软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(5):1526−1546 [doi: 10.13328/j.cnki.jos.006204]

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

面向智能攻击的行为预测研究*

马钰锡[1], 张全新[1], 谭毓安[1], 沈 蒙[2,3]

[1](北京理工大学 计算机学院,北京 海淀 100081)

[2](北京理工大学 网络空间安全学院,北京 海淀 100081)

[3](鹏城实验室 网络空间安全研究中心,广东 深圳 518055)

通讯作者: 沈蒙, E-mail: shenmeng@bit.edu.cn

摘 要: 人工智能的迅速发展和广泛应用促进了数字技术的整体跃升.然而,基于人工智能技术的智能攻击也逐渐成为一种新型的攻击手段,传统的攻击防护方式已经不能满足安全防护的实际需求.通过预测攻击行为的未来步骤,提前部署针对性的防御措施,可以在智能攻击的对抗中取得先机和优势,有效保护系统安全.首先界定了智能攻击和行为预测的问题域,对相关研究领域进行了概述;然后梳理了面向智能攻击的行为预测的研究方法,对相关工作进行分类和详细介绍;之后,分别阐述了不同种类的预测方法的原理机制,并从特征及适应范围等角度对各个种类的方法做进一步对比和分析;最后,展望了智能攻击行为预测的挑战和未来研究方向.

关键词: 攻击预测;行为预测;智能攻击;攻击行为;人工智能

中图法分类号: TP309

中文引用格式: 马钰锡,张全新,谭毓安,沈蒙.面向智能攻击的行为预测研究.软件学报,2021,32(5):1526−1546. http://www.jos.org.cn/1000-9825/6204.htm

英文引用格式: Ma YX, Zhang QX, Tan YA, Shen M. Research on behavior-prediction for intelligent attacks. Ruan Jian Xue Bao/Journal of Software, 2021,32(5):1526−1546 (in Chinese). http://www.jos.org.cn/1000-9825/6204.htm

Research on Behavior-Prediction for Intelligent Attacks

MA Yu-Xi[1], ZHANG Quan-Xin[1], TAN Yu-An[1], SHEN Meng[2,3]

[1](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[2](School of Cyberspace Security, Beijing Institute of Technology, Beijing 100081, China)

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

Abstract: The rapid development and broad application of artificial intelligence have promoted the overall leap in digital technology. However, intelligent attacks based on artificial intelligence technology have gradually become a new type of attack method. Traditional attack protection methods have been far from meeting the requirements of security protection. By predicting the future steps of the attack behavior and deploying targeted defense measures in advance, the opportunities and advantages can be obtained in the confrontation of intelligent attacks and system security is effectively protected. This study first defines the problem domain of behavior-prediction and intelligent attacks and outlines its related research areas. Then it combs the research methods of behavior-prediction for intelligent attacks, and introduces the classification and related work in detail. After that, the principle and mechanism of different types of prediction methods are explained, respectively. Each type's methods are further compared, discussed, and analyzed from the perspective of characteristics and adaptation scope. Finally, the challenges and future directions of intelligent attack behavior-prediction are prospected.

* 基金项目: 国家自然科学基金(61876019, 61972039); 北京市自然科学基金(4192050); 北京市科技新星计划(Z201100006820006); 广东省重点领域研发计划(2019B010136001); 之江实验室开放课题(2020AA3AB04)

Foundation item: National Natural Science Foundation of China (61876019, 61972039); Natural Science Foundation of Beijing Municipality (4192050); Beijing Nova Program of Science and Technology (Z201100006820006); Key Research and Development Program of Guangdong Province (2019B010136001); Opening Fund of Zhejiang Lab. (2020AA3AB04)

收稿时间: 2020-06-17; 修改时间: 2020-09-16, 2020-10-26; 采用时间: 2020-11-17; jos 在线出版时间: 2020-12-02

Key words: attack prediction; behavior prediction; intelligent attack; attack behavior; artificial intelligence

近年来,人工智能(artificial intelligence,简称 AI)蓬勃发展,在文本、语音、计算机视觉等领域取得了重要进展.AI 驱动的系统实现了手工任务的自动化、生产效率的提高和自主决策的增强,AI 技术覆盖了社会行业的各个方面,促进了新一轮的产业变革和科技革命[1].

然而,AI 技术如果被恶意利用,同样会带来巨大的负面影响,攻击者不仅会针对 AI 系统本身,还会自行利用 AI 技术来增强自己在其他领域的犯罪活动[2].具体表现在以下方面.

(1) 数字安全:AI 技术降低了网络攻击门槛,同时丰富了攻击模式,以自动化方式提升复杂攻击的速度与效率,加剧了鱼叉式网络钓鱼等劳动密集型网络攻击的危害.由 AI 技术驱动的自主程序能够探测安全系统和网络,搜索可能被利用的未发现的漏洞,使网络钓鱼(phishing)、语音合成、拒绝服务(distributed denial of service,简称 DDoS)等网络攻击趋于智能化,如利用语音合成冒充本人、利用现有软件漏洞进行自动黑客攻击,以及利用 AI 系统漏洞样本投毒、生成对抗样本等[3].

(2) 物理安全:利用 AI 自动操控无人机以及全自动和半自动驾驶系统构成新型风险,包括多辆无人车故意撞击,多架无人机进行协同攻击,控制关键基础设施等攻击行为.此外,还会出现新型攻击方式,如破坏网络物理系统,使无人车轨道漂移或目标检测系统故障;利用无法远程控制的物理系统进行攻击,如微型无人机群协同攻击等[4].

(3) 政治安全:利用 AI 分析收集到的海量社交、媒体数据并进行自动监测,进而开展高度精准的针对性欺骗;或利用 AI 分析人类的行为、情绪和信仰,发起逼真的虚假宣传活动等,使社交工程攻击更加复杂化,涉及隐私入侵和社交操纵的威胁增大[5,6].

基于 AI 技术的智能攻击正逐渐发展为社会各个领域中的隐患,成为众多科研工作者的研究热点[7].然而,智能攻击的随机性和不确定性,使得以此为基础的系统或网络安全态势变化为复杂的非线性过程,传统安全防护的方式,例如端对端拦截、攻击检测等已经远远不能满足安全防护的实际需求[8,9].传统的安全防护通常在攻击发生之后,根据攻击的规则、特征等进行快速防御,其主要的防护对象是系统终端和数据库等,主要目的是阻止权限劫持、信息数据泄露以及系统侵占等;而随着部分网络攻击基本要素的更新,攻击手段的自动化、智能化,传统防护技术的局限性逐渐显露,例如不可追踪、滞后响应、被动防御等,在与智能攻击方式的对阵中已经处于劣势地位,使得越来越多的研究正在朝智能预测的方向发展[10,11].

智能攻击一般由多个事件构成,具有一定的顺序和逻辑关系,而行为预测技术根据当下已检测到的报警日志,例如攻击动作、目标、步骤等信息,预测该智能攻击未来即将发生的攻击行为,建立动态的响应机制,以检测、预测、响应、防护为组成过程,使系统提前做出针对性的防御措施.同时,通过训练模型的方式建立机器的自动感知和自学习机制,能够使防护系统具有自主行为和思维能力,包括对复杂的复合式攻击行为的识别和预测,为防护系统安全提供主动、动态、实时、快速响应的安全屏障,有效提高系统的安全性[12].

本文对面向智能攻击的行为预测方法进行了全面的调查和论述.第 1 节首先给出智能攻击和行为预测的基础定义以及可行性分析等相关内容,归纳总结不同类别智能攻击的攻击方式和引发原因,界定智能攻击行为预测的问题域,并对其相关研究领域进行概述.第 2 节从技术原理的角度将目前主流的智能攻击行为的预测方法分类为基于神经网络的预测方法、基于博弈论的预测方法、基于攻击图的预测方法、基于数据挖掘的预测方法以及其他预测方法.第 3 节中根据第 2 节的研究工作概述,分别阐述不同种类预测方法的原理机制,并从特征和适应范围等方面对各种方法进行对比、讨论和分析.第 4 节展望未来的研究方向和发展趋势.最后,第 5 节总结全文.

1 智能攻击与行为预测

1.1 智能攻击

智能攻击的内涵丰富而复杂,目前对智能攻击的广泛定义包括:(1) 通过输入恶意样本欺骗 AI 算法,使得 AI

模型输出非预期的结果,包括黑盒攻击和白盒攻击等多种方式[13];(2) 基于 AI 技术,对网络传输、智能设备、社交媒体等不同领域进行的入侵行为,例如高级鱼叉式钓鱼攻击、利用 AI 算法破解验证码等[14];(3) 传统的攻击方式经过推移和演变,呈现出大规模、协同、多阶段的特征,形成趋于自动化和智能化的新一代智能攻击行为[15].

与传统的攻击方式相比,智能攻击的隐蔽性更强,危险性更高,因而对系统或网络造成的威胁更大.例如,传统网络攻击中,攻击规模和攻击效率难以兼顾,而使用 AI 自动执行网络攻击在很大程度上可以兼顾多个方面,使鱼叉式网络钓鱼、分布式拒绝服务等传统劳动密集型和成本高昂的网络攻击转型并带来更大的威胁;同时,智能攻击会利用 AI 技术使恶意软件获取对上下文环境的理解,然后通过学习被感染系统的环境信息,尝试尽可能地将攻击目标的环境融入到攻击中[16].

基于以上定义,本节分别挑选 3 种类型中具有代表性的智能攻击作详细阐述.首先是对 AI 系统或模型本身进行攻击的对抗攻击,攻击者利用“数据投毒”“部署后门”等方式设计恶意对抗样本,干扰 AI 算法的正确输出.然后介绍基于 AI 技术的智能攻击代表:恶意软件逃逸和智能 APT 攻击,二者均利用 AI 技术训练自身模型,对特定的用户或系统漏洞发动针对性攻击,例如,恶意软件逃逸通过在良性程序和软件中混淆或隐藏恶意代码片段躲避防御模型的安全检测,而智能 APT 攻击中典型的钓鱼攻击,通过学习社交网络中用户数据,模拟用户邮件书写风格、手机操作行为等发起指定性的攻击,却很难被防护系统检测和识别.最后阐述的是由传统攻击方式演变的自动化僵尸网络的智能攻击,传统僵尸网络随着网络要素的更新和 AI 技术的发展,实现了在攻击过程中自我学习,进而自动化决策进行繁殖和传播,构成新型的自动化僵尸网络,在以下内容将进行具体阐述.

1.1.1 对抗攻击

智能攻击首先体现在对人工智能系统本身的攻击,攻击者通过设计针对性的数值型向量(numeric vectors)在原始样本中,干扰机器学习模型的识别和分类,达到其攻击目的.这样的攻击通常称为对抗攻击,攻击样本称为对抗样本[17].具体而言,对抗样本的产生主要因为训练模型的样本集只能覆盖部分,并非包含所有的可能性,因而不能训练出覆盖全部样本特征的模型,这就导致训练的模型边界与真实决策边界不完全一致,两者之间产生的差异即为对抗样本的空间.

图 1 中曲线与分界线包裹区域为对抗样本空间,模型训练对 A,B 两类样本分类曲线的边界与真实决策边界并不重合,其中的样本会导致模型识别出现错误,因此不重合区域即被定义为对抗样本空间.对抗攻击就是寻求特殊的途径生成上述区域中对抗样本,干扰模型的分类过程.

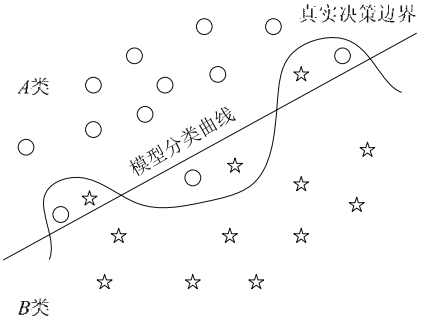

Fig.1 Adversarial sample space

图 1 对抗样本空间

当前,基于机器学习、深度学习的多数 AI 算法严重依赖于训练集的样本分布.而在真实场景下,攻击者通过人为制造噪声,改变数据分布,或是生成恶意的对抗样本,即“投毒攻击”,干扰模型的分类过程,造成 AI 系统的识别错误.例如,无论是在交通标志上贴上噪声图片以欺骗无人车目标检测的分类器[18],还是骗取无人机在侦察任务中寻找敌人的活动[19],抑或颠覆内容过滤器以将恐怖分子招募宣传发布到社交网络[20]等,都给社会安全带来了极大的威胁.

1.1.2 恶意软件逃逸

恶意软件通常采用人工方式生成,由攻击者通过编写脚本来产生特洛伊木马和网络病毒,然后利用密码抓取、Rootkit 和其他工具协助执行和散播.对恶意软件的检测通常使用机器学习方法,利用从恶意软件样本中检索到的数据,如指令序列、特殊字段、或者原始字节进行学习,从而建立区分恶意或良性软件的模型[21].

然而,攻击者通过观察和学习安全防护系统的过滤规则和决策策略,利用其作为基础知识开发“最小程度被检测出”的恶意软件,例如在良性程序和软件中混淆或隐藏恶意代码片段等,这样的攻击行为被称为恶意软件逃逸,又称规避攻击.Kolosnjaji 等人[22]提出对输入数据的微小更改会导致在测试时发生错误分类,他们使用深度网络(deep neural network,简称 DNN)从应用程序的原始字节中学习恶意代码检测方法所存在的漏洞,提出了一种基于梯度的攻击:通过在每个恶意软件样本末尾更改少量特定字节,既能够实现其入侵功能,同时又可以逃避 DNN 的安全检测.实验结果表明:仅修改少于 1%的特定字节,对抗性恶意软件的二进制文件就能够高概率躲避安全检测.

1.1.3 自动化僵尸网络

由于具备分布式特性,僵尸网络的出现对 Internet 构成了严重的威胁,它不但可以使用从计算机到任何已连接设备的任何未打补丁的机器,而且经常被用作执行各种网络威胁的机制,例如进行 DDoS 攻击、网络钓鱼、生成和分发勒索软件等.

对僵尸网络的检测和防御工作通常利用被动监视技术(例如蜜罐)实现[23],但随着网络要素的更新和 AI 技术的发展,攻击者利用僵尸网络可以实现在攻击过程中自我学习,自行做出决定,实现自动化、智能化传播,构成新型智能僵尸网络[24]:(1) AI 驱动的恶意软件会通过一系列的自动化决策进行自繁殖,根据被感染系统的参数进行智能调整,假设一个蠕虫版的攻击可以理解目标的环境并选择对应的攻击技术,如果它利用 EternalBlue 漏洞被修复了,那么它可以转向暴力破解 SMB 凭证,加载 Mimikatz 或安装键盘记录器来获取凭证;(2) 恶意软件可以在受感染的系统中安全地学习系统的环境知识,比如受感染设备通信的内部设备、使用的端口和协议以及账户信息;(3) 除了自动化攻击过程中的决策外,攻击者中也不需要 C2 来进行传播和完成目标,攻击会变得更加静默和危险.

例如,在 2017 年 5 月,一种名为 WannaCry[25]的新型勒索软件利用以前已修复的漏洞,以社会工程学作为主要的攻击媒介,像网络蠕虫一样广泛传播,微软反恶意软件使用预测模型立即将其检测到.WannaCry 的传播机制从著名的、已公开的 SMB 漏洞利用而来,已修复的 SMB EternalBlue 漏洞(CVE2017-0145)的公开可用代码,为该常规勒索软件提供了蠕虫般的功能.具体来说,WannaCry 通过至少 3 个阶段的自动化程序实现:(a) 搜索易受攻击的计算机;(b) 利用 SMB 漏洞代码;(c) 安装 WannaCry.

1.1.4 智能 APT 攻击

高级可持续威胁(advanced persistent threat,简称 APT)攻击自身具有极强的隐蔽性和针对性,通过运用受感染的各种介质、社会工程学、供应链等若干手段,实施定向、持久且有效的威胁和攻击.APT 攻击通常包含多种攻击形式,例如鱼叉式网络钓鱼、0day 漏洞、物理欺骗等,整个攻击过程包括收集定向情报、内部横向渗透、构建控制通道、数据收集上传和单点突破攻击等多个步骤[26].

攻击者将 AI 技术应用其中,APT 攻击也逐渐智能化:(1) 利用机器学习分析电子邮件、社交媒体通信资料或者智能设备操作特征,生成模仿用户行为(邮件书写风格、手机操作行为等)的恶意软件,以此对私人信息发起针对性攻击,能够大幅提高钓鱼的成功率;(2) 通过 AI 模型破解用来防护自动化攻击的验证码 CAPTCHA 机制,可以轻松绕过网页安全检测.

2013 年,Vicarious 等人[27]声明破解了 Captcha.com、Yahoo、Google 和 PayPal 基于文本的验证码检测,其攻击的准确率达到了 90%.之后,2016 年黑帽安全技术大会(black hat Conf.)上,Seymour 等人[28]提出一种时间递归神经网络 SNAP_R,该模型可以自主学习如何向特定用户发布钓鱼帖进行鱼叉式钓鱼,通过在社交网络 Twitter 上测试发现,钓鱼帖的点击率达到了有史以来的最高.

1.2 行为预测

行为预测是指对目标对象行为的未来状态或未知状况的预计和推测,具体的讲,根据已发生或出现的历史信息、主观经验和教训作为先验知识,经过特定理论、公式等的推导和分析,对不确定或未知的事件做出定量或定性的表述,同时寻求出行为的事件发生规律,为之后的决策和规划提供支持[29].

预测的目标、对象、期限和内容通常是差异性的,因此形成了多种分类和方法,例如定性、定量以及定性和定量相结合的分类,时间序列、贝叶斯网络(Bayesian network)、灰色理论等方法.目前,主流、应用广泛的预测方法是对目标研究对象产生的相关数据进行收集和整理,利用机器学习、数据挖掘等方法进行分析和学习,构建对应的预测模型,进而基于模型对行为的未来事件进行预测和推理,做出对应的决策措施.

1.2.1 问题域

网络安全具有复杂的体系,内部可以分为执行层、感知层以及任务层和战略层.目前,AI 技术已经在执行层和感知层有广泛的应用,例如在执行层上面实用化,显著提升规则化安全工作的效率,弥补专业人员人手的不足;在感知层面,把原本依赖于人(不可靠)的安全体系标准化,实现大规模的推广,包括人脸识别和图像识别等,但在任务层和战略层开始摸索,仍处于初期的阶段.

然而,智能攻击以自动化方式提升复杂攻击的速度与效率,同时尽可能地将攻击目标的环境融入到攻击中,利用 AI 技术使恶意软件获取对上下文环境的理解,然后通过学习被感染系统的环境信息,调整其攻击行为以融入到目标环境中.例如,智能恶意软件不会去猜测环境中主要使用的是 Windows 系统还是 Linux 系统,也不会去猜测使用 Twitter 或 Instagram 作为 C2 通信信道,而是直接学习和理解目标网络中的通信,并融入到攻击活动或攻击模型中.

因此,事实上,人工智能和机器学习已经被防御者用来寻找非基于签名而是基于行为的恶意软件,即在任务层和战略层利用行为分析和行为预测等基于“行为”的方法进行反制.例如,通过查找异常的用户活动来识别被劫持的账户,并自动发现系统和应用程序的异常流量;通过监控系统和人类行为,以检测潜在的恶意偏离,或基于现有的模式预测新的威胁和恶意软件,提前做出针对性的防御措施等.

通常,对智能攻击行为的检测是反应性的,并只对特定模式或观察到的异常有响应;而预测则主动、先行推断即将发生的恶意事件,在该事件造成任何损害之前对其做出反应,使系统可以做出针对性的防御措施,从而有效提高系统安全性.攻击行为预测方面的研究工作和进展并不像攻击检测那样突出,但是正受到越来越多的关注,该领域的突破将使整个安全领域受益[30].

1.2.2 可行性分析

本节根据智能攻击行为的特性和现有方法的类型,从攻击行为的相关性、时间序列的自相似性和攻击意图目的性这 3 个方面对智能攻击行为预测的可行性进行分析.

(1) 攻击行为的相关性

基于第 1.1 节的表述可以看出:大部分智能攻击均包含多个攻击行为或攻击事件,在目前阶段的网络和系统安全防护机制下,独立单一的攻击行为往往难以成功实施;同时,智能攻击的多个攻击行为和事件之间存在一定的逻辑、时间、顺序等关联性,例如,攻击者一个攻击行为成功实施与否,决定了该攻击下一步行为是否正常进行,甚至未来所有的攻击步骤的发生.因此,作为防御者(预测人员),根据历史报警日志记录的攻击信息进行学习和分析,建立算法模型,可以探索智能攻击的行为相关性以及发生规律,从而了解攻击的未来事件.

(2) 时间序列的自相似性

时间序列是指将同一统计指标的数值按其发生的时间排列而成的数列集合,而对时间序列分析的主要目的是根据已有的历史数据对未来进行预测.由于安全系统或网络得预警日志通常也按时间顺序收集,进而可以生成具有时序的报警序列,利用模型或理论归纳出报警的触发规律及趋势走向,并以此进行类推,可以预测未来攻击的行为事件.

时间序列的可预测性可以通过自相似理论证明,自相似性是指在不同时间尺度上均具有一致统计特性的随机过程,而衡量和验证自相似性的通用技术则是计算 Hurst 指数.孟锦等人[31]采用自相似理论表明了网络安

全态势时序数据的可预测性,他们通过利用小波分析法计算Hurst指数,推导出网络态势时间序列是偏随机游走序列,具有一定的自相似性,并根据可预测性与 Hurst 指数的关系,判断态势时序数据的可预测性,从而证实了报警时间序列的关联性和可预测性,利用网络报警时间序列数据进行攻击预测具备一定的可行性.

(3) 攻击意图目的性

实际的智能攻击攻防对抗的场景中,攻击者在实施攻击之前,例如恶意软件或钓鱼邮件的散播,在设定的总体入侵目标下,通常都会提前计划攻击对象和相应的攻击步骤,每一步均有明确的攻击目标和攻击方式,具备极强的针对性和目的性;同时,随着每一步的成功实施,智能攻击将持续进行,越接近攻击整体目标,其攻击意图越明显,这直接导致其可推测性也越来越高,越来越可靠.

2 智能攻击的行为预测方法概述

目前,对于攻击预测方法的分类有多种,根据预测方法的属性,可以分为定量预测方法、定性预测方法和混合预测方法[32];根据模型类别,可以分为离散模型、连续模型、数据挖掘模型等[9].基于第 1.1 节中智能攻击的定义,面向不同类别的代表性的智能攻击行为,调查了目前主要的研究工作,本文根据预测方法的模式,将主流的方法分为基于神经网络的预测方法、基于博弈论的预测方法、基于攻击图的预测方法、基于数据挖掘的预测方法和其他预测方法.如表 1 所示,其中,“攻击类别”表明了智能攻击的类别和属性,“攻击行为”表明了该智能攻击的具体行为事件,“模型类别”和“方法原理”简单描述了该研究的解决方案.

Table 1 Methods for behavior-prediction of intelligent attacks

表 1 智能攻击行为的预测方法

文献	攻击类别	攻击行为	模型类别	方法原理
Shen[33]	智能僵尸网络	多事件网络入侵	神经网络	原始 LSTM+监督学习
Kishioka[34]	智能僵尸网络	自进化僵尸网络	神经网络	卷积神经网络
Lu[35]	恶意软件逃逸	恶意软件动态行为	神经网络	生成式对抗网络+家庭聚类算法
Ali[36]	智能 APT 攻击	高级鱼叉式网络钓鱼	神经网络	深度神经网络+遗传算法
Bruckner[37]	对抗攻击	混淆恶意数据逃避检测	博弈论	静态预测博弈
Li[38]	智能僵尸网络	复杂网络攻击	博弈论	双人零和静态博弈
Bulò[39]	恶意软件逃逸	随机字节隐藏漏洞	博弈论	随机预测博弈
Haopu[40]	智能 APT 攻击	智能 APT 攻击	博弈论	动态贝叶斯博弈
Yassin[41]	智能僵尸网络	变体特征僵尸网络	图网络	依赖关系图
GhasemiGol[42]	智能僵尸网络	大规模网络入侵	图网络	不确定性感知(uncertainty-aware)攻击图
Nandi[43]	智能僵尸网络	系统漏洞入侵	图网络	攻击图+博弈论
Amer[44]	智能 APT 攻击	恶意软件 API 调用	图网络	马尔可夫链
Hernández[45]	智能 APT 攻击	社交网络钓鱼	SVM	数据挖掘+SVM
Banerjee[46]	智能僵尸网络	智能僵尸网络	数据挖掘	关联规则挖掘
Lim[47]	恶意软件逃逸	恶意代码伪装	数据挖掘	序列相似性比对算法
Vaishnavi[48]	恶意软件逃逸	恶意可执行文件代码伪装	数据挖掘	随机森林,随机树和 Rep 树
Zhi[49]	恶意软件逃逸	恶意软件逃逸	Venn 算法	Venn-Abers 预测器
Mursleen[50]	恶意软件逃逸	变态恶意软件	SVM	SVM 结合水波优化算法
Roseline[51]	恶意软件逃逸	恶意软件逃逸	机器学习	多层随机森林集成技术
Kou[52]	智能 APT 攻击	多事件网络入侵	图网络	事件因果分析,攻击意图识别

基于攻击图的方法以构建有向图的形式描述攻击行为信息,可以清晰地建立攻击行为的逻辑关系和概率分布;基于博弈论的方法更侧重于与攻击者的博弈对抗,对复杂攻击意图的推理和预测;基于神经网络和基于数据挖掘的预测方法均依赖于大量历史数据,但神经网络模型利用数据集训练模型拟合智能攻击行为特征,而数据挖掘直接对数据进行统计分析,分析智能攻击的行为规律.在第 3 节中会对不同种类的预测方法做进一步详细的分析.

2.1 基于神经网络的行为预测

神经网络(neural network,简称 NN)一般指人工神经网络(artificial neural network,简称 ANN),是一种模仿生物神经网络的结构和功能的数学模型或计算模型,用于对函数进行估计或近似[53].神经网络具有函数逼近能

力、自学习能力、复杂分类功能、联想记忆功能、快速优化计算能力以及高度并行分布信息存贮方式带来的强鲁棒性和容错性等优点,在处理非线性问题方面具有无法比拟的优势[54].智能攻击本身基于 AI 技术,例如应用神经网络算法训练深度模型进行攻击,或是针对神经网络的特点绕过安全检测机制等.将基于神经网络的算法应用于智能攻击的行为预测中,不仅可以发挥上述神经网络本身的优势,也可以对智能攻击做出针对性的防御措施.

静态代码分析易受代码混淆技术的影响,在文件执行期间收集的行为数据容易区分,但捕获时间相对较长,恶意有效负载可能在被检测到时已经修改和传递.Rhode 等人[55]根据行为数据的简短快照,使用循环神经网络(recurrent neural network,简称 RNN)预测可执行文件是否为恶意软件的可能性.他们在研究中发现,一组循环神经网络能够以 94%的准确度预测可执行文件在执行的前 5s 内是恶意的还是良性的.这是首次在执行期间对一般类型的恶意文件进行预测,而不是在执行后使用活动日志文件预测,有效增强了网络安全端点保护.

长短期记忆网络(long short-term memory,简称 LSTM)解决了一般的 RNN 存在的长期依赖问题,一般应用在自然语言处理领域,通过结合上下文信息理解语义,预测文本输出等[56],最近几年,研究人员开始尝试将 LSTM 在自然语言处理中的语义模型迁移至安全防护领域.Shen 等人[33]认为:对于网络安全事件的预测,不仅是预测事件是否会发生(即二分类任务),而是去预测在进行攻击时攻击者会采取的具体行为,比如在多步攻击中攻击者会使用的 CVE(common vulnerabilities & exposures),或者在早期的攻击发生时刻就评估攻击的潜在严重性.因此,他们提出一种基于 LSTM 模型的预测系统,按照安全事件发生的时间顺序建立安全事件序列,使用已知的安全事件序列来预测未来将发生的可能事件,并在模型训练过程中加入性能监控模块,设定阈值迭代训练,以提高模型的准确率.

此后,研究人员开始对 LSTM 模型进行更多的尝试.Fang 等人[57]通过利用具有长短期记忆的双向循环神经网络(BRNN-LSTM)开发了深度学习框架,该框架可以容纳数据集展现的复杂现象,包括长期依赖性和高度非线性,实现了更高的预测准确性.为了提高 LSTM 的抗噪能力和序列关联分析能力,Fan 等人[58]提出了一种称为 ALEAP 的方法,将事件嵌入和注意机制纳入 LSTM 模型,并从多源安全设备收集的数百万个安全事件进行测试,证明了该方法在网络安全事件预测中的有效性.

此外,其他神经网络在智能攻击的行为预测中也有较好的表现.Kishioka 等人[34]提出了一种使用卷积神经网络(convolutional neural networks,简称 CNN)的动态预测方法,通过使用主机日志的邻接矩阵作为 CNN 的输入数据,来预测自演化僵尸网络的传播程度.Lu 等人[35]为恶意软件动态行为分类任务构建了基于深度学习的家庭聚类算法和名为 MalDeepNet 的特殊深度神经网络,该模型可以检测通过恶意生成对抗网络(Mal-GAN)实现的未来恶意软件.Hasan 等人[59]采用深度卷积神经网络(deep convolