ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2018/12(02)-0171-14
doi: 10.3778/j.issn.1673-9418.1708038

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

机器学习安全性问题及其防御技术研究综述

李 盼,赵文涛,刘 强+,崔建京,殷建平
国防科技大学 计算机学院,长沙 410073

Security Issues and Their Countermeasuring Techniques of Machine Learning: A Survey

LI Pan, ZHAO Wentao, LIU Qiang+, CUI Jianjing, YIN Jianping
College of Computer, National University of Defense Technology, Changsha 410073, China
+ Corresponding author: E-mail: qiangliu06@nudt.edu.cn

LI Pan, ZHAO Wentao, LIU Qiang, 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.

Abstract: Machine learning has already become one of the most widely used techniques in the field of computer science, and it has been widely applied in image processing, natural language processing, network security and other fields. However, there has been many security threats that need to be overcome on current machine learning algorithms and training data set, which will affect the security of several practical applications, such as facial detection, malware detection and automatic driving, etc. According to the known security threats, which aim to a variety of machine learning algorithms, such as the support vector machine (SVM) classifier, clustering and deep neural networks, this paper introduces the issues that happen in the training, testing/inference phase of machine learning, which include privacy leaking and attacks of poisoning, evasion, impersonate and inversion based on the adversarial samples. Then, this paper sums up the machine learning adversary model as well as its safety assessment mechanism and concludes a certain number of countermeasures and privacy protection techniques on training and testing processes. Finally, this paper looks forward some correlative problems worthy of further discussion.
Key words: machine learning; adversarial sample; security threats; countermeasuring techniques

摘 要: 机器学习已经成为当前计算机领域研究和应用最广泛的技术之一,在图像处理、自然语言处理、网络安全等领域被广泛应用。然而,一些机器学习算法和训练数据本身还面临着诸多安全威胁,进而影响到基于

Received 2017-07, Accepted 2017-09.
CNKI网络优先出版: 2017-09-11, http://kns.cnki.net/kcms/detail/11.5602.TP.20170911.1551.002.html

机器学习的面部检测、恶意程序检测、自动驾驶汽车等实际应用系统的安全性。由目前已知的针对支持向量机(support vector machine,SVM)分类器、聚类、深度神经网络(deep neural networks,DNN)等多种机器学习算法的安全威胁为出发点,介绍了在机器学习的训练阶段和测试/推理阶段中出现的基于对抗样本的投毒、逃逸、模仿、逆向等攻击和隐私泄露等问题,归纳了针对机器学习的敌手模型及其安全评估机制,总结了训练过程和测试过程中的若干防御技术和隐私保护技术,最后展望了下一步机器学习安全研究的发展趋势。

关键词:机器学习;对抗样本;安全威胁;防御技术

文献标志码:A **中图分类号**:TP181

1 引言

近年来,机器学习受到广泛关注,并在很多领域中取得了很好的应用成果。比如:网络安全检测(垃圾邮件检测、恶意程序检测等)、图像识别(人脸识别、图片分类等)、自动驾驶以及其他与人们日常生活密切相关的领域,都有机器学习的应用实例。由此可见,机器学习将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,在机器学习带给人们巨大便利的同时,其本身也暴露了一些安全性问题。早期在垃圾邮件检测系统[1]和入侵检测系统[2]等应用机器学习算法的安全领域中发现了针对系统模型特点来逃避检测的问题,给机器学习在安全检测领域带来了很大的挑战。迄今为止,越来越多威胁机器学习安全的问题被发现,有针对面部识别系统(face recognition system,FRS)[3-4]缺陷来模仿受害者身份的非法认证危害,也有涉及医疗数据[5]、人物图片数据[6]的隐私窃取危害,更有针对自动驾驶汽车[7]、语音控制系统[8]的恶意控制危害。随着机器学习应用领域的不断扩大,有关机器学习的安全性问题将受到更为广泛的关注。

Dalvi等人在2004年提出敌手分类(adversarial classification)的概念[9],指出了早期机器学习在垃圾邮件检测系统中的一些恶意邮件逃避分类检测问题。2005年Lowd等人进一步提出对抗学习(adversarial learning)的概念[10]。Barreno等人在2006年首次较为明确地提出机器学习安全问题的有关概念和知识[11],包括机器学习系统的攻击分类和敌手建模,并在2010年进一步完善了有关机器学习安全的一些概念[12]。目前,有关机器学习安全的更加系统、规范化的概念和机制还在不断完善。

本文组织结构如下:第2章简单介绍了机器学习技术及其分类,总结了当前针对机器学习敌手模型和安全性问题的分类方法,并从机器学习的两个重要阶段,即训练和测试/推理阶段,具体介绍了当前机器学习面临的安全威胁;第3章针对第2章的安全性问题,归纳了已有的对机器学习算法进行安全评估的框架,以及应对安全问题的防御技术和隐私保护技术;第4章对机器学习安全威胁和防御技术的下一步研究进行了展望。

2 机器学习安全性问题

2.1 机器学习技术及其分类

机器学习是结合了计算机、概率学、统计学、心理学以及类脑科学等多个学科的交叉研究领域,是人工智能的核心问题之一,主要研究如何更好地让计算机模拟和实现人类的学习行为,从而实现知识的自动获取和产生。

根据反馈的不同,机器学习技术可以分为监督学习、无监督学习和强化学习3类[13]。其中,监督学习的主要特点是要在训练模型时提供给学习系统训练样本以及样本对应的类别标签,因此其又称为有导师学习。典型的监督学习方法有决策树、支持向量机(support vector machine,SVM)、监督式神经网络等分类算法和线性回归等回归算法。无监督学习方法主要特点是训练时只提供给学习系统训练样本,而没有样本对应的类别标签信息。典型的无监督学习方法有聚类学习和自组织神经网络学习。强化学习方法的特点是通过试错(try-and-error)来发现最优行为策略,而不是带有标签的样本学习。

2.2 机器学习安全性问题的分类和敌手模型

在介绍机器学习的安全性问题之前,先引入机器学习安全性问题的分类体系以及其威胁发起者的一些知识——敌手模型。

2.2.1 机器学习安全性问题分类体系

针对机器学习安全性问题的分类,Barreno等人提出从3个不同角度对攻击行为进行分类,之后经过不断的完善[7-8,14],形成图1所示的分类体系。第一个角度是依据攻击对分类器的影响分为诱发型攻击(causative attack)(影响训练集)和探索性攻击(exploratory attack)(不影响训练集)。第二个角度依据攻击造成的安全损害(security violation)将其分为完整性攻击(integrity attack)、可用性攻击(availability attack)和隐私窃取攻击(privacy violation attack)。第三个角度从攻击的专一性(specificity of an attack)将其分为针对性攻击(discriminate attack)和非针对性攻击(indiscriminate attack)。

