T. Ma, and Y. Tian, "PPSDT: A Novel Privacy-Preserving Single Decision Tree Algorithm for Clinical Decision-Support Systems Using IoT Devices," *Sensors*, vol. 19, no. 1, pp. 142, 2019.

[44] X.H. Liu, "Study on the Algorithms of Privacy Preserving Support Vector Machine," [MA.Sc. dissertation] Shandong University of Science and Technology, 2011.
(刘晓红, "隐私保护支持向量机的算法研究", 山东科技大学, 2011。)

[45] Y. Rahulamathavan, R.C. Phan, S. Veluru, K. Cumanan, and M. Rajarajan, "Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud," *IEEE Trans. Dependable Sec. Comput*, vol. 11, no. 5, pp. 467-479, 2014.

[46] R. Bost, R.A. Popa, S. Tu, and S. Goldwasser, "Machine Learning Classification over Encrypted Data," *The Network and Distributed System Security Symposium* (*NDSS*), pp. 4324-4325, 2015.

[47] A. Wood, V. Shpilrain, K. Najarian, A. Mostashari, and D. Kahrobaei, "Private-Key Fully Homomorphic Encryption for Private Classification," *International Congress on Mathematical Software* (*ICMS*), pp 475-481, 2018.

[48] X.D. Zhu, H. Li, and F.H. Li, "Privacy-preserving logistic regression outsourcing in cloud computing," *International Journal of Grid and Utility Computing*, vol. 4, no. 2-3, pp. 144-150, 2013.

[49] Y. Aono, T. Hayashi, L.T. Phong, and L.H. Wang, "Scalable and Secure Logistic Regression via Homomorphic Encryption," *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy* (*CODASPY*), pp. 142-144, 2016.

[50] M. Kim, Y. Song, S. Wang, Y.H. Xia, and X.Q. Jiang, "Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation," *JMIR medical informatics*, vol. 6, no. 2, pp. e19, 2018.

[51] Y.C. Yao, L. Song, and C. E, "Investigation on distributed k-means clustering algorithm of homomorphic encryption," *Computer Technology and Development*, vol. 27, no. 2, pp. 81-85, 2017.

[52] X. Meng, D. Papadopoulos, A. Oprea, and N. Triandopoulos, "Privacy-Preserving Hierarchical Clustering: Formal Security and Efficient Approximation," *Computer Research Repository* abs/1904.04475, 2019.

[53] M. Barni, C. Orlandi, and A. Piva, "A privacy-preserving protocol for neural-network-based computation," *Proceedings of the 8th workshop on Multimedia and security* (*MM&Sec*), pp. 146-151, 2006.

[54] C. Orlandi, A. Piva, and M. Barni, "Oblivious Neural Network Computing via Homomorphic Encryption," *EURASIP Journal on Information Security*, vol. 2007, no. 1, pp. 1-11, 2007.

[55] I. Damgård, and M. Jurik, "A generalisation, a simpli. cation and some applications of paillier's probabilistic public-key system," *International Workshop on Public Key Cryptography* (*PKC'01*), pp. 119-136, 2001.

[56] R. Gilad-Bachrach, N. Dowlin, K. Laine, K.E. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," *International Conference on Machine Learning* (*ICML*), pp. 201-210, 2016.

[57] H. Chabanne, A.D. Wargny, J. Milgram, C. Morel, and E. Prouff, "Privacy-Preserving Classification on Deep Neural Network," *IACR Cryptology ePrint Archive*, vol. 2017, pp. 35, 2017.

[58] E. Chou, J. Beal, D. Levy, S. Yeung, A. Haque, and F.F. Li, "Faster CryptoNets: Leveraging Sparsity for Real-World Encrypted Inference," *Computer Research Repository* abs/1811.09953, 2018.

[59] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa, "Oblivious Multi-Party Machine Learning on Trusted Processors," *USENIX Security Symposium*, pp. 619-636, 2016.

[60] Z.F. Lu, and J. Li, "FHE* KDFRS: FHE-compatible kernel based face recognition system" *Journal of Yunnan University*, vol. 40, no. 6, pp. 1116-1127, 2018.
(陆正福, 李佳, "FHE*KDFRS:全同态加密相容的核基人脸识别系统", *云南大学学报*, 2018, 40(6): 1116-1127。)

[61] P. Li, J. Li, Z.A. Huang, T. Li, C.Z. Gao, S.M. Yiu, and K. Chen, "Multi-key privacy-preserving deep learning in cloud computing," *Future Generation Comp. Syst*, vol. 74, pp. 76-85, 2017.

[62] P.T. Xie, M. Bilenko, T. Finley, R. Gilad-Bachrach, K.E. Lauter, and M. Naehrig, "Crypto-Nets: Neural Networks over Encrypted Data," *computer science*, 2014.

