软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(2):406−423 [doi: 10.13328/j.cnki.jos.006147]

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

机器学习安全攻击与防御机制研究进展和未来挑战*

李欣姣[1,2], 吴国伟[1,2], 姚 琳[1,3], 张伟哲[4], 张 宾[3]

[1](大连理工大学 软件学院,辽宁 大连 116620)

[2](辽宁省泛在网络与服务软件重点实验室(大连理工大学),辽宁 大连 116620)

[3](鹏城实验室 网络空间安全中心,广东 深圳 518055)

[4](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 吴国伟, E-mail: wgwdut@dlut.edu.cn

摘 要: 机器学习的应用遍及人工智能的各个领域,但因存储和传输安全问题以及机器学习算法本身的缺陷,机器学习面临多种面向安全和隐私的攻击.基于攻击发生的位置和时序对机器学习中的安全和隐私攻击进行分类,分析和总结了数据投毒攻击、对抗样本攻击、数据窃取攻击和询问攻击等产生的原因和攻击方法,并介绍和分析了现有的安全防御机制.最后,展望了安全机器学习未来的研究挑战和方向.

关键词: 机器学习;安全和隐私;攻击分类;防御机制

中图法分类号: TP18

中文引用格式: 李欣姣,吴国伟,姚琳,张伟哲,张宾.机器学习安全攻击与防御机制研究进展和未来挑战.软件学报,2021,32(2):406−423. http://www.jos.org.cn/1000-9825/6147.htm

英文引用格式: Li XJ, Wu GW, Yao L, Zhang WZ, Zhang B. Progress and future challenges of security attacks and defense mechanisms in machine learning. Ruan Jian Xue Bao/Journal of Software, 2021,32(2):406−423 (in Chinese). http://www.jos.org.cn/1000-9825/6147.htm

Progress and Future Challenges of Security Attacks and Defense Mechanisms in Machine Learning

LI Xin-Jiao[1,2], WU Guo-Wei[1,2], YAO Lin[1,3], ZHANG Wei-Zhe[4], ZHANG Bin[3]

[1](School of Software Technology, Dalian University of Technology, Dalian 116620, China)

[2](Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province (Dalian University of Technology), Dalian 116620, China)

[3](Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen 518055, China)

[4](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Machine learning applications span all areas of artificial intelligence, but due to storage and transmission security issues and the flaws of machine learning algorithms themselves, machine learning faces a variety of security- and privacy-oriented attacks. This survey classifies the security and privacy attacks based on the location and timing of attacks in machine learning, and analyzes the causes and attack methods of data poisoning attacks, adversary attacks, data stealing attacks, and querying attacks. Furthermore, the existing security defense mechanisms are summarized. Finally, a perspective of future work and challenges in this research area are discussed.

Key words: machine learning; security and privacy; attack classification; defense mechanism

* 基金项目: 国家自然科学基金(61872053); 中央高校基本科研业务费专项资金(DUT19GJ204); 广东省重点领域研发计划(2019B010136001); 广东省重点科技计划(LZC0023)

Foundation item: National Natural Science Foundation of China (61872053); Fundamental Research Funds for the Central Universities (DUT19GJ204); Key-Area Research and Development Program of Guangdong Province (2019B010136001); Key Science and Technology Program of Guangdong Province (LZC0023)

收稿时间: 2019-08-12; 采用时间: 2019-12-01; jos 在线出版时间: 2020-10-12

机器学习当前被应用在各个领域,如恶意检测、图像识别分类、语音指令识别、自动驾驶、推荐系统、医疗系统等等.但是机器学习的安全和隐私问题随着其应用的推广日渐突出,成为阻碍其发展的重要因素.机器学习面临的攻击会导致机器学习算法分类出错、计算出错(如将恶意软件设别为正常软件导致木马攻击、在自动驾驶中计算出错导致交通事故等),从而降低机器学习算法的可信度.同时,机器学习的训练数据往往包含用户隐私数据(如健康数据和位置信息、身份数据和图像内容等),用户希望在保证隐私的条件下进行训练,但面向隐私的攻击会导致用户数据的隐私泄露(如攻击者基于推理结果分析或计算出用户的隐私数据),从而降低机器学习算法的隐私性.因此,保证机器学习算法的安全性和隐私性是机器学习发展的重要课题.

保证机器学习的安全性,指保证模型面临安全攻击仍能推理出正确结果的能力.保证机器学习的隐私性,指保证模型面临隐私攻击仍能保证模型数据、训练数据和由此引申出的用户隐私数据不被泄露的能力.从使用对抗训练和防御精馏来提高模型安全性,到使用加密和扰动来提高模型隐私性,虽然机器学习的安全问题从被发现和提出到现在发展时间较短,但已经取得了一定的进展.本文从机器学习面临的攻击出发,基于攻击发生的位置和时序分类,结合攻击者能力分析了各种攻击产生的原因,并介绍了现有的解决方案.

本文第 1 节介绍机器学习及其遭受攻击的原因,并按照位置和时序对攻击进行分类.第 2 节详细介绍机器学习面临的攻击类型、攻击产生的原因和攻击手段.第 3 节阐述机器学习现有的安全机制.最后给出现有机器学习安全机制存在的问题和未来的方向.

1 机器学习与安全威胁

1.1 机器学习简介及分类

机器学习是人工智能发展至今最重要的学科,旨在利用数据和经验改进智能算法的性能.机器学习的目标是从给定的训练集中学习到一个模型,当新的未知数据到来时,根据这个模型预测结果.机器学习算法的训练集称为一组样本,每个样本包含标签和一组特征值,每个样本可以由特征值计算一个特征向量,机器学习算法用训练样本的特征向量和标签构建模型的过程称为训练或学习的过程,使用这个模型对新的测试样本预测标签的过程称为推理过程.模型能够正确推理出测试样本标签的能力是模型泛化性能的指标.

机器学习按照形式可以分为监督学习、无监督学习和强化学习(如图 1 所示).监督学习的训练数据带有人为标注的标签,主要应用于分类和预测任务,常见于统计分类和回归分析,如垃圾邮件分类、房价预测等.在分类学习中,样本归属于两个或多个离散的类,模型的目标是将新的样本划分到这些类中.分类模型的构建可以通过寻找不同类之间的超平面,或使用支持向量机(SVM)、神经网络或逻辑回归等方法实现.回归模型指训练标签为连续值的训练样本产生的模型,通过拟合一个与训练样本的数据距离最近的模型(通常为一条曲线),可以实现对新的测试数据基于特征值的预测.当训练数据难以分类或人工标注类别成本太高时,使用机器学习解决模式识别中的各种问题称为无监督学习,又称为归纳性学习.无监督学习常见于聚类和数据降维,通过循环和递减运算来减小误差,达到分类的目的,*K*-means 是最常用的聚类方法之一.强化学习模型是基于奖励值学习状态和动作之间映射的模型,主要应用于决策选择中,如 AlphaGO[1]围棋机器人的决策.强化学习模型根据动作的奖励值调整其权重,等新的状态到来时,采取奖励值最高的策略.

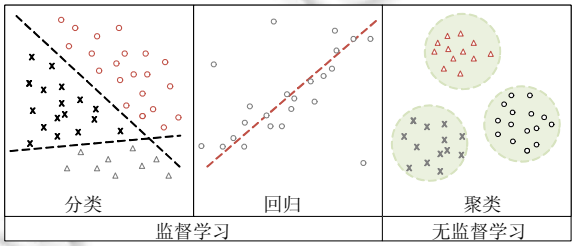


Fig.1 Classification of machine learning

图 1 机器学习分类

1.2 机器学习面临的安全威胁和攻击

机器学习的过程由数据提供者、机器学习算法及模型训练和机器学习的服务使用者三方参与.当数据提供者既提供数据又训练和使用模型时,模型和数据安全得到一定保障,但往往机器学习的参与方是分离的,因而硬件设备存储的安全性、网络数据传播的安全性、学习算法模型的安全性都是决定机器学习安全性的重要因素.

终端设备用户在使用设备时会产生大量关于姓名、位置、喜好甚至医疗记录等隐私相关的数据,这些数据从设备发出便脱离了数据提供者的掌控,因而在机器学习的过程中,数据提供者设备的安全性、机器学习计算方法的可信性、服务使用者的询问和计算能力都给数据提供者的隐私性带来隐患.攻击者除利用存储和传输安全窃取用户隐私外,还可以作为服务使用者进行询问攻击,结合询问统计信息和背景知识,对用户隐私发动模型提取攻击、去匿名化攻击和成员推理攻击等.

数据拥有者在提交数据到机器学习计算方法的过程中,攻击者通过对数据进行窃取或污染,可以改变模型训练的结果,降低模型的预测正确率.当攻击者是服务使用者时,可以根据询问结果和对模型的背景知识,提取或重构模型参数,进而通过推理训练数据集或构造对抗样本来对模型进行多种攻击.

结合机器学习的安全威胁,部分文献[2,3]首先对机器学习中存在的安全问题进行了总结和讨论;随后,文献[4–11]对更多模型的安全性进行了越来越系统的分析和讨论.