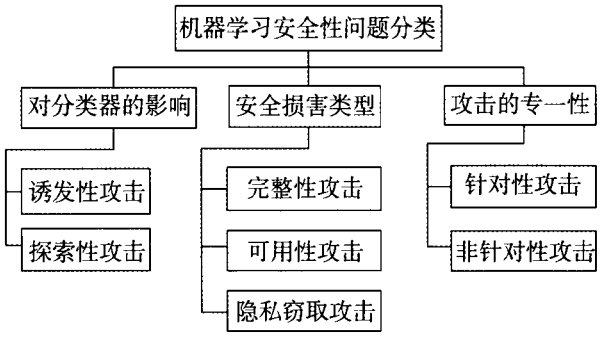

Fig.1 Taxonomy of security threats towards machine learning

图1 针对机器学习的安全威胁分类体系

2.2.2 机器学习敌手模型

Lowd等人最早在2005年提出研究机器学习安全威胁时需要考虑敌手的知识[10];Barreno等人在2010年提出考虑攻击者的目的、能力的敌手模型[14];Biggio等人在此基础上于2014年进一步完善了机器学习的敌手模型[15],提出从敌手目标、敌手知识、敌手能力、敌手策略四方面来建立机器学习敌手模型。具体内容为:

敌手目标(adversary's goal):根据文献[10,12,15],敌手目标可以从两个角度描述,即攻击者期望造成的安全破坏程度(完整性、可用性或隐私性)和攻击的专一性(针对性、非针对性)。例如:攻击者的目标可以是产生一个非针对性的破坏完整性的攻击,来最大化分类器的错误率;也可以产生针对性的窃取隐私的攻击,来从分类器中获得具体的客户隐私信息。

敌手知识(adversary's knowledge):敌手的知识可以从分类器的具体组成来考虑,从敌手是否知道分类器的训练数据、特征集合、学习算法和决策函数的种类及其参数、分类器中可用的反馈信息(敌手通过输入数据得到系统返回的标签信息)等方面将敌手知识划分为有限的知识和完全的知识。

敌手能力(adversary's capability):敌手的知识主要是指攻击者对训练数据和测试数据的控制能力。可以从以下几方面定义:第一是攻击对分类器造成的影响是探索性的还是诱发性的;第二是敌手控制训练数据或者测试数据的程度;第三是敌手操纵的特征的内容及具体程度。

攻击策略(attack strategy):敌手的攻击策略是指攻击者为了最优化其攻击目的会对训练数据和测试数据进行的修改措施。具体包括攻击哪些样本类型,如何修改类别信息,如何操纵特征等。

2.3 机器学习面临的安全性问题

已知的机器学习安全性问题主要集中在监督学习中,其中分类算法的安全性问题居多,有少量相关研究在无监督学习的聚类算法和强化学习中。

对于传统的监督学习算法而言,朴素贝叶斯和SVM算法是两种经典的学习方法,应用十分广泛。而机器学习的安全性问题最早也是在这两种经典分类算法中出现的,暴露的主要安全问题是在以传统监督学习算法为基础的网络安全检测系统上,多数攻击采取训练期间注入恶意数据或者精心制作恶意数据来逃避分类器检测这两种手段。如:最早提出向基于朴素贝叶斯算法的垃圾邮件检测系统中注入恶意数据[1],以及逃避应用线性核SVM的PDF文件中恶意程序检测系统[16]的检测等攻击手段。

聚类算法是一种典型无监督学习算法,它可以发现数据分布的隐含模式,目前已经在很多领域中

使用，尤其是在恶意DNS(domain name system)检测、恶意程序检测、收集网络攻击来源信息等安全领域广泛应用。同样，这些用于安全领域的聚类算法本身也存在安全问题。针对聚类算法的攻击手段主要是训练期间注入恶意数据，进而影响聚类结果。文献[17-20]介绍了针对聚类算法的恶意数据注入威胁；除此之外，针对聚类算法的还有迷惑攻击(obfuscation attacks)[17]，即攻击者的目的是在不改变其他样本聚类结果的前提下，通过混淆对抗样本(adversarial sample)与其他类别的内容来隐藏对抗样本集合。

近几年来，深度学习是机器学习领域中快速发展的研究方向之一，引起了学术界和工业界的广泛关注。而深度神经网络(deep neural networks，DNN)作为深度学习的重要组成，在很多模式识别的任务中取得了优异的性能，尤其是在视觉分类和语音识别上表现尤为突出。虽然DNN在分类性能上高于其他的分类方法，但近期研究表明其具有反直觉(counterintuitive)的特性[21]。具体来讲，在图片和语音识别任务中，DNN只提取了其中很少的特征，导致其无法识别甚至误分类具有部分差异的图片。攻击者利用这个DNN弱点能够逃避系统检测，甚至模仿受害者来获取其权限。2013年末，Szegedy的研究团队[21]首次提出了用他们产生的轻微扰动的图片来欺骗训练好的DNN。随后，其他研究团队[3,22-25]也提出了很多对目前取得优异分类性能的DNN进行模仿攻击的实例，甚至对物理世界的FRS实现了攻击[3,26]。除了FRS面临的安全威胁之外，与DNN相关的诸如语音识别系统[8]、自动驾驶系统[3,26]等领域也面临着安全威胁。

表1总结了针对机器学习算法的安全威胁研究现状。从表1可知，早期的研究主要针对SVM、朴素贝叶斯、聚类和特征选择等学习方法，而目前大量的安全威胁主要针对DNN。

Table 1 Summary of related works regarding security threats against machine learning

表1 机器学习安全威胁研究现状小结

算法	投毒攻击	欺骗攻击		逆向攻击
		逃避攻击	模仿攻击	
朴素贝叶斯		[1]		
SVM	[16,27]	[28]		
特征选择算法(PCA、LASSO)	[29-30]			
DNN	[31]	[3,7,22-26]	[3,7-8,22-26]	[5-6,32]
聚类	[17-19,33]	[17]		

2.3.1 针对训练过程的安全威胁

对数据进行训练是机器学习的一个重要过程，训练过程对实际用于分类和预测模型的好坏有着直接关系。由此可见，训练数据对于机器学习模型的重要程度不言而喻。正因如此，很多攻击者把针对机器学习模型的攻击重点放在了训练数据上，最常见的针对训练过程的攻击就是投毒攻击。