[63] Stone, and H. Marshall, "The generalized weierstrass approximation theorem," *Mathematics Magazine*, vol. 21, no. 5, pp. 167-184, 1948.

[64] Q.C. Zhang, L.T. Yang, and Z.K. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1351-1362, 2015.

[65] A. Wiesberg, "Machine learning on encrypted data," University of Mannheim, Germany, 2018.

[66] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the SuLQ framework," *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (*PODS*), pp. 128-138, 2005.

[67] A. Friedman, and A. Schuster, "Data mining with differential privacy," Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), pp. 493-502, 2010.

[68] N. Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu, "Differentially private data release for data mining," *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (*KDD*), pp. 493-501, 2011.

[69] A.D. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," *Proceedings of the forty-third annual ACM symposium on Theory of computing* (*STOC*), pp. 813-822, 2011.

[70] J. Lei, "Differentially Private M-Estimators," *Advances in Neural Information Processing Systems* (*NIPS*), pp. 361-369, 2011.

[71] J. Zhang, Z.J. Zhang, X.K. Xiao, Y. Yang, and M. Winslett, "Functional Mechanism: Regression Analysis under Differential Privacy," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1364-1375, 2012.

[72] K. Nissim, S. Raskhodnikova, and A.D. Smith, "Smooth sensitivity and sampling in private data analysis," *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (*STOC*), pp. 75-84, 2007.

[73] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86-95, 2011.

[74] Y.L. Zhang, N. Liu, and S.P. Wang, "A differential privacy protecting K-means clustering algorithm based on contour coefficients," https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0206832, 2018.

[75] M. Abadi, A. Chu, I.J. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (*CCS*), pp. 308-318, 2016.

[76] N. Papernot, M. Abadi, Ú. Erlingsson, I.J. Goodfellow, and K. Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," *the international conference on Learning Representations* (ICLR), 2017.

[77] C.G. Xu, J. Ren, D.Y. Zhang, Y.X. Zhang, Z. Qin, and K. Ren, "GANobfuscator: Mitigating Information Leakage Under GAN via Differential Privacy," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2358-2371, 2019.

[78] L. Yu, L. Liu, C. Pu, M.E. Gursoy, and S. Truex, "Differentially Private Model Publishing for Deep Learning," *Computer Research Repository* abs/1904.02200, 2019.

[79] B.Z. Wu, S.W. Zhao, G.Y. Sun, X.L. Zhang, Z. Su, C.H. Zeng, and Z.H. Liu, "P3SGD: Patient Privacy Preserving SGD for Regularizing Deep CNNs in Pathological Image Classification," *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 2099-2108, 2019.

[80] R. Shokri, and V. Shmatikov, "Privacy-Preserving Deep Learning," *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (*CCS*), pp. 1310-1321, 2015.

[81] M. Liu, H.T. Jiang, J. Chen, A. Badokhon, X.T. Wei, and M.C. Huang, "A Collaborative Privacy-Preserving Deep Learning System in Distributed Mobile Environment," *2016 International Conference on Computational Science and Computational Intelligence* (*CSCI*), pp. 192-197, 2016.

[82] L.T. Phong, Y. Aono, T. Hayashi, L.H. Wang, and S. Moriai, "Privacy-Preserving Deep Learning via Additively Homomorphic Encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333-1345, 2017.

[83] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft, "Privacy-preserving ridge regression on hundreds of millions of records," *IEEE Symposium on Security and Privacy* (*S&P*), pp. 334-348, 2013.

[84] B.D. Rouhani, M.S. Riazi, and F. Koushanfar, "Deepsecure: Scalable provably-secure deep learning," *Proceedings of the 55th Annual Design Automation Conference. ACM*, no. 2, 2018.

[85] E. Makri, D. Rotaru, N.P. Smart, and F. Vercauteren, "EPIC: efficient private image classification (or: learning from the masters)," *Cryptographers' Track at the RSA Conference* (*CT-RSA*), pp. 473-492, 2019.

赵镇东 于2018年在长沙理工大学网络工程专业获得学士学位。现在北京交通大学网络空间安全专业攻读硕士学位。研究领域为机器学习隐私保护。研究兴趣包括：同态加密。Email: 18120490@bjtu.edu.cn

王逸翔 于2018年在北京交通大学信息安全专业获得学士学位。现在北京交通大学网络空间安全专业攻读博士学位。研究领域为对抗机器学习。研究兴趣包括：机器学习，对抗样本。Email: 18112047@bjtu.edu.cn

常晓林 于2005年在香港科技大学计算机科学技术专业获得博士学位。现任北京交通大学计算机与信息技术学院教授。研究领域包括：网络空间安全和人工智能安全。Email: xlchang@bjtu.edu.cn