本文按照机器学习的学习和推理两个阶段分析其面临的攻击,如图 2 所示.根据攻击发生的逻辑和时序,将机器学习面临的攻击分为训练数据面临的数据投毒攻击、测试数据面临的对抗样本攻击、训练和测试数据的数据窃取攻击和由推理结果面临的询问攻击带来的成员测试攻击、数据窃取攻击、模型提取攻击.

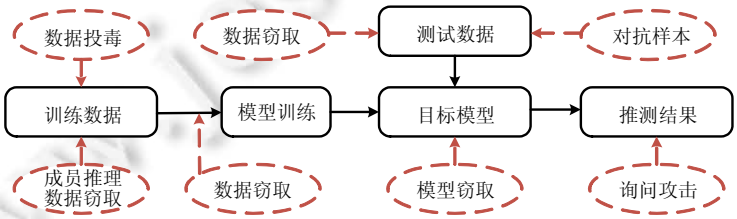


Fig.2 Attacks on machine learning

图 2 机器学习面临的攻击

2 机器学习面临的攻击

2.1 相关术语

首先简要列出一些在机器学习安全与隐私领域的术语.

- 对抗样本(adversarial example):为了让模型混淆出错而对原始样本经过精心扰动的样本.
- 对抗扰动(adversarial perturbation):为使原始样本称为对抗样本而增加的扰动.
- 对抗训练(adversarial training):使用原始训练集和对抗样本共同训练机器学习模型.
- 敌手(adversary):特指制作对抗样本的攻击者.
- 白盒攻击(write-box attack):攻击者拥有目标模型全部知识的攻击,包括其参数值、模型结构、训练方法、训练数据等.
- 黑盒攻击(black-box attack):攻击者仅拥有模型有限知识的攻击,例如攻击者通过在训练阶段产生对抗样本进行对抗样本攻击.
- 检测器(detector):检测样本是否为对抗样本的机制.
- 出错率(fooling ratio):模型被攻击后的出错率.
- 靶向攻击(targeted attack):指定模型输出的攻击,如为对抗样本指定分类标签等.
- 威胁模型(threat model):模型可能遭受的攻击方式,例如黑盒攻击.

- 对抗样本转移性(transferability):对抗样本在其生成模型之外的有效性.
- 通用干扰(universal perturbation):使用于任意样本得到的对抗样本都能有效使模型出错的干扰.

2.2 攻击者分析

攻击者对机器学习发动不同攻击需要不同的访问权限和背景知识,因而我们根据攻击者能力的不同对其进行分析.攻击者能力的标准包括攻击者知识、攻击者目标和攻击者策略这 3 个维度[12].

(1) 攻击者知识.攻击者知识指攻击者对目标模型的背景知识,包括对训练集的背景知识、对训练集标签的背景知识、对模型算法的背景知识、对模型决策函数的背景知识、对训练完成的模型的背景知识和对训练结果的背景知识等;同时包括攻击者能否修改训练和测试数据和能否修改训练数据标签等.攻击者知识决定了攻击者能够发动的攻击类型.

(2) 攻击者目标.攻击者目标指攻击者要降低模型的性能指标,包括模型的隐私性指标和模型的正确性指标.模型的隐私性攻击目标指攻击者根据其能力推测的信息类型和数量,如去匿名化和成员推理攻击是推测用户身份信息,而模型提取是提取模型参数信息.模型的正确性攻击目标是攻击者要改变模型的分类边界使模型分类出错,并增大模型最小化的损失函数使模型出错率提高.

(3) 攻击者策略.攻击者策略指攻击者发动攻击的措施,例如使用何种对抗样本生成算法进行对抗攻击、通过何种修改手段、修改哪些数据发动污染攻击等.

2.3 训练数据投毒攻击

训练数据投毒攻击[10,13−17]也称为训练数据污染攻击,发生在训练阶段,攻击者具有获取、修改或创造训练数据集的能力,知晓训练数据集的标签等背景知识,攻击目标是训练数据集,旨在通过修改一定数量的训练数据,使模型训练到错误的对应关系,从而使模型训练出错.例如,在在线垃圾邮件过滤模型的训练过程中,通过产生大量投毒邮件,使垃圾邮件判别器无法进行正常判断,或使人脸识别系统出错[18]等.通过操纵模型输入[5]、修改训练数据[4]或使特征丢失和破坏[19],都能够发动投毒攻击(如图 3 所示).

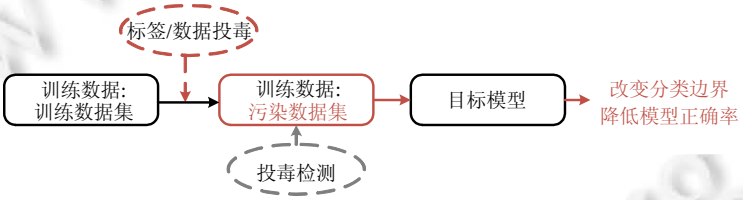


Fig.3 Data poisoning attack

图 3 数据投毒攻击

因为数据的收集、传输途径不可靠,或进入模型前的存储不安全,攻击者可以修改训练数据,对其发动投毒攻击,降低模型的正确性.PAC 模型理论[20]可以评估一个机器学习模型能否在一定范围内正确学习到预定内容,基于 PAC 理论,Kearns[4]指出:对于任意学习算法,训练数据的修改会影响模型的安全性,要想达到一定的学习准确率ε,修改训练数据的比例β需要满足$\beta \leqslant \varepsilon/(1+\varepsilon)$,即学习准确率与训练数据修改比例成反比关系,被修改数据的比例越小,模型的学习准确率越高.数据投毒攻击会修改部分训练数据的标签或直接改变部分模型输入,当达到一定比例时,模型的分类边界被修改,如图 4 所示,模型的安全性降低.训练数据有标签和数据两个部分,因而对训练数据的投毒攻击也可以分为对标签的投毒和对数据的投毒.

(1) 标签投毒

攻击者通过直接修改训练数据的标签信息,使训练数据对应到错误的标签,模型学习到错误的对应关系,在面对新的测试数据时偏离正常判断,准确率降低.在攻击者拥有训练数据访问权时,修改训练数据标签是很容易发动的攻击.Binggio 等人[21]在 SVM 分类训练机中,通过修改 40%的训练数据标签,使模型仅有 30%的准确性.

Mozaffarikermani 等人[22]对医疗数据标签进行修改,大大降低了模型学习的准确性.

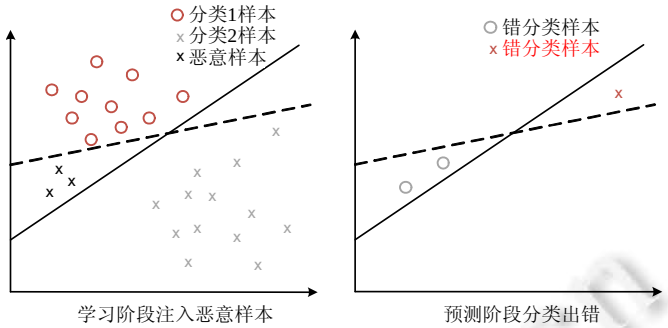


Fig.4 Change of classification boundaries by poisoning attack

图 4 投毒攻击改变分类边界

(2) 数据投毒

在训练数据进入模型之前,攻击者通过修改原有的训练数据或创造新的错误的数据发动污染攻击,使模型的准确率降低.根据机器学习模型输入数据的特点,可以将学习模型分为离线输入学习模型和定时更新在线输入的学习模型.

当模型为离线输入时,Biggio 等人[14]基于测试错误有固定公式的 SVM 学习算法,通过修改其训练数据,降低了训练精度;Mei 等人[23]针对图损失训练(包含线性或逻辑回归或 SVM 训练)和连续输入的模型,给出了更一般化的污染框架,指出:只要模型符合条件,就能找到最佳的污染方式.

对于在线的或需要定时更新的学习模型,Kloft 等人[24]指出:基于模型定时收集训练数据的规律,对新收集的数据进行污染以降低模型安全性;也指出,可以通过寻找使训练数据经验平均值差异最大的点来检测被污染数据.文献[25,26]也对在线学习模型的新输入数据进行了修改,降低了模型的安全性.

当攻击者无权获取预处理数据,可以间接污染训练数据时,Perdisci 等人[27]在蠕虫签名生成器的数据流中增加扰动,使其修改判别阈值降低签名识别的准确性,也属于数据投毒攻击.

2.4 对抗样本攻击

对抗样本攻击[6,28,29]的攻击者具有获取和修改测试样本的能力,知晓标签等背景知识,攻击目标是测试数据集,旨在通过构造对抗样本,使模型推测结果出错.对抗样本攻击发生在推理阶段.攻击者为实施攻击而特意构造的样本称为“对称样本”.对抗样本能够降低模型正确性,例如使垃圾邮件或恶意文档逃避检查,或使分类任务得到攻击者的目标分类.

对抗样本攻击首由 Szegedy 等人[5]提出和命名,发生在模型推理阶段,是攻击者根据其背景知识通过构造对抗样本,使模型结果出错的攻击.Binggio 等人[12]根据攻击者的目标、知识和能力提出了攻击者模型,对攻击者进行了不同的划分.Papernot 等人[8]根据攻击者的目标不同,对攻击者目标进行了分类.

基于攻击者能够获取的背景知识的不同,对抗样本攻击又分为白盒和黑盒两种:白盒攻击指攻击者能够获取训练机制或训练的参数,黑盒攻击指攻击者仅能多次查询或收集训练数据.白盒攻击中,攻击者通过获取模型的参数和数值,使用数学方法构造对抗样本,使模型得到攻击者想要的结果,使模型安全性降低.黑盒攻击中,攻击者通过多次查询,分析输入和输出的对应关系,或通过多次对抗尝试,修改测试数据,得到对抗样本,使模型得到攻击者想要的结果,降低模型的安全性.基于攻击者能否直接修改模型输入的不同,对抗样本攻击又可以分为精致构造的对抗样本攻击[5]和物理世界的数据管道对抗样本攻击[30,31].基于攻击者攻击目的的不同,对抗样本攻击又可以分为对抗样本造成的逃逸攻击、对抗样本造成的靶向错误分类攻击和对抗样本造成的源/目标错误分类攻击[8].按照攻击者攻击手段的不同,对抗样本攻击又可以分为基于梯度的攻击(白盒)、基于分数的攻击(白盒)、基于迁移的攻击(黑盒)、基于决策的攻击(黑盒).

对抗样本攻击如图 5 所示.

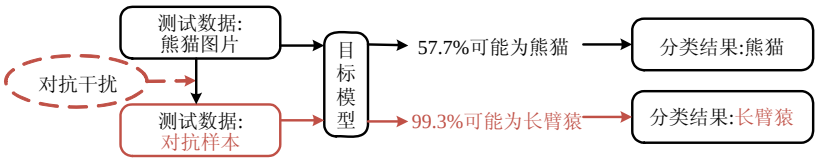


Fig.5 Adversary attack

图 5 对抗样本攻击

(1) 白盒攻击

根据攻击者想要达到的目的不同,白盒的对抗样本攻击可以分为错误分类、源/目标对应的错误分类和靶向错误分类[8]这 3 种:错误分类一般指二分类任务中,使结果出错的攻击,通常发生在恶意软件、邮件或文件的识别学习任务中;源/目标对应的错误分类一般指在多分类标签的分类任务中,通过优化算法修改原始数据得到最小修改的对抗样本,使其通过学习模型得到攻击者指定的错误分类,一般发生在图像识别的任务中,对抗样本能够得到次级可能性的标签分类;靶向错误分类攻击一般指攻击者生成的对抗样本对于人类而言是无意义的,但通过学习模型能够得到攻击者指定的分类,一般发生在图像识别或语音识别中,无意义的图像和语音得到了明确的分类和识别.

- 错误分类攻击

错误分类攻击也称逃逸攻击,攻击目标是二分类的学习模型,目的是给训练数据或原始数据增加扰动,使学习到的结果偏离原始学习结果.Biggio 等人[12]指出:在机器学习中,攻击者与学习者的损失函数是对抗性的,寻找最小扰动的对抗样本可以进一步转换为最优化问题,使用优化理论,通过改动最小数据量,达到篡改分类结果的目的,使恶意邮件逃避检查,并给出了最优化改动的计算公式.Grosse 等人[32]使用文献[12]中的最优化方法进行恶意软件对抗样本的生成,并指出:在恶意软件检测学习模型中,输入数据的熵明显少于图形分类学习的输入.如何在不改变软件性能的情况下构造对抗样本,是一个相对困难的问题.恶意软件检测系统的分类与图形分类的区别在于输入熵更小,且输入是离散和受限的.文章提出了能够降低学习模型对改动敏感度的方案,但是只有非常依赖参数选择的再次训练(re-training)有明显效果.

- 源/目标对应的错误分类攻击

源/目标对应的错误分类攻击指针对具有特定输出的特定测试样本的对抗攻击,攻击者针对有特定分类的特定测试样本进行改动,得出明确错误分类的对抗样本.攻击利用的是机器学习模型在泛化时的缺陷,文献[5]以针对 MNIST 手写识别数据集为例,指出了机器学习模型存在的两个特点.

- 一是传统神经网络认为,最后一个隐含层的每一个神经元代表数据的一种语义特征,因而模型通过最大化激活某一神经元完成分类.但实验证明,随机的神经元组合表现出了相同的特点.因而,代表数据语义特征的是整个神经元激活空间而不是单个的神经元.
- 二是机器学习模型存在肉眼不可分辨的微小改动导致结果不一样的对抗样本,作者利用 L-BFGS[6]来求解最优化问题,计算得到了对抗样本.

Papernot 等人[8]也以 MNIST 数据集为例,针对所有的前向 DNN 提出了基于雅克比矩阵即前向导数生成的对抗样本构造方法,减少影响的特征数量,仅改动 4.02%输入数据,就可达到 97%的攻击成功率,并给出了输入和输出的直接映射.Goodfellow 等人[6]针对文献[12]中提到的非线性模型易被细微修改的对抗样本攻击的问题,提出即使是线性模型也存在对抗样本,提出了 FGSM 算法.并给出了针对对抗样本的对抗训练方法(adversarial training),有效降低了欺骗成功率.Moosavidezfooli 等人[33]针对 Goodfellow 等人[6]提出的 FGSM 算法,提出了计算最小扰动ϵ的计算方法.Huang 等人[34]针对 Goodfellow 等人[6]提出的 FGSM 算法,提出了更快寻找扰动的方法,减少攻击需要添加的扰动.

- 靶向错误分类攻击

靶向错误分类攻击指具有特定错误的目标输出的对抗攻击,攻击者生成人类无法识别或认为是毫无意义的对抗样本,也可以对已经有正确分类的测试样本进行细微的修改,但却能够通过机器学习模型高概率、高可信度地分类到目标分类.Nguyen 等人[35]以 AlexNet 深度神经网络为模型,对 MNIST 数据库使用进化算法和梯度上升算法两种算法生成具有目标分类的图片,实验结果仅有 0.94%的出错率.Carlini 等人[36]指出:经由人工产生的人类无法识别的语音片段,可以被机器识别为指令的音频.

- 其他情况

在现实世界中,数据从各种摄像头或传感器进入学习模型,攻击者无法产生细粒度修改的对抗样本.在这种条件下,对抗样本攻击也是普遍存在的.Smith 等人[37]指出:攻击者通过多种途径捕捉合法的人脸图像,可以使人脸识别系统遭受重放攻击.Sharif 等人[38]通过对人脸图像的细微修改,使机器混淆无法识别,实现错误分类攻击.相比于文献[38],Kurakin 等人[28]通过科学观察的方法减少了干扰项,使计算更方便,且不固定修改像素的位置,可以少量修改所有像素.Sharif 等人[38]也指出:在通过摄像头等设备录入测试样本的情境中,捕捉到的镜片或其他设备上反射的图像可以作为对抗样本使分类结果出错,实现源/目标对应的错误分类攻击.

(2) 黑盒攻击

现有的机器学习云服务平台多给用户提供测试用的询问接口,攻击者只能通过观察测试数据输入模型后返回的结果进行攻击.黑盒攻击与白盒攻击不同,攻击者无法确定要干扰和错误分类到的目标分类,也无法获取训练数据和模型数据,因而无法根据背景知识设计生成最优对抗样本.最早针对机器学习的黑盒对抗样本攻击由 Lowd 等人[3]提出,在垃圾邮件过滤器以单词为变量判别垃圾邮件的条件下,通过多次问询,给变量增加标签,通过逆向工程(adversarial classifier reverse engineering)和使用成本函数来获得垃圾邮件通过过滤器的最小改动,实现了对抗样本攻击.同样的攻击模式也发生在文献[39,40]中.在回归机器学习任务中,Alfeld 等人[41]针对预测的回归模型,通过对模型的推理,修改输入数据,达到预期输出.

根据攻击者的知识,当攻击者的知识不同,黑盒攻击对模型的影响力也不相同.当攻击者已知模型标签输出概率时,攻击仅稍弱于白盒攻击.Xu 等人[42]针对已知模型的基于内容 PDFrate[43]和结构 Hidost[44]对恶意 pdf 进行分类的学习模型,提出只要向 pdf 中植入少量可执行代码就可以实现攻击,并给出了基于遗传编程的随机修改恶意软件生成对抗样本的方法.Rndic 等人[45]指出:主动防御方法只对特定攻击防御有效,攻击者利用其知识可以重建模型和训练数据,一旦攻击方法稍有变动,恶意 pdf 识别成功率降低,并给出了通过向 PDF 文件中增加被 pdf renderer 忽略的内容发起攻击的实例和实验报告[46].

(3) 对抗样本的转移性

对抗样本攻击不仅仅局限于攻击目标模型,它的影响面更广.Papernot 等人[47]指出:对抗样本具有转移性,即对抗样本在其生成模型之外的有效性.已知攻击者通过将自己生成的数据输入模型进行分类,并将模型输出作为标签,可以训练出一个代替模型,并通过代替模型生成对抗样本.利用转移性,将对抗样本返回原模型,可以实现对抗样本攻击.Papernot 等人[47]利用转移性发动对抗样本攻击,使原模型达到 82.24%的出错率.Papernot 等人[48]通过将 Amazon 训练的逻辑回归数据库错分类到 96%,证明利用对抗样本转移性的黑盒攻击可以推广到很多机器学习模型中.

2.5 数据窃取攻击

数据窃取攻击指通过存储和通信机制的漏洞、查询或反演技术等多种手段窃取机器学习隐私信息(如隐私的训练数据、模型的训练方法和训练参数)的攻击.数据窃取攻击针对机器学习的隐私性,大部分发生在黑盒攻击中,因此,攻击者仅具有窃取部分数据的能力(如图 6 所示).

训练阶段的数据窃取攻击是攻击者利用数据存储和数据传输的不安全性发动的攻击,如经由安全信道传输到云端服务器的隐私数据未经加密或采取其他安全措施[8,49,50],被攻击者窃取.推理阶段的数据窃取指因学习到模型后未及时删除隐私数据或用户的测试数据在进入模型前后被攻击者窃取,如指纹重构[51]、移动设备触摸手势重构[52]、人脸重放[37]等.

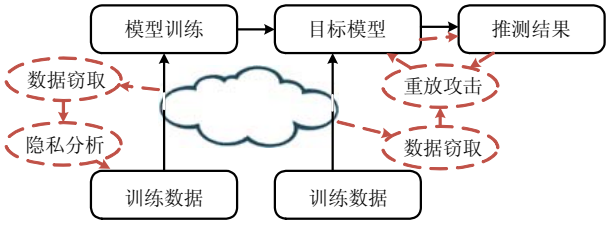


Fig.6 Data stealing attack

图 6 数据窃取攻击

2.6 隐私询问攻击

询问攻击指攻击者无法获取训练数据和模型数据,只能通过观察测试数据输入模型后返回的结果,即询问结果,进行计算和推测而发动的攻击.主要攻击类型有 3 种:成员推理攻击、训练数据提取攻击和模型提取攻击,如图 7 所示.

Fig.7 Querying attack

图 7 询问攻击

(1) 成员推理攻击

成员推理攻击的目标是训练数据的个体,攻击者根据询问结果判断出某个个体是否参与模型训练.成员推理攻击基于模型提取攻击、统计的综合数据和有噪声的真实数据来建立攻击模型,破坏需要保证训练数据隐私性的机器学习模型,Shokri 等人[53]利用成员推理攻击判断出某个疾病相关的数据是否参与了模型训练、破坏了模型的隐私性.

(2) 训练数据提取攻击

训练数据提取攻击的目标是训练数据的条目,是攻击者利用询问数据与已有知识推测训练数据隐私的攻击.Fredrikson 等人[54]指出:攻击者利用模型输出与某些特定属性的关联,可以推测训练数据的隐私信息,文章使用模型输出与人口统计学信息,在药物剂量预测模型中成功恢复训练数据的基因信息.Fredrikson 等人[55]进一步指出:模型反演攻击可以让攻击者提取模型输入,但相比于模型输入,数据与分类的统计信息更具有隐私性.在训练数据提取攻击中,攻击者利用大量询问的结果获得模型的分类和每个分类输出的概率,以此创建与模型相似的特征向量,每个特征向量代表某个类别的平均特征向量值,当某个类别仅有一个个体时,该个体隐私泄露.如人脸识别模型中,攻击者可以获取这个人的人脸信息.

(3) 模型提取攻击

模型提取攻击的目标是机器学习模型的数据,指攻击者利用询问接口获得模型的分类与测试输入输出数据,从而重构一个与原模型相似的模型的攻击.Vorobeychik 等人[56]在已知分类和问询权限的基础下,重构出了拥有相似训练数据量的模型.在已知模型类型、不知模型参数的情况下,Ateniese 等人[57]通过分析模型可以知道训练数据的某些统计属性,并利用这些统计属性构造一个新的模型,从而进一步实现训练数据提取攻击.Tramer 等人[58]指出:依靠询问接口,仅靠观察模型推测过程中的输入输出,对攻击者就可以提取模型信息,甚至构建一个相似的辅助模型用以进行进一步的攻击.

3 安全防御机制及分析

本节总结当前针对机器学习针对数据投毒攻击、对抗样本攻击、数据窃取攻击和隐私询问攻击的防御机

制,并分析这些机制的优缺点,如图 8 所示.根据针对的攻击不同,现有的安全机制主要分为正则化、对抗训练、防御精馏、模型隐私改造、加密和扰动.输入空间的正则化主要针对训练数据的污染攻击,而模型参数的正则化、对抗训练和防御精馏主要针对推测阶段的对抗样本攻击.模型隐私改造、加密和扰动主要用来防御由数据窃取和询问攻击带来的多种安全问题.

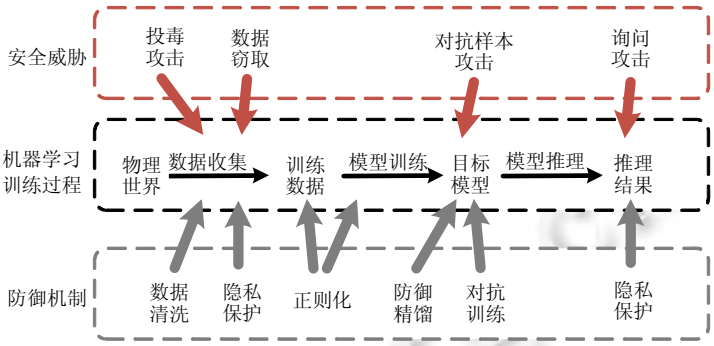


Fig.8 Security attacks and defense mechanisms in machine learning

图 8 机器学习中的安全攻击和防御机制

3.1 数据集和模型正则化

正则化是对机器学习进行规则化的过程,即通过对模型和模型输入的规范化操作,降低模型的出错率.正则化分为对训练数据的正则化和对模型的正则化:训练数据的正则化可以防御训练数据投毒攻击,而对模型的正则化可以防御对抗样本攻击.

(1) 输入正则化

在模型学习能力足够高而训练数据不足时,模型训练容易过拟合,从而在面对新的数据时出错率更高,因而容量足够的训练数据对模型训练至关重要;同时,在训练数据足够训练模型的情况下,攻击者可以通过污染训练数据的方式降低训练数据质量,从而使模型训练出错降低模型的准确性,因而要保证模型的性能,必须要保证训练数据的质量.高质量的训练数据有合理的特征空间和数据分布,同时有足够的数量.对训练数据的正则化可以理解为在保证训练数据存储安全的情况下,提升训练数据的质量.提升训练数据的质量称为数据集增强,即通过特征提取改变数据集的特征空间和数据分布,或通过注入噪声进行数据扩充,从而生成新的训练样本,创建具有更大容量甚至无限容量的增强数据集,从而提升模型的泛化能力.

机器学习认为:更多的训练数据可以降低模型出错率,使模型更具泛化能力.因而,扩充训练数据集是提高模型性能的重要手段.而攻击者通过修改训练数据(例如,在训练数据的垃圾邮件中增加某个词,而在测试数据的垃圾邮件中排除这个词的使用,可以规避检查)或测试数据(例如,垃圾邮件攻击者会增加积极的具有隐含意义的词语来逃避垃圾邮件过滤[40])的分布,就可以达到投毒攻击或对抗样本攻击的目的.这源于模型对未知数据不具有鲁棒性.

为了防御投毒攻击,多数集中式学习的防御机制建立在查找不在预期输入域内的样本上[59,60],以提升模型遇到未知数据的抵抗能力.Rubinstein 等人[61]指出:规范训练数据分布空间,可以减少投毒攻击超出模型输入预期.Biggio 等人[21]也使用正则化输入空间降低攻击者修改训练标签导致的逃逸攻击.而分布式学习的防御机制建立在查找参与者训练出的不在于期内的模型上[62,63].

(2) 模型正则化

模型正则化是利用正则化项对模型参数和训练方式进行规范化,进而提升模型泛化能力的过程.Barreno 等人[64]指出:使用去噪自动编码器(DAE)可以去除大部分对抗性噪声,但更小改动的新对抗样本对去噪自编码器和深度神经网络堆叠的网络依然攻击有效.因而提出了深度收缩网络(DCN),在损失函数中加入了平滑度惩罚,即正则化项,旨在最小化经验风险的同时,降低细微改动对模型输出的影响,以提高模型对对抗样本的鲁棒性.

- 参数正则化

模型参数正则化是利用正则化项,使模型参数满足某些约束的过程.机器学习中,训练数据特征向量对模型输出的影响受到模型参数数值的影响,模型参数的数量由特征向量决定,当参数数值为 0 时,代表该特征向量为噪声特征,模型参数数值的大小决定了对应特征向量对模型输出的影响大小.为了达到正则化的目的,降低数据改变对模型输出的影响,模型训练过程倾向于让参数数值尽可能稀疏(即非零参数尽可能少),各个参数数值尽可能小.

机器学习的损失函数是模型预测与真实结果的差异值,风险函数是损失函数的期望值.模型的损失函数越小,说明模型对训练数据学习的越充分.为了模型能够更好地学习到训练数据,需要损失函数足够小.为了在保证学习足够充分的同时达到输入正则化的目的,同时防止模型过拟合,在损失函数中加入正则化项$\lambda J(f)$来衡量模型的复杂度.目前常用的正则化项有 L_0,L_1 和 L_2 范式.L_0 范式要求参数数值总和要小于某个数值,L_1 范式为要求参数数值的绝对值总和在一定范围内以保证模型参数的稀疏性,L_2 范式要求参数数值的平方和在一定范围内以保证模型参数数值尽可能小.

损失函数:$L(Y,f(X))=(Y,f(X))^2$.

风险函数:$\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$.

最小化目标函数:$\min\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$.

- Dropout

Dropout[65]也是模型参数的一种正则化过程,它改变的不是参数的数值,而是参数的数量.机器学习的 ensemble 方法在提升模型泛化能力的同时带来了过量的计算代价,Dropout 认为,不训练多个模型也可以达到同样的效果.即通过在训练期间随机丢弃神经元及其连接来构造简化的网络,通过强迫神经元和其他随机挑选剩下来的神经元共同工作,减弱神经元之间的联合适应性,提高模型的泛化能力.而在测试阶段,一个简化的网络就可以逼近所有简化网络预测的平均结果.实验表明:Dropout 方法对数据集容量要求很高,在大容量数据集和增强数据集上能够获得很好的效果.

3.2 对抗训练机制

对抗样本攻击的产生,是机器学习模型输入的维度高而模型过于线性导致的,即是模型泛化能力不足,因而无法充分学习到训练数据和标签的映射关系.在一定的背景知识下,可以通过添加少量干扰产生对抗样本来跨越模型的决策边界,达到对抗攻击的目的.为了防御对抗样本攻击,除了提升训练数据质量外,研究从训练数据和模型改进两个方面,分别产生了对抗训练和防御精馏等安全模型.

对抗训练指使用对抗模型产生带有完全标注的对抗样本和合法样本混合起来对原模型进行训练,以提升模型鲁棒性的防御机制.Pinto 等人[66]提出了机器人互搏模型,一方尝试抓取物体,另一方尝试破坏其平衡性,从而使双方的模型得到优化.Kurakin 等人[10]指出:利用传递性,可以将小的数据库的对抗样本训练扩展到大数据库的对抗样本训练.Goodfellow 等人[67]利用对抗训练,将在 MNIST 数据集上的错误识别率从 89.4%降到 17.9%.Huang 等人[68]通过惩罚错分类的对抗样本,增加模型的顽健性.Tramèr 等人[69]提出了联合对抗训练(ensemble adversarial training),增加对抗样本多样性,但是也提出:在对抗训练中引入所有未知攻击的对抗样本是不现实的,对抗训练的非适应性导致对抗训练的局限性.

从对抗样本存在被提出,研究根据模型特征提出了 L-BFGS[6]、FGSM[35]、DeepFool[33]、Carlini-Wagner[70]等通过多种对抗样本生成算法,而 Xu 等人[42]提出,基于遗传算法可以不断产生新的对抗样本.对抗训练旨在将对抗样本和正确的标签关联学习,使模型能够学习到正确的映射关系.因而寻找更多的对抗样本生成算法,产生足量的对抗样本,可以有效抵御对抗样本攻击,提高模型性能.

3.3 防御精馏

精馏[71]是通过一个模型的输出训练另一个模型的机器学习算法,是在保证训练精度的条件下压缩模型的方法.防御精馏是 Papernot 等人[72]在精馏方案的基础上,通过两个相同模型之间的训练,达到梯度掩码[73],从而增强模型面对对抗样本的鲁棒性的方案.在随后的研究中,Papernot 等人[74]指出,面对黑盒攻击防御精馏存在缺陷,并提出了可扩展的防御精馏技术.实验证明:使用防御精馏技术可以产生输出表面更平滑的、对扰动不敏感的模型提高模型的顽健性,且能够将对抗样本攻击的成功率从 95%降到不足 0.5%.但很快,Carlini[75]就指出了防御精馏存在缺陷,并提出了大量破坏防御精馏安全性的攻击.

3.4 隐私保护机制

数据集和模型正则化机制能够防御数据集投毒攻击和对抗样本攻击;对抗训练机制能够提升学习模型对对抗样本的鲁棒性;防御精馏能够提升模型应对扰动的能力,提升模型输出的平滑性.以上 3 种方案分别针对数据投毒攻击和对抗样本攻击进行了防御.而数据窃取攻击和隐私询问攻击是针对模型和数据的隐私攻击,通过使用加密、扰动方案,可以在根本上保护数据和模型的隐私,而模型的隐私改造可以使模型以保护隐私的机制进行学习.

(1) 加密方案

加密是保障数据安全性和隐私性的重要手段,在用户数据进入机器学习服务提供商之前,使用加密手段可以防止因存储和传输的安全漏洞导致的数据窃取攻击.同态加密、乱码电路[76,77]、秘密共享机制[78,79]和安全处理器机制是最常使用的加密方法.

考虑到机器学习服务提供商有窃取和利用用户隐私的嫌疑,同态加密技术[80,81]通过对训练数据和模型数据的加密实现了对数据隐私的保护,允许用户直接对密文进行相应的加法或乘法运算,得到数据仍是加密的结果,与对明文进行同样的操作再将结果加密一样.使用同态加密,用户加密的内容到达机器学习服务提供商后无法被解密,直接进入机器学习模型中,提升了模型的隐私性.同时,通过使用同态加密保证数据安全性不可避免地带来了效率问题,因此,同态加密多被应用于密文加性计算,而乘法仍在明文进行.

乱码电路是指需要保护的双方或多方要获得某项计算的结果时,将计算转换为乱码电路,并将自己的乱码输入发送给另一方,另一方可以根据电路和收到的乱码输入,结合自己的乱码输入获得计算结果并分享给发送方的方法.Bost 等人[81]结合同态加密和乱码电路实现了超平面决策、朴素贝叶斯和决策树这 3 种经典分类机器学习任务.

秘密共享机制是利用 shamir 门限方案的特性,即 w 个参与者共享一个秘钥,任意 t(门限值)个参与者都能计算出秘钥的值,而任何 t–1 个参与者都无法计算出秘钥的值.其中,为了共享秘钥,秘钥服务器需秘密地向每一个参与者发送一部分信息,这些信息称为共享(share).将秘密共享机制应用在机器学习中,可以使用户训练的模型通过无共谋的服务器传输,最终通过门限值以上的子模型构建新的模型更新.Bonawitz 等人[79]结合分布式网络特征提出了模型加密聚合方案,通过用户将使用秘密共享秘钥和用户私钥双重加密的内容互相传递的方式,在获得门限值以上的共享后解密得到最终结果,可实现高效率、高维度的数据加密传输计算.但加密聚合方案使得服务器无法根据数值判断收集到的子模型是否被污染或破坏,使得攻击者通过攻击分布式子模型从而破坏全局模型成为可能.如何在加密聚合的环境下检测和过滤异常子模型,还没有很好的解决方案.如何实时检测聚合学习机制中的子模型质量,是保证全局安全性的关键和重要研究方向.

安全处理器机制是通过硬件设备的安全性保证计算安全性的方案,其中,SGX(Intel software guard extensions)[82]是 Intel 公司的软件安全性增强技术,通过一组 CPU 指令,隔离应用程序代码和数据的特定可信区域,为开发人员提供安全可信空间,使敏感数据或代码免受外部的干扰或检查.

(2) 扰动方案

针对推理阶段的询问攻击带来的各种安全问题,安全防御机制的重点在于保证输入数据和模型数据的隐私性.对于本文第 2 节中提出的机器学习,容易遭受成员推理攻击.这是因为具有某个特征向量的个体在攻击者

具有一定背景知识的条件下,容易被去匿名化.在数据库系统中,Dwork 等人[83]针对这种问题提出了差分隐私机制,为数据库分析算法提供了很好的隐私标准.对于一种随机化算法 M,其分别作用于两个仅相差一个样本的相邻数据集,差分隐私形式化的定义为

$$\Pr[M(d)\in S]\leqslant e^{\epsilon}\Pr[M(d')\in S]+\delta.$$

差分隐私指出:通过合理的数学计算和对数据添加干扰噪声的方式保护所发布数据中潜在的用户隐私信息,可以使攻击者在拥有完美背景知识的情况下,通过询问攻击无法识别单个个体.基于差分隐私的这种特性,将其应用于机器学习的数据和模型保护中,可以防止成员推理攻击,也可以在分布式学习中保证原始数据的隐私性.根据应用位置的不同,可以分为输入扰动、模型扰动和输出扰动.

输入扰动中,Dwork 等人[84]和 Yu 等人[85]将差分隐私引入训练数据中,对模型的输入进行扰动,提高了隐私性.模型扰动中,Hardt[86]、Abadi[87]和 Song 等人[88]将差分隐私引入模型算法中,对模型的梯度进行扰动,提高了隐私性.Geyer 等人[89]将差分隐私引入分布式的随机梯度下降算法的梯度干扰中.输出扰动中,Chaudhuri 等人[90]引入差分隐私对模型的输出进行扰动,减低了攻击者成员推理攻击和模型反演攻击的可能性.Chaudhuri 等人[91]对文献[90]进行了改进,提出了使用差分隐私的经验风险最小化算法.Kung 等人[92]提出了将差分隐私应用于机器学习中可用性高隐私性高最优化算法.

(3) 模型的隐私改造

针对机器学习隐私的攻击多发生在模型训练之前的训练数据攻击和训练之后的黑盒攻击中,攻击者通过直接盗取训练数据发动隐私攻击,或通过询问接口多次询问,反推训练数据和模型隐私,从而重构训练数据和模型.加密和扰动作为经典的安全机制,虽然提升了机器学习的安全性,但也因计算复杂度和处理数据量的限制无法应用于所有机器学习算法中.为了应对多种使用环境和攻击类型,对模型结构的改造和创新能够在提高效率的同时满足安全性要求.

大多数机器学习服务提供商采用集中式学习的方案,集中收集用户数据进行计算并提供给用户询问接口.尽管用户可以在上传过程中进行加密,但服务提供商进行学习和计算之前这些数据都将变成原始数据.这种方案使服务提供商无条件获取用户隐私.随着分布式网络的大范围应用,分布式机器学习应运而生.Hitaj 等人[93]利用生成对抗网络对联合分布式训练深度学习模型发起攻击,提高网络的隐私性.这使得任意参加训练的用户都可能成为敌手,生成与其他参与者训练数据无限逼近的假样本来窃取他人隐私.此外,Papernot 等人[94]在精馏方法的基础上提出了 PATE 模型,如图 9 所示.精馏是使用一个教师模型的输出训练另一个学生模型以进行模型压缩的方法,而 PATE 模型通过将集中式模型分解成多个老师模型最终聚合成学生模型的方式,降低了模型的敏感度,通过给聚合过程增加差分隐私(differential privacy)[83]的方式提升了隐私性,并通过隐蔽前置操作,仅向用户提供学生模型询问接口的方式降低了模型反演攻击的可能,提升了安全性.

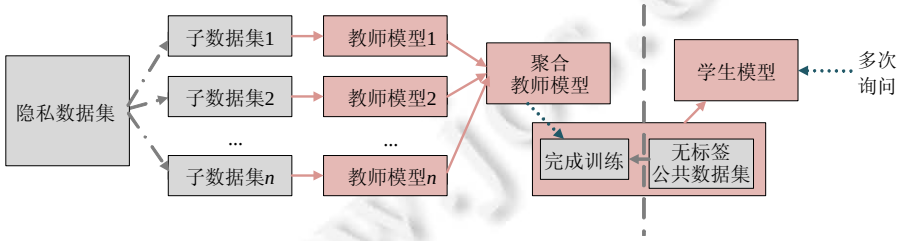


Fig.9 PATE model

图 9 PATE 模型

当然,提升硬件和软件系统的安全性,也是提升机器学习隐私性的重要部分.Ohrimenko 等人[95]提出了可以让多方安全地进行模型训练的安全处理器.区块链技术除了在商业领域的广泛使用,也被用来进行隐私数据的保护[96]和机器学习的参数存储[97]等.

安全攻击和防御机制见表 1.

Table 1 Security attacks and defense mechanisms

表 1 安全攻击和防御机制

攻击位置	攻击方式	攻击分类	分类	文献	防御方法
训练阶段	数据窃取攻击	数据窃取	数据窃取	[37,51,52]	加密[76–82], 扰动[83–92]
	数据投毒攻击	投毒攻击	标签投毒 数据投毒	[21,22] [14,23–27]	输入空间正则化[21,59–63]
推理阶段	对抗样本攻击	白盒攻击	错误分类 源/目标对应错误分类 靶向错误分类	[12,28,32,38] [5,6,8,34,38] [35,36]	模型参数正则化[64], Dropout[65], 对抗训练[10,66–69], 防御精馏[72,74]
		黑盒攻击	黑盒攻击 转移性攻击	[39–42,45,46] [47,48]	
	询问攻击	成员推理攻击 数据提取攻击 模型推理攻击	成员推理 数据推理 模型推理	[53] [54,55] [56–58]	生成对抗网络[93], PATE[94], 软硬件安全[95–97], 扰动[83–92]

4 安全机器学习研究挑战和方向

在机器学习前后使用加密和扰动机制,也为机器学习的安全性提供了保障.但是计算复杂度和可操作数据量,使其无法大范围的使用.同时,针对多种学习模型的对抗样本攻击,不同的应对方案通常只对特定攻击防御有效,在模型面对新的攻击形式时依然脆弱.对抗样本的转移性有利有弊,在可以进行对抗训练的同时,也能对新的模型产生干扰.Dropout 和 PATE 机制是机器学习结构的创新,在一定程度上解决了机器学习的安全和隐私问题.但新的问题应运而生,例如,复杂的算法结构会降低学习效率、同时提升对训练数据的容量和维度要求,而训练数据的容量和维度本身就是机器学习面临的问题.出于对计算成本和安全收益的考量,如何平衡性能与安全,是所有安全机制面临的问题.除了已知的数据污染攻击、对抗样本攻击和询问攻击等,新的攻击形式正在产生,亟需研究更有力的安全机制.

综上,现有的机器学习安全机制还有很大的发展空间,总结未来的研究方向如下.

(1) 数据和模型的异常检测

一方面,训练数据的数量不足会导致模型过拟合,使模型面对新数据时出错率更高;训练数据的投毒攻击会降低数据质量,从而降低模型的正确性.因而,保证模型的性能必须保证训练数据的数量与质量.为了保证训练数据的数量,合理地收集、整理和扰动是扩充数据集的重要途径;为了保证训练数据及的质量,合理的数据清洗和正则化策略可以减少和防御训练数据的投毒攻击.

另一方面,在多个子模型共同参与学习的机器学习模型中,攻击者通过偷渡或破坏少数模型的训练结果发动对整体模型的投毒攻击,影响模型的正确性.为了防止模型参与方出错,Bonawitz 等人[79]结合分布式网络的特征提出了模型的加密聚合方案,保证了模型参与方在传输中的安全和隐私,但同样给聚合方检测和过滤异常子模型带来了困难,无法防御参与方发动的异常攻击.因此,如何在保证传输安全的条件下检测和过滤异常子模型,是保证全局安全性的关键和研究方向.

(2) 对抗样本生成算法和模型输出平滑性

通过在模型训练过程中不断将对抗样本和正确的标签作为训练数据,对抗训练机制能够提高模型应对对抗样本的鲁棒性.L-BFGS[6]、FGSM[35]、DeepFool[33]、Carlini-Wagner[70]、Xu[42]这些对抗样本生成算法不仅被用来发动对抗样本攻击,其生成的对抗样本也被用在对抗训练中.寻找和发现更多的对抗样本生成算法以扩充对抗训练的训练数据集,是机器学习应对对抗样本攻击的重要手段.

对抗样本生成算法不断更新,研究者无法及时将对抗样本应用到对抗训练中.为了应对新兴攻击手段,防御精馏可以提高模型对细微数据改动的鲁棒性,提高模型输出的平滑性,一定程度上防御对抗样本攻击.为了提升

模型应对新的对抗样本的能力,更多平滑输出的方法是重要的研究方向.

(3) 模型隐私性提升

对机器学习的隐私攻击多发生在推理阶段的黑盒询问攻击中,攻击者通过多次询问反推训练数据和模型隐私,从而重构训练数据和模型数据.为保证集中式机器学习的数据隐私,PATE[94]模型在精馏的基础上,通过对多个教师模型进行扰动聚合,保证了单个模型输出的隐私,从而防御攻击者对数据的去匿名化攻击.而同态加密、零知识证明、差分隐私等机制也广泛应用于分布式学习中,在带来隐私性的同时,多种加密和扰动方案带来的计算代价和安全问题仍需解决.因此,寻找更安全高效的加密和扰动算法以及隐私保护模型是未来的研究方向.

References:

[1] Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587):484−489. [doi: 10.1038/nature16961]

[2] Dalvi N, Domingos P, Sanghai S, Verma D, *et al*. Adversarial classification. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2004. 99−108. [doi: 10.1145/1014052.1014066]

[3] Lowd D, Meek C. Adversarial learning. In: Proc. of the 11th ACM Sigkdd Int'l Conf. on Knowledge Discovery in Data Mining. 2005. [doi: 10.1145/1081870.1081950]

[4] Kearns MJ, Li M. Learning in the presence of malicious errors. SIAM Journal on Computing, 1993,22(4):807−837. [doi: 10.1137/0222052]

[5] Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[6] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the IEEE Symp. on Security and Privacy (SP). 2017. 39−57. [doi:10.1109/sp.2017.49]

[7] Papernot N, Mcdaniel P, Sinha A, *et al*. SoK: Security and privacy in machine learning. In: Proc. of the IEEE European Symp. on Security and Privacy. 2018. 399−414. [doi: 10.1109/EuroSP.2018.00035]

[8] Papernot N, Mcdaniel PD, Jha S, *et al*. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy. 2016. 372−387. [doi: 10.1109/EuroSP.2016.36]

[9] Song L, Ma CG, Duan GH. Machine learning security and privacy: A survey. Chinese Journal of Network and Information Security, 2018,4(8):1−11 (in Chinese with English abstract). [doi: 10.11959/j.issn.2096-109x.2018067]

[10] Kurakin A, Goodfellow IJ, Bengio S, *et al*. Adversarial machine learning at scale. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[11] Li P, Zhao WT, Liu Q, *et al*. Security issues and their countermeasuring techniques of machine learning: A survey. Journal of Frontiers of Computer Science and Technology, 2018,12(2):171−184 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.1708038]

[12] Biggio B, Corona I, Maiorca D, *et al*. Evasion attacks against machine learning at test time. In: Proc. of the European Conf. on Machine Learning. 2013. 387−402. [doi: 10.1007/978-3-642-40994-3_25]

[13] Powers DM. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2011,2(1):37−63. [doi: 10.9735/2229-3981]

[14] Biggio B, Nelson B, Laskov P, *et al*. Poisoning attacks against support vector machines. In: Proc. of the Int'l Conf. on Machine Learning. 2012. 1467−1474.

[15] Mahloujifar S, Mahmoody M, Mohammed A, *et al*. Multi-party poisoning through generalized *p*-tampering. arXiv:1809.0347, 2018.

[16] Rubinstein BI, Nelson B, Huang L, *et al*. ANTIDOTE: Understanding and defending against poisoning of anomaly detectors. In: Proc. of the Internet Measurement Conf. 2009. 1−14. [doi: 10.1145/1644893.1644895]

[17] Jacob S, Pang WK, Percy L. Certified defenses for data poisoning attacks. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. 2017. 3520−3532.

[18] Sharif M, Bhagavatula S, Bauer L, *et al*. Accessorize to a Crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proc. of the Computer and Communications Security. 2016. 1528−1540. [doi: 10.1145/2976749.2978392]

[19] Globerson A, Roweis ST. Nightmare at test time: Robust learning by feature deletion. In: Proc. of the Int'l Conf. on Machine Learning. 2006. 353−360. [doi: 10.1145/1143844.1143889]

[20] Valiant LG. A theory of the learnable. Symp. on the Theory of Computing, 1984,27(11):1134−1142. [doi: 10.1145/1968.1972]

[21] Biggio B, Nelson B, Laskov P, *et al*. Support vector machines under adversarial label noise. In: Proc. of the Asian Conf. on Machine Learning. 2011. 97−112.

[22] Mozaffarikermani M, Surkolay S, Raghunathan A, *et al*. Systematic Poisoning attacks on and defenses for machine learning in healthcare. IEEE Journal of Biomedical and Health Informatics, 2014,19(6):1893−1905. [doi: 10.1109/JBHI.2014.2344095]

[23] Mei S, Zhu X. Using machine teaching to identify optimal training-set attacks on machine learners. In: Proc. of the National Conf. on Artificial Intelligence. 2015. 2871−2877.

[24] Kloft M, Laskov P. Online anomaly detection under adversarial impact. In: Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS). 2010. 405−412.

[25] Kloft M, Laskov P. Security analysis of online centroid anomaly detection. Journal of Machine Learning Research, 2012,13(1): 3681−3724. [doi: 10.1016/j.dss.2012.08.019]

[26] Biggio B, Didaci L, Fumera G, *et al*. Poisoning attacks to compromise face templates. In: Proc. of the Int'l Conf. on Biometrics. 2013. 1−7. [doi: 10.1109/ICB.2013.6613006]

[27] Perdisci R, Dagon D, Lee W, *et al*. Misleading worm signature generators using deliberate noise injection. In: Proc. of the IEEE Symp. on Security and Privacy. 2006. 17−31. [doi: 10.1109/SP.2006.26]

[28] Kurakin A, Goodfellow IJ, Bengio S, *et al*. Adversarial examples in the physical world. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[29] Papernot N, Mcdaniel P, Goodfellow I, *et al*. Practical black-box attacks against machine learning. In: Proc. of the Computer and Communications Security. 2017. 506−519. [doi: 10.1145/3052973.3053009]

[30] Athalye A, Engstrom L, Ilyas A, *et al*. Synthesizing robust adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning. 2018. 284−293.

[31] Moosavidezfooli S, Fawzi A, Fawzi O, *et al*. Universal adversarial perturbations. In: Proc. of the Computer Vision and Pattern Recognition. 2017. 86−94. [doi: 10.1109/CVPR.2017.17]

[32] Grosse K, Papernot N, Manoharan P, *et al*. Adversarial perturbations against deep neural networks for malware classification. arXiv: 1606.04435, 2016.

[33] Moosavidezfooli S, Fawzi A, Frossard P, *et al*. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the Computer Vision and Pattern Recognition. 2016. 2574−2582. [doi: 10.1109/CVPR.2016.282]

[34] Huang R, Xu B, Schuurmans D, *et al*. Learning with a strong adversary. arXiv:1511.03034, 2015.

[35] Nguyen A, Yosinski J, Clune J, *et al*. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the Computer Vision and Pattern Recognition. 2015. 427−436. [doi: 10.1109/CVPR.2015.7298640]

[36] Carlini N, Mishra P, Vaidya T, *et al*. Hidden voice commands. In: Proc. of the Usenix Security Symp. 2016. 513−530.

[37] Smith DF, Wiliem A, Lovell BC, *et al*. Face recognition on consumer devices: Reflections on replay attacks. IEEE Trans. on Information Forensics and Security, 2015,10(4):736−745. [doi: 10.1109/TIFS.2015.2398819]

[38] Sharif M, Bhagavatula S, Bauer L, *et al*. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proc. of the Computer and Communications Security. 2016. 1528−1540. [doi: 10.1145/2976749.2978392]

[39] Wittel GL, Wu SF. On attacking statistical spam filters. In: Proc. of the Conf. on Email and Anti-Spam. 2004. 1−7.

[40] Lowd D, Meek C. Good word attacks on statistical spam filters. In: Proc. of the Conf. on Email and Anti-Spam. 2005. 1−8.

[41] Alfeld S, Zhu X, Barford P, *et al*. Data poisoning attacks against autoregressive models. In: Proc. of the National Conf. on Artificial Intelligence. 2016. 1452−1458.

[42] Xu W, Qi Y, Evans D, *et al*. Automatically evading classifiers: A case study on PDF malware classifiers. In: Proc. of the Network and Distributed System Security Symp. 2016. [doi: 10.14722/ndss.2016.23115]

[43] Smutz C, Stavrou A. Malicious PDF detection using metadata and structural features. In: Proc. of the Annual Computer Security Applications Conf. 2012. 239−248. [doi: 10.1145/2420950.2420987]

[44] Srndic N, Laskov P. Detection of malicious PDF files based on hierarchical document structure. In: Proc. of the Network and Distributed System Security Symp. 2013. [doi: 10.1145/2420950.2420987]

[45] Rndic N, Laskov P. Practical evasion of a learning-based classifier: A case study. In: Proc. of the IEEE Symp. on Security and Privacy. 2014. 197−211. [doi: 10.1109/SP.2014.20]

[46] Smutz C, Stavrou A. Malicious PDF detection using metadata and structural features. In: Proc. of the Annual Computer Security Applications Conf. 2012. 239−248. [doi: 10.1145/2420950.2420987]

[47] Papernot N, Mcdaniel PD, Goodfellow IJ, *et al*. Practical black-box attacks against deep learning systems using adversarial examples. arXiv:1602.02697v2, 2016.

[48] Papernot N, Mcdaniel PD, Goodfellow IJ, *et al*. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv:1605.07277, 2016.

[49] Papernot N, Mcdaniel PD, Sinha A, *et al*. Towards the science of security and privacy in machine learning. arXiv:1611.03814, 2016.

[50] Alrubaie M, Chang JM. Privacy-preserving machine learning: Threats and solutions. IEEE Symp. on Security and Privacy, 2019, 17(2):49−58. [doi: 10.1109/MSEC.2018.2888775]

[51] Feng J, Jain AK. Fingerprint reconstruction: From minutiae to phase. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(2):209−223. [doi: 10.1109/TPAMI.2010.77]

[52] Alrubaie M, Chang JM. Reconstruction attacks against mobile-based continuous authentication systems in the cloud. IEEE Trans. on Information Forensics and Security, 2016,11(12):2648−2663. [doi: 10.1109/TIFS.2016.2594132]

[53] Shokri R, Stronati M, Song C, *et al*. Membership inference attacks against machine learning models. In: Proc. of the IEEE Symp. on Security and Privacy. 2017. 3−18. [doi: 10.1109/SP.2017.41]

[54] Fredrikson M, Lantz E, Jha S, *et al*. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proc. of the Usenix Security Symp. 2014. 17−32.

[55] Fredrikson M, Jha S, Ristenpart T, *et al*. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the Computer and Communications Security. 2015. 1322−1333. [doi: 10.1145/2810103.2813677]

[56] Vorobeychik Y, Li B. Optimal randomized classification in adversarial settings. In: Proc. of the Int'l Conf. on Autonomous Agents and Multi-agent Systems. 2014. 485−492.

[57] Ateniese G, Mancini LV, Spognardi A, *et al*. 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. [doi: 10.1504/ijsn.2015.071829]

[58] Tramer F, Zhang F, Juels A, *et al*. Stealing machine learning models via prediction APIs. In: Proc. of the Usenix Security Symp. 2016. 601−618.

[59] Qiao M, Valiant G. Learning discrete distributions from untrusted batches. In: Proc. of the Conf. on Innovations in Theoretical Computer Science. 2018. [doi: 10.4230/LIPIcs.ITCS.2018.47]

[60] Steinhardt J, Koh PW, Liang P, *et al*. Certified defenses for data poisoning attacks. In: Proc. of the Neural Information Processing Systems. 2017. 3517−3529.

[61] Rubinstein BI, Nelson B, Huang L, *et al*. ANTIDOTE: Understanding and defending against poisoning of anomaly detectors. In: Proc. of the Internet Measurement Conf. 2009. 1−14. [doi: 10.1145/1644893.1644895]

[62] Fung C, Yoon CJ, Beschastnikh I, *et al*. Mitigating sybils in federated learning poisoning. arXiv: 1808.04866, 2018.

[63] Shen SQ, Tople S, Saxena P. Auror: Defending against poisoning attacks in collaborative deep learning systems. In: Proc. of the Conf. on Computer Security Applications. 2016. 508−519. [doi: 10.1145/2991079.2991125]

[64] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[65] Srivastava N, Hinton GE, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014,15(1):1929−1958.

[66] Pinto L, Davidson J, Gupta A, *et al*. Supervision via competition: Robot adversaries for learning tasks. In: Proc. of the Int'l Conf. on Robotics and Automation. 2017. 1601−1608. [doi: 10.1109/ICRA.2017.7989190]

[67] Goodfellow I, Shlens J, Szegedy C, *et al*. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations. 2015.

[68] Huang R, Xu B, Schuurmans D, *et al*. Learning with a strong adversary. arXiv:1511.03034, 2015.

[69] Tramer F, Kurakin A, Papernot N, *et al*. Ensemble adversarial training: Attacks and defenses. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[70] Carlini N, Wagner DA. Towards evaluating the robustness of neural networks. In: Proc. of the IEEE Symp. on Security and Privacy. 2017. 39−57. [doi: 10.1109/SP.2017.49]

[71] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv: 1503.02531, 2015.

[72] Papernot N, Mcdaniel P, Wu X, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the IEEE Symp. on Security and Privacy. 2016. 582−597. [doi: 10.1109/SP.2016.41]

[73] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning. 2018.

[74] Papernot N, Mcdaniel P. Extending defensive distillation. arXiv:1705.05264, 2017.

[75] Carlini N, Wagner DA. Defensive distillation is not robust to adversarial examples. arXiv:1607.04311, 2016.

[76] Bost R, Popa RA, Tu S, *et al*. Machine learning classification over encrypted data. In: Proc. of the Network and Distributed System Security Symp. 2015. [doi: 10.14722/ndss.2015.23241]

[77] Nikolaenko V, Weinsberg U, Ioannidis S, *et al*. Privacy-preserving ridge regression on hundreds of millions of records. In: Proc. of the IEEE Symp. on Security and Privacy. 2013. 334−348. [doi: 10.1109/SP.2013.30]

[78] Bogdanov D, Kamm L, Laur S, *et al*. Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2018,15(5):1427−1432. [doi: 10.1109/TCBB.2018.2858818]

[79] Bonawitz KA, Ivanov V, Kreuter B, *et al*. Practical secure aggregation for privacy-preserving machine learning. In: Proc. of the Computer and Communications Security. 2017. 1175−1191. [doi: 10.1145/3133956.3133982]

[80] Erkin Z, Veugen T, Toft T, *et al*. Generating private recommendations efficiently using homomorphic encryption and data packing. IEEE Trans. on Information Forensics and Security, 2012,7(3):1053−1066. [doi: 10.1109/TIFS.2012.2190726]

[81] Bost R, Popa RA, Tu S, *et al*. Machine learning classification over encrypted data. In: Proc. of the Network and Distributed System Security Symp. 2015. [doi: 10.14722/ndss.2015.23241]

[82] Ohrimenko O, Schuster F, Fournet C, *et al*. Oblivious multi-party machine learning on trusted processors. In: Proc. of the Usenix Security Symp. 2016. 619−636.

[83] Dwork C. Differential privacy. In: Proc. of the Int'l Colloquium on Automata Languages aAnd Programming. 2006. 1−12. [doi: 10.1007/11787006_1]

[84] Dwork C, Talwar K, Thakurta A, *et al*. Analyze gauss: Optimal bounds for privacy-preserving principal component analysis. In: Proc. of the Symp. on the Theory of Computing. 2014. 11−20.

[85] Yu H, Vaidya J, Jiang X, *et al*. Privacy-preserving SVM classification on vertically partitioned data. In: Proc. of the Knowledge Discovery and Data Mining. 2006. 647−656. [doi: 10.1007/11731139_74]

[86] Hardt M, Price E. The noisy power method: A meta algorithm with applications. In: Proc. of the Neural Information Processing Systems. 2014. 2861−2869.

[87] Abadi M, Chu A, Goodfellow IJ, *et al*. Deep learning with differential privacy. In: Proc. of the Computer and Communications Security. 2016. 308−318. [doi: 10.1145/2976749.2978318]

[88] Song S, Chaudhuri K, Sarwate AD, *et al*. Stochastic gradient descent with differentially private updates. In: Proc. of the IEEE Global Conf. on Signal and Information Processing. 2013. 245−248. [doi: 10.1109/GlobalSIP.2013.6736861]

[89] Geyer RC, Klein T, Nabi M, *et al*. Differentially private federated learning: A client level perspective. arXiv:1712.07557v2, 2017.

[90] Chaudhuri K, Sarwate AD, Sinha K, *et al*. A near-optimal algorithm for differentially-private principal components. Journal of Machine Learning Research, 2013,14(1):2905−2943. [doi: 10.1016/j.robot.2013.06.001]

[91] Chaudhuri K, Monteleoni C, Sarwate AD, *et al*. Differentially private empirical risk minimization. Journal of Machine Learning Research, 2011,12(3):1069−1109. [doi: 10.1109/MIS.2011.2]

[92] Kung S. Compressive privacy: From information/estimation theory to machine learning. IEEE Signal Processing Magazine, 2017,34(1):94−112. [doi: 10.1109/MSP.2016.2616720]

[93] Hitaj B, Ateniese G, Perezcruz F, *et al*. Deep models under the GAN: Information leakage from collaborative deep learning. In: Proc. of the Computer and Communications Security. 2017. 603−618. [doi: 10.1145/3133956.3134012]

[94] Papernot N, Abadi M, Erlingsson U, *et al*. Semi-supervised knowledge transfer for deep learning from private training data. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[95] Ohrimenko O, Schuster F, Fournet C, *et al*. Oblivious multi-party machine learning on trusted processors. In: Proc. of the Usenix Security Symp. 2016. 619−636.

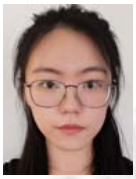
[96] Zyskind G, Nathan O, Pentland A, *et al*. Decentralizing privacy: Using blockchain to protect personal data. In: Proc. of the IEEE Symp. on Security and Privacy. 2015. 180−184. [doi: 10.1109/SPW.2015.27]

[97] Outchakoucht A, Essamaali H, Leroy JP, *et al*. Dynamic access control policy based on blockchain and machine learning for the Internet of things. Int'l Journal of Advanced Computer Science and Applications, 2017,8(7):417−424. [doi: 10.14569/IJACSA.2017.080757]

附中文参考文献:

[9] 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):1−11. [doi: 10.11959/j.issn.2096-109x.2018067]

[11] 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.计算机科学与探索,2018,12(2):171−184. [doi: 10.3778/j.issn.1673-9418.1708038]



李欣姣(1992−),女,博士生,主要研究领域为大数据安全和隐私.

张伟哲(1976−),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为网络空间安全,网络安全,系统安全,内容安全,云计算,高性能计算.

吴国伟(1973−),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为自主智能系统,智能边缘计算.

张宾(1976−),男,博士,高级工程师,CCF 专业会员,主要研究领域为网络测量,拓扑发现,异常检测.

姚琳(1976−),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为大数据安全和隐私.