投毒攻击(poisoning attack)主要是对机器学习在训练模型时需要的训练数据进行投毒，是一种破坏模型可用性和完整性的诱发型攻击。攻击者通过注入一些精心伪造的恶意数据样本(带有错误的标签和攻击的性质)，破坏原有的训练数据的概率分布，从而使训练出的模型的分类或者聚类精度降低，达到破坏训练模型的目的。由于实际中应用机器学习算法的系统的原始训练数据大多是保密的，一般不会被攻击者轻易修改，但很多系统为了增强适应能力需要定期重新训练实现模型更新，从而给了攻击者可趁之机。比如，自适应生物面部识别系统[4,29,34-35]、恶意软件分类系统[36]、垃圾邮件检测系统[1]等，都需要定期重新训练。以针对主成分分析(principal component analysis，PCA)算法为基础的自适应生物面部识别系统的攻击为例[29]，攻击者利用系统需要定期更新的特点，如图2所示，在重新训练期间针对特定的受害者对训练数据注入伪造的对抗样本，使原来模型产生用于识别受害者特征的中心值(centroid)Xc在对抗样本的影响下逐步向攻击者特征的中心值Xa移动，进而实现用攻击者图像来代替受害者图像进行验证的攻击目的。对于聚类学习，虽然没有办法对标签进行修改，但文献[17-19]介绍了针对单连接分层聚类和完全连接分层聚类的投毒攻击，采用一种启发策略，通过引入桥(bridge)的概念，来度量对抗样本落在不同类别之间对聚类精度造成的影响，从而找到最优的对抗样本，造成聚类精度的降低。除此之外，目前在实际中广泛应用的SVM算法[27]、神经

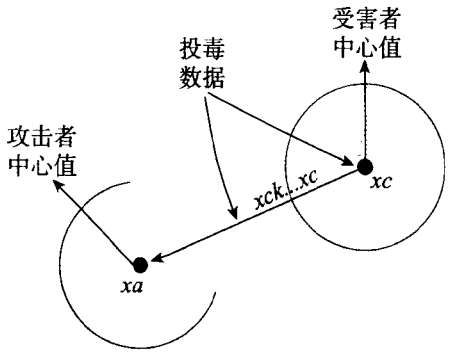

Fig.2 Illustration of a poisoning attack

图2 一种投毒攻击示意图

网络算法[31]、LDA[37](latent Dirichlet allocation)以及一些特征选择算法[30],如PCA、LASSO(least absolute shrinkage and selection operator)等,也面临着投毒攻击的威胁。

而对于实现投毒攻击的技术来说,生成合适的对抗样本是投毒攻击能否成功实现的关键。文献[38]提出了一种针对普遍机器学习算法的攻击算法,该攻击是在已知机器学习训练数据分布和算法的情况下实施攻击,通过产生的对抗样本对模型在验证集上使分类精度下降的程度来筛选对抗样本,并且在决策树、最近邻分类器、多层感知机、SVM算法中均进行了实验,但没有给出该攻击有效性的证明,也无法确保攻击的有效性。而一种比较常用的产生对抗样本的方法——梯度下降策略(gradient ascent strategy)[27,30],通过对度量样本对抗性能的目标函数梯度的计算,进而产生满足要求的最优对抗样本,在对SVM、LASSO算法以及PDF中的恶意程序检测系统的投毒攻击中均有使用。最新的研究[31]提出了一种更有效的产生对抗样本的方法,它采用生成对抗网络(generative adversarial network,GAN)中生成模型(generative model)和判别模型(discriminative model)的思想,即先用生成模型产生候选的对抗样本,再用判别模型筛选最优对抗样本。该方法与一般的直接梯度法(direct gradient)在MNIST和CIFAR-10两个数据集上的实验结果对比可知,GAN产生对抗样本的速度更快,识别准确率(accuracy)更低,损失值(loss)更大。同时,GAN也可以产生针对恶意软件分类器的对抗样本[39]。

除了训练过程,在机器学习的测试/推理过程中也存在着很多安全性问题,给系统造成很大的安全性威胁。如表2所示,不同的攻击类型会给分类器造成不同程度的安全性威胁和损害。

Table 2 Comparison among security threats occurred in training and testing/inference

表2 训练、测试/推理过程中的不同安全威胁比较

过程	攻击类型	对数据集影响	安全损害
训练过程	投毒	诱发性	完整性/可用性
测试/推理过程	逃避攻击	探索性	完整性/可用性
	模仿攻击	探索性	完整性/可用性
	逆向攻击	探索性	隐私性

2.3.2 针对测试/推理过程的安全威胁

测试/推理过程主要指通过训练出的模型来对新数据进行分类或者聚类的过程,是机器学习模型发挥作用的阶段。攻击者利用训练模型的缺陷精心制作对抗样本来逃避检测,或者模仿受害者以获得非法访问权限,甚至通过基于机器学习的相关应用程序接口来获取内部隐私的训练数据,达到获取受害者隐私信息的目的。常见的将针对测试/推理过程的攻击分为欺骗攻击(spoofing attack)(包括逃避攻击、模仿攻击)和逆向攻击。

逃避攻击(evasion attack)是一种比较常见的针对测试/推理过程的欺骗攻击。逃避攻击最早是针对机器学习在安全领域中的一些应用提出的,是早期针对机器学习的攻击形式,比如垃圾邮件检测系统[1]、PDF文件中的恶意程序检测系统[16]等。其主要的攻击策略是攻击者通过产生一些可以成功地逃避安全系统检测的对抗样本,实现对系统的恶意攻击,给系统的安全性带来严重威胁。Biggio的研究团队针对这方面的攻击和防御措施做了一些研究[28,40]。比如:对于文本分类问题[41],文献[28]提出的梯度法来产生最优化的逃避对抗样本,这种方法在垃圾邮件检测系统和PDF文件中的恶意程序检测上成功实现了攻击。最新的研究[3]表明,逃避攻击可以在物理世界中逃避FRS的检测,从而引出了另一大潜在的威胁,即针对一些公众场合的监控系统的安全威胁。

模仿攻击(impersonate attack)是一种和逃避攻击很类似的欺骗攻击,侧重于对受害者样本的模仿,目前主要出现在基于机器学习的图像识别系统和语音识别系统中。攻击者通过产生特定的对抗样本,使机器学习错误地将人类看起来差距很大的样本错分类为攻击者想要模仿的样本,从而达到获取受害者权限(基于面部识别、语音控制的实际系统中)的目的。目前,该类攻击主要发生在DNN算法中,这是因为DNN往往通过提取样本的极少特征来实现目标的识别,所以攻击者很容易通过修改关键特征实现模仿攻击。在图像模仿攻击中比较典型的模仿攻击实例有很多,Nguyen等人[22]提出的利用改进的遗传算法——MAP-Elites(multi-dimensional archive of phenotypic elites)来产生多个类别图片进化后的最优对抗样本,并用这些对抗样本对Goollge的AlexNet和基于Caffe架构的Le-Net-5网络进行模仿攻击,从而欺骗DNN实现误分类。针对物理世界的模仿攻击,Kurakin等人[26]先通过最小相似类(least likely class)的方法产生电子版的对抗样本,再将其打印出来并通过手机相机拍摄来欺骗实际的图片分类系统——GeekPwn 2016,这个过程中由于打印和相机拍摄的过程损失了很多敌手样本原来在电子世界中一些微小的像素特征,因而在物理世界对抗样本的攻击成功率大大低于电子世界的攻击成功率,但证实了真实物理世界模仿攻击实现的可能性。Sharif更是提出一种在实际生活中通过对攻击者的面部戴一副特定的眼镜来躲避最先进的FRS的检测[3],甚至是通过这种方式模仿其他受害者。这种攻击方式经实验验证是物理上可实现的,而且不易察觉,对FRS造成很大的安全威胁。此外,最新的研究[25]表明,采用将多个网络集成的算法来产生可迁移的对抗样本(针对一种DNN产生的对抗样本可以威胁其他DNN),这种方法既可以产生不随目标标签迁移的非针对性对抗样本,也能产生随着目标标签迁移的针对性对抗样本,并且在大规模数据集ILSVRC 2012和最先进的商业图片分类系统Clarifai.com中进行了实验,从而实现对大规模数据集的有效攻击。此外,还存在对声音信息的模仿攻击,文献[8]提到语音控制系统也受到了模仿攻击的困扰,并通过实验证实可以用人类感觉毫无意义的语音来实现对人类语音控制命令的模拟。

逆向攻击(inversion attack)是利用目前机器学习系统提供的一些API(application program interface)来获取系统模型的一些初步信息,进而通过这些初步信息对模型进行逆向分析,从而获取模型内部的一些隐私数据(病人医疗数据、客户生活调查数据、用户面部识别数据等)[5-6,32]。而根据敌手模型中敌手知识的多少,可将其分为白盒攻击(white-box attack)和黑盒攻击(black-box attack)[6]。白盒攻击是指攻击者可以访问下载系统模型或者掌握其他更多的信息,而黑盒攻击是指攻击者只有模型提供的API进行一些访问之后得到的反馈或者有一定的标签知识。这种攻击需要比较大的计算开销,严重威胁了一些关键数据的隐私,有些数据泄漏甚至会给别人带来生命威胁(个性化定制病人药物系统的逆向攻击可能导致错误配药致使病人死亡)[5]。Fredrikson等人用一种通用方法实现了对基于线性回归算法的定制化个人药物系统的逆向攻击[5];之后,Fredrikson又在此基础上实现了对决策树算法的白盒攻击和黑盒攻击,并进一步针对FRS使用梯度下降方法实现了对特定面部图像的重建攻击(reconstruction attack)[6]。Tramer等人[32]利用一些黑盒模型机器学习云服务平台(Google、Amazon等)接口输出的置信值(confidence values),实现了对其的等式求解攻击(equation-dolving attacks),并在模型只有输出标签信息的情况下实现了对多类逻辑回归(multiclass LR)、DNN、RBF核的SVM算法的模型提取攻击(model extraction attacks)。

3 机器学习安全防御技术

针对机器学习的防御技术是根据其训练和推理过程中的具体安全威胁考虑对策的。从图3的系统框图可以看出,安全防御技术主要有针对机器学习算法的安全评估、针对训练过程的防御技术、针对测试/推理过程的防御技术以及对数据安全的隐私保护技术这几方面。

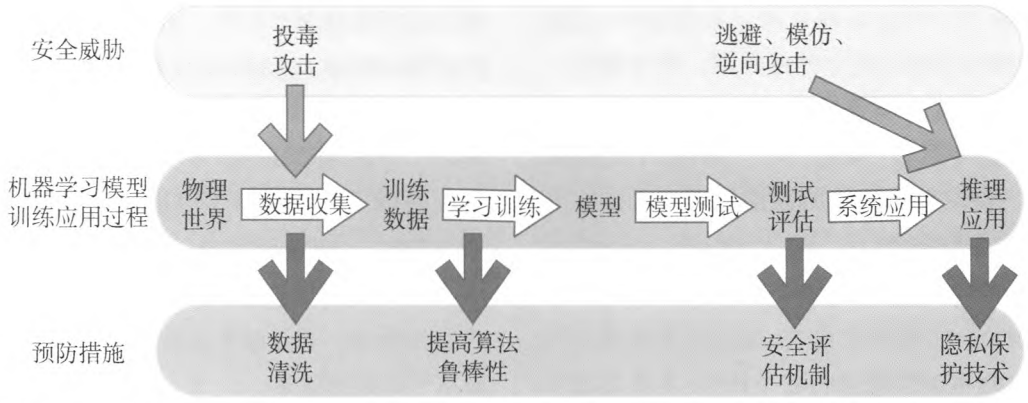

Fig.3　Security threats and their countermeasures in machine learning model training and application

图3　机器学习训练应用过程中安全威胁及防御措施

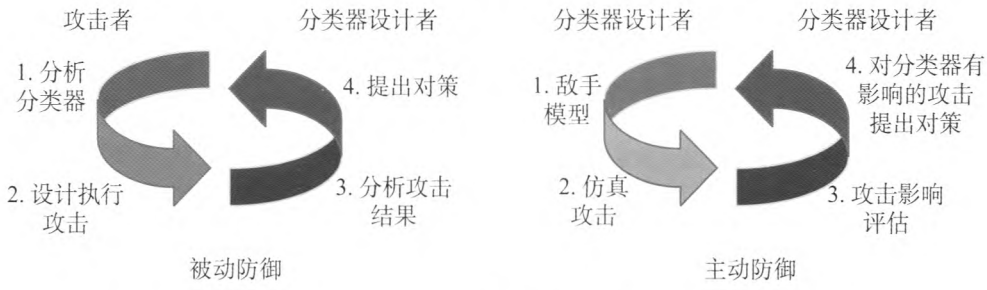

Fig.4　A conceptual representation of reactive defense and proactive defense

图4　被动防御和主动防御概念图

3.1　机器学习安全评估机制

面对如此多的针对机器学习的攻击方法，传统的针对机器学习算法的评估机制显出一些不足。传统的评估机制主要是面向分类性能评估，没有对分类器的安全性进行评估。因此，很多科学家提出了包含安全性/鲁棒性评估的机器学习算法评估机制[15,42]。

安全评估机制主要是基于假设-对策（what-if）分析方法[43]，设计者对分类器的缺陷提出攻击假设，再针对特定的攻击寻找对策来完善分类器，其实质是一种主动防御（proactive defense）的安全评估机制。主动防御是在被动防御（reactive defense）的基础上发展而来的。图4给出了这两种防御机制的概念图[15]。

被动防御是一种比较常见的防御机制，攻击者通过对分类器的分析，不断尝试发现合适的攻击策略，进而实施最优的攻击；而设计者又会尽快地对出现的新的对抗样本进行分析，尽可能地更新分类器，典型的更新方法有重新收集数据进行训练或者加入新的特征来检测最新出现的攻击。如此攻击和防御的过程循环交替进行，形成一种竞赛（race）。主动防御的过程与被动攻击基本一致，只是开展攻防技术竞赛（arms race）的双方主体都是分类器设计者。在部署设计好的分类器之前，设计者通过假设存在敌手对分类器进行渗透。首先通过2.2节中机器学习的敌手模型假设具体的敌手目标、知识、能力和策略，进而找出特定敌手模型下分类器可能存在的缺陷和攻击威胁，然后再提出合适的对策加入分类器的设计。这种机制在系统发布前对其进行渗透测试，大大提升了系统的安全性，因此可以作为很好的安全评估机制。

在攻击数据的干扰下，会使训练数据和测试数据在分布上有比较大的差异，形成一种不平稳（non-stationary）的数据分布。可以通过数据分布的异常情况来考虑是否存在潜在的攻击数据，这可以作为系统安全性能评估的一个指标。在主动防御机制的基础上，Biggio等人提出考虑训练数据和测试数据的分布情况进行分类器的安全评估[15,42,44]。图5给出一种

在特定敌手模型下考虑训练和测试数据分布，应用what-if策略进行分析的安全评估框架。首先假定一个可以产生任何攻击场景的敌手模型；然后建立具体攻击场景（敌手目标、知识、能力、策略）下的数据分布模型 P；接着根据原始的数据集用重采样技术生成 k 对数据的数据集（D_{TRi},D_{TSi}），$i=1,2,\cdots,k$；然后从 D_{TRi} 和 D_{TSi} 中筛选符合具体攻击场景下数据分布模型 P 的训练集 TR 和测试集 TS；最后用重新筛选的数据集 TR 和 TS 来对系统进行安全评估；在此基础上假设不同的敌手模型来进行循环的渗透和评估。

Fig.5 Security evaluation mechanism considering data distribution

图5 考虑数据分布的安全评估机制

3.2 训练过程的防御技术

针对训练过程的投毒攻击行为，其主要采用在系统的训练过程中注入恶意数据的手段，因此其防御技术主要采用数据清洗（data sanitization）[12]或者提高学习算法鲁棒性[45-47]来对抗恶意训练数据。

数据清洗主要是对恶意的训练数据进行筛选移除，是一种比较直接和实际的防御手段。文献[48]提到一个针对垃圾邮件检测系统的拒绝消极影响（reject on negative impact）防御方法。该方法先将一个候选样本加入基础的训练数据集以得到一个新训练数据集，之后用新训练数据集学习出新分类器，再用同一个测试数据集来评估新分类器和原分类器（基础训练集学习得到）的分类性能。如果新分类器的错误率和原分类器的错误率相比差很多，则认为这个候选数据是一个恶意数据，将其移除训练集，否则加入训练集。该方法在垃圾邮件检测系统上进行实验，取得了较好的效果，但其操作比较繁琐，计算量比较大，不适合大规模候选集合的筛选。

另一种防御方法是通过提高算法的鲁棒性来限制恶意训练数据的影响。Biggio等人提出了[46-47]Bagging（bootstrap aggregating）和RSM（random subspace method）两种防御投毒攻击的方法。Rubinstein等人提出了一种ANTIDOTE的方法[45]。该方法主要针对原来的PCA算法进行了改进，形成了最大化中位值绝对偏差（median absolute deviation）的PCA-GRID算法，并结合Laplace截断阈值的方法来降低恶意训练数据的影响。这3种方法在拒绝恶意训练数据上都取得了较好的效果。

3.3 测试/推理过程的防御技术

相比训练过程，针对测试/推理过程的防御技术主要集中在对学习算法鲁棒性的提高上。从安全评估机制可以看出，针对攻击的防御技术可以很自然地想到攻击者和防御者之间的博弈。而在对测试过程的防御技术中，提出了很多博弈论的方法。Globerson等人最早[33,49]根据min-max方法提出了一个称为Invariant SVM的算法来处理在测试时最坏情况（worse-case）下的特征操纵（加入特征、删减特征和修改特征）行为。Bruckner等人提出的斯塔克尔伯格博弈（Stackelberg games）[50]和基于纳什均衡博弈[51]的NashSVM算法，都在一定程度上提高了算法的鲁棒性。Biggio等人基于这些工作进行拓展，提出了随机预测博弈（randomized prediction games）[52]。该方法考虑了攻击者和防御者双方策略集合概率分布的参数、边界等，还结合随机化的思路设计随机化分类函数，迫使攻击者选择低成效的攻击策略。

另一类防御措施是采用考虑数据分布的主动防御机制，称作对抗训练（adversarial training）。测试/推理过程出现的对抗样本实质上是改变了测试数据的样本分布，使其与训练数据集的样本分布相差过大，导致分类器的误分类。因此，结合主动防御的策略，可以通过在训练时加入对抗样本[21,23]的方式仿真测试时可能的数据分布。Szegedy等人[21]提出在训练集中注入带有完全标注的对抗样本，这种混合了合法样本和对抗样本训练出的模型有更强的鲁棒性。Goodfellow等人也通过对抗训练的方法在MNIST数据集上，使模型对对抗样本识别的错误率从89.4%降到17.9%[53]。

相对于在训练集中加入对抗样本，也有通过平滑模型输出的机制来加强模型在小扰动下的鲁棒性。文献[54]提出一种深度收缩网络(deep contractive networks)的机器学习模型，通过使用更平滑的惩罚项来训练模型提高系统的鲁棒性。此外，防御精馏(defensive distillation)技术也可以实现平滑的目的。

Papernot等人在之前提出的精馏(distillation)技术的基础上，针对DNN中的对抗样本提出了针对训练过程的防御精馏技术[55]。该方法通过从原来的DNN精馏提取出来的类别概率知识，加入到一个小的DNN中，从而在保持分类精度的同时提高模型泛化能力，降低了输入扰动对分类器模型的影响，可以训练出对输入扰动不敏感的平滑分类器模型，提高模型的鲁棒性。在MNIST和CIFAR10两个数据集上，对初始分类性能优异的两种DNN在应用防御精馏技术前后对敌对样本的辨别能力进行了对比：针对MNIST数据集的DNN的前后对比，使对抗样本成功欺骗率从95.86%降到0.45%，而对CIFAR-10数据集则从87.89%降到5.11%，说明了这种方法对特定敌对样本的有效检测。但这种防御精馏方法只能对抗有限的对抗样本，文献[56]提到一种对抗样本的变种精馏是解决不了的。由此可见，需要进一步研究更加有效的安全学习算法。

3.4 隐私保护的机器学习技术

现在机器学习模型的训练需要大量的数据，尤其是DNN的训练。很多数据的收集方式都是通过众包(crowdsourcing)技术获得的，里面包含大量的用户隐私信息(比如照片和语音信息)，甚至是敏感信息(比如医疗数据)。出于对这些信息安全性的考虑，数据收集方需要对这些数据无限期保存，如何低成本且高效地保护这些数据是一个重要的问题。此外，这些训练数据还面临着被攻击者窃取的可能，前面介绍的逆向攻击已经说明了这种威胁的存在。因此，对数据的隐私保护是机器学习防御的一个关键技术。

加密技术是保护数据隐私的关键，差分隐私(differential privacy)技术就是一种通过加密数据来保护隐私的方式。差分隐私模型是一种被广泛认可的严格的隐私保护模型，它通过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。差分隐私最早是由Dwork[57]提出的一个概念，该算法的安全性和鲁棒性有很好的理论支撑。Erlingsson等人提出了一种从用户终端软件中获得匿名的具有强壮的隐私保证的众包技术——RAPPOR(randomized aggregatable privacy-preserving ordinal response)[58]。该技术采用随机响应(randomized response)的安全机制结合差分隐私技术，实现了高效的隐私安全保证。其具体的实现过程为：对于给定客户的变量 v，RAPPOR算法通过客户的机器执行，在真实值 v 的基础上叠加编码形成一些“噪音”，这些噪音形成一个数组，再将这个数组上传给服务器，而 v 的这种噪音形式的表达就可以透露出变量 v 的一些需要的可控的信息量，在保留重要信息的同时保证了数据的安全性。

此外，同态加密(homomorphic encryption)技术[59]也通过对数据的加密实现了对数据隐私的保护。同态加密允许用户直接对密文进行特定的代数运算，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，因此其在对云端存储数据的隐私保护方面有很大的应用前景。近年来，很多基于同态加密的安全多方计算[60]和对原有机器学习算法近似这两种策略的加密机器算法被提出。比如：2015年Aslett等人[61]提出了对全同态加密数据进行分类的完全随机森林(complete random forest, CRF)算法、朴素贝叶斯算法和逻辑回归算法；2017年姚禹丞等人[62]提出了一种基于同态加密的分布式 K 均值聚类算法；2016年Dowlin等人[63]提出了一种可应用于加密数据的近似神经网络CryptoNets，并在MNIST手写识别数据集上进行测试，高效地实现了准确率为99%的分类性能。

除了用数据加密方法来防御一些攻击行为之外，尽量减少机器学习模型API接口的敏感信息输出[32]也是一种保护数据隐私性的思路。

4 结束语

随着大数据、物联网、云计算、人工智能等应用

领域的蓬勃发展,机器学习技术扮演着越来越重要的角色。相应地,针对机器学习的安全性问题及其防御机制受到了学术界和工业界的极大关注。本文对相关工作进行调研和分析,认为机器学习安全性问题及其防御技术研究有如下发展趋势:

(1)针对机器学习的新型安全威胁将不断涌现。当前,机器学习技术研究处在一个空前火热的阶段,大量学习框架、算法及其优化机制被提出。然而现有机器学习模型和算法很少考虑到自身的安全性问题,面临着较大的安全风险[16-26,28,32]。另一方面,基于统计的机器学习方法严重依赖于数据,容易遭受对抗样本和数据统计不完全的影响。此外,对抗样本难以收集和预测的特点使得设计基于机器学习技术的对抗样本检测方法变得困难。因此,针对机器学习的新型安全威胁必将引起更多关注。

(2)敌手环境中机器学习系统安全评估成为一个新的研究热点。随着机器学习安全威胁的不断涌现,机器学习系统设计者将越来越重视其系统安全性能的评估。目前机器学习系统安全评估的标准还在探索和讨论之中[15,42],需要制定更加完善且具有权威性的安全评估机制。因此,考虑建立完善且权威的安全评估机制也是亟待解决的一个问题。

(3)隐私保护的机器学习方法是提升学习模型安全性的重要途径。当前,差分隐私[58]、同态加密[59-63]等隐私保护技术已取得了一些研究成果。然而现有的隐私保护技术依然面临着效率代价比不高的问题。因此,研究高效费比的模型隐私保护技术也是一个重要的课题。

(4)安全的深度学习模型是机器学习安全性问题研究中的一个新增长点。当前,已经有大量的研究工作指出DNN的反直觉特性将导致DNN面临很多的安全威胁[7-8,22-26],虽然已有一些学者提出使用在训练过程加入对抗样本[21,53]以及提高算法鲁棒性的方式[54-55]来弥补上述特性所带来的安全缺陷,但是这些解决方案依然没能很好地解决这一安全问题。因此,考虑发展更接近人脑思考方式的神经网络算法(如加入了先验信息的贝叶斯深度网络[64]),进一步建立安全的深度学习模型是一个很值得研究的课题。

(5)在新型机器学习算法设计过程中,安全性、泛化性能和算法开销需要综合考虑、联合优化。通常来讲,增强安全性将增加算法开销,甚至降低泛化能力,从而影响算法的实用性[65]。因此,在安全的机器学习算法设计过程中,要合理地权衡算法的安全性和实用性,以更好地满足实际应用需求。

References:

[1] Wittel G L, Wu S F. On attacking statistical spam filters[C]// Proceedings of the 1st Conference on Email and Anti-Spam, Mountain View, Jul 30-31, 2004.

[2] Kayacik H G, Zincir-Heywood A N, Heywood M I. Automatically evading IDS using GP authored attacks[C]//Proceedings of the 2007 IEEE Computational Intelligence in Security & Defense Applications, Honolulu, Apr 1-5, 2007. Piscataway: IEEE, 2007: 153-160.

[3] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Oct 24-28, 2016. New York: ACM, 2016: 1528-1540.

[4] Biggio B, Didaci L, Fumera G, et al. Poisoning attacks to compromise face templates[C]//Proceedings of the 6th International Conference on Biometrics, Madrid, Jun 4-7, 2013. Piscataway: IEEE, 2013: 1-7.

[5] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing[C]//Proceedings of the 23rd USENIX Security Symposium, San Diego, Aug 20-22, 2014. Berkeley: USENIX Association, 2014: 17-32.

[6] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, Oct 12-16, 2015. New York: ACM, 2015: 1322-1333.

[7] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. arXiv: 1602.02697, 2016. https://arxiv.org/pdf/1602.02697.

[8] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands

[C]//Proceedings of the 25th USENIX Security Symposium, Austin, Aug 10-12, 2016. Berkeley: USENIX Association, 2016: 513-530.

[9] Dalvi N, Domingos P, Mausam, et al. Adversarial classification[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Aug 22-25, 2004. New York: ACM, 2004: 99-108.

[10] Lowd D, Meek C. Adversarial learning[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Aug 21-24, 2005. New York: ACM, 2005: 641-647.

[11] Barreno M, Nelson B, Sears R, et al. Can machine learning be secure? [C]//Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, Taipei, China, Mar 21-24, 2006. New York: ACM, 2006: 16-25.

[12] Huang Ling, Joseph A D, Nelson B, et al. Adversarial machine learning[C]//Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, Chicago, Oct 21, 2011. New York: ACM, 2011: 43-58.

[13] Yan Youbiao, Chen Yuanyan. A survey on machine learning and its main strategy[J]. Application Research of Computers, 2004, 21(7): 4-10.

[14] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning[J]. Machine Learning, 2010, 81(2): 121-148.

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

[16] Srndic N, Laskov P. Detection of malicious PDF files based on hierarchical document structure[C]//Proceedings of the 20th Annual Network & Distributed System Security Symposium, San Diego, Feb 24-27, 2013: 1-16.

[17] Biggio B, Pillai I, Ariu D, et al. Is data clustering in adversarial settings secure?[C]//Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security, Berlin, Nov 4, 2013. New York: ACM, 2013: 87-98.

[18] Biggio B, Bulò S R, Pillai I, et al. Poisoning complete-linkage hierarchical clustering[C]//LNCS 8621: Proceedings of the 2014 International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, Joensuu, Aug 20-22, 2014. Berlin, Heidelberg: Springer, 2014: 42-52.

[19] Biggio B, Rieck K, Ariu D, et al. Poisoning behavioral malware clustering[C]//Proceedings of the 2014 Artificial Intelligent and Security Workshop, Scottsdale, Nov 7, 2014. New York: ACM, 2014: 27-36.

[20] Zhao Wentao, Long Jun, Yin Jianping, et al. Sampling attack against active learning in adversarial environment[C]//LNCS 7647: Proceedings of the 9th International Conference on Modeling Decisions for Artificial Intelligence, Catalonia, Nov 21-23, 2012. Berlin, Heidelberg: Springer, 2012: 222-233.

[21] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. Computer Science, arXiv: 1312.6199, 2013.

[22] Nguyen A M, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//Proceedings of the 2015 Conference on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 427-436.

[23] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks [C]//Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 2574-2582.

[24] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 European Symposium on Security and Privacy, Saarbrücken, Mar 21-24, 2016. Piscataway: IEEE, 2016: 372-387.

[25] Liu Yanpei, Chen Xinyun, Liu Chang, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. arXiv: 1611.02770, 2016. https://arxiv.org/pdf/1611.02770.

[26] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[EB/OL]. arXiv: 1607.02533, 2016. https://arxiv.org/pdf/1607.02533.

[27] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines[C]//Proceedings of the 29th International Conference on International Conference on Machine Learning, Edinburgh, Jun 26-Jul 1, 2012. Madison: Omnipress, 2012: 1467-1474.

[28] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time[C]//LNCS 8190: Proceedings of the 2013 European Conference on Machine Learning and Knowledge Discovery in Databases, Prague, Sep 23-27, 2013. Berlin, Heidelberg: Springer, 2013: 387-402.

[29] Biggio B, Fumera G, Roli F, et al. Poisoning adaptive biometric systems[C]//LNCS 7626: Proceedings of the 2012 International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Hiroshima, Nov 7-9, 2012. Berlin, Heidelberg: Springer, 2012: 417-425.

[30] Huang Xiao, Biggio B, Brown G, et al. Is feature selection secure against training data poisoning?[C]//Proceedings of the 32nd International Conference on Machine Learning, Lille, Jul 6-11, 2015: 1689-1698.

[31] Yang Chaofei, Wu Qing, Li Hai, et al. Generative poisoning attack method against neural networks[EB/OL]. arXiv: 1703.01340, 2017. https://arxiv.org/pdf/1703.01340.

[32] Tramèr F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Security Symposium, Austin, Aug 10-12, 2016. Berkeley: USENIX Association, 2016: 601-618.

[33] Globerson A, Roweis S. Nightmare at test time: robust learning by feature deletion[C]//Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006. New York: ACM, 2006: 353-360.

[34] Biggio B, Fumera G, Roli F. Pattern recognition systems under attack: design issues and research challenges[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28(7): 1460002.

[35] Zhu Xinzhong. Super-class discriminant analysis: a novel solution for heteroscedasticity[J]. Pattern Recognition Letters, 2013, 34(5): 545-551.

[36] Kloft M, Laskov P. Security analysis of online anomaly detection[J]. Journal of Machine Learning Research, 2010, 13 (1): 3681-3724.

[37] Mei Shike, Zhu Xiaojin. The security of latent Dirichlet allocation[C]//Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, San Diego, May 9-12, 2015: 681-689.

[38] Mozaffari-Kermani M, Sur-Kolay S, Raghunathan A, et al. Systematic poisoning attacks on and defenses for machine learning in healthcare[J]. IEEE Journal of Biomedical and Health Informatics, 2015, 19(6): 1893-1905.

[39] Hu Weiwei, Tan Ying. Generating adversarial malware examples for black-box attacks based on GAN[EB/OL]. arXiv: 1702.05983, 2017. https://arxiv.org/pdf/1702.05983.

[40] Zhang Fei, Chan P P, Biggio B, et al. Adversarial feature selection against evasion attacks[J]. IEEE Transactions on Cybernetics, 2016, 46(3): 766-777.

[41] Liu Qiang, Zhou Sihang, Zhu Chengzhang, et al. MI-ELM: highly efficient multi-instance learning based on hierarchical extreme learning machine[J]. Neurocomputing, 2016, 173: 1044-1053.

[42] Biggio B, Corona I, Nelson B, et al. Security evaluation of support vector machines in adversarial environments[M]// Ma Yunqian, Guo Guodong. Support Vector Machines Applications. Berlin, Heidelberg: Springer, 2014: 105-153.

[43] Rizzi S. What-if analysis[M]//Liu Ling, Özsu M T. Encyclopedia of Database Systems. Springer US, 2009: 3525-3529.

[44] Laskov P, Kloft M. A framework for quantitative security analysis of machine learning[C]//Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence, Chicago, Nov 9, 2009. New York: ACM, 2009: 1-4.

[45] Rubinstein B I P, Nelson B, Huang Ling, et al. ANTIDOTE: understanding and defending against poisoning of anomaly detectors[C]//Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference, Chicago, Nov 4-6, 2009. New York: ACM, 2009: 1-14.

[46] Biggio B, Fumera G, Roli F. Multiple classifier systems for robust classifier design in adversarial environments[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1/4): 27-41.

[47] Biggio B, Corona I, Fumera G, et al. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks [C]//LNCS 6713: Proceeding of the 10th International Workshop on Multiple Classifier Systems, Naples, Jun 15-17, 2011. Berlin, Heidelberg: Springer, 2011: 350-359.

[48] Nelson B, Barreno M, Chi F J, et al. Misleading learners: co-opting your spam filter[M]//Yu P S, Tsai J J P. Machine Learning in Cyber Trust. Springer US, 2009: 17-51.

[49] Teo C H, Globerson A, Roweis S T, et al. Convex learning with invariances[C]//Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, Dec 3-6, 2007. Red Hook: Curran Associates, 2007: 1489-1496.

[50] Brückner M, Scheffer T. Stackelberg games for adversarial prediction problems[C]//Proceedings of the 17th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, San Diego, Aug 21-24, 2011. New York: ACM, 2011: 547-555.

[51] Brückner M, Kanzow C, Scheffer T. Static prediction games for adversarial learning problems[J]. Journal of Machine Learning Research, 2012, 13(1): 2617-2654.

[52] Bulò S R, Biggio B, Pillai I, et al. Randomized prediction games for adversarial machine learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 99: 1-13.

[53] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[EB/OL]. arXiv: 1412.6572, 2014. https://arxiv.org/pdf/1412.6572.

[54] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples[EB/OL]. arXiv: 1412.5068, 2014. https://arxiv.org/pdf/1412.5068.

[55] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the 2016 IEEE Symposium on Security and Privacy, San Jose, May 22-26, 2016. Washington: IEEE Computer Society, 2016: 582-597.

[56] Carlini N, Wagner D. Defensive distillation is not robust to adversarial examples[EB/OL]. arXiv: 1607.04311, 2016. https://arxiv.org/pdf/1607.04311.

[57] Dwork C. Differential privacy[C]//LNCS 4052: Proceedings on the 33rd International Colloquium on Automata, Languages and Programming, Venice, Jul 10-14, 2006. Berlin, Heidelberg: Springer, 2006: 1-12.

[58] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: randomized aggregatable privacy-preserving ordinal response[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, Nov 3-7, 2014. New York: ACM, 2014: 1054-1067.

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

[60] Zhang Wenke, Yang Yong, Yang Yu. Secure multi-party computation[J]. Information Security and Communications Privacy, 2014(1): 97-99.

[61] Aslett L J M, Esperança P M, Holmes C C. Encrypted statistical machine learning: new privacy preserving methods [EB/OL]. arXiv: 1508.06845, 2015. https://arxiv.org/pdf/1508.06845.

[62] Yao Yucheng, Song Ling, E Chi. Investigation on distributed *k*-means clustering algorithm of homomorphic encryption [J]. Computer Technology and Development, 2017, 27(2): 81-85.

[63] Gilad-Bachrach R, Dowlin N, Laine K, et al. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, Jun 19-24, 2016: 201-210.

[64] Wang Hao, Yeung D Y. Towards Bayesian deep learning: a framework and some existing methods[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3395-3408.

[65] Kifer D, Machanavajjhala A. No free lunch in data privacy [C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, Athens, Jun 12-16, 2011. New York: ACM, 2011: 193-204.

附中文参考文献:

[13] 闫友彪, 陈元琰. 机器学习的主要策略综述[J]. 计算机应用研究, 2004, 21(7): 4-10.

[60] 张文科, 杨勇, 杨宇. 安全多方计算研究[J]. 信息安全与通信保密, 2014(1): 97-99.

[62] 姚禹丞, 宋玲, 鄂驰. 同态加密的分布式*K*均值聚类算法研究[J]. 计算机技术与发展, 2017, 27(2): 81-85.

LI Pan was born in 1992. He is an M.S. candidate at National University of Defense Technology. His research interests include machine learning and cyber security.

李盼(1992—),男,河北邯郸人,国防科技大学硕士研究生,主要研究领域为机器学习,网络安全。

ZHAO Wentao was born in 1973. He received the Ph.D. degree from National University of Defense Technology in 2008. Now he is a professor at National University of Defense Technology. His research interest is cyberspace security situation awareness.

赵文涛(1973—),男,内蒙凉城人,2008年于国防科技大学获得博士学位,现为国防科技大学教授,主要研究领域为网络空间安全态势感知。

LIU Qiang was born in 1986. He received the Ph.D. degree from National University of Defense Technology in 2014. Now he is a lecturer at National University of Defense Technology. His research interests include 5G network, Internet of things, wireless network security and machine learning.

刘强(1986—),男,江西临川人,2014年于国防科技大学获得博士学位,现为国防科技大学讲师,主要研究领域为5G网络,物联网,无线网络安全,机器学习。

CUI Jianjing was born in 1994. He is an M.S. candidate at National University of Defense and Technology. His research interests include cyber security and machine learning.

崔建京(1994—),男,山东即墨人,国防科技大学硕士研究生,主要研究领域为网络安全,机器学习。

YIN Jianping was born in 1963. He received the Ph.D. degree from National University of Defense Technology in 1990. Now he is a professor at National University of Defense Technology. His research interests include artificial intelligence, network algorithm and information security.

殷建平(1963—),男,湖南益阳人,1990年于国防科技大学获得博士学位,现为国防科技大学教授,主要研究领域为人工智能,网络算法,信息安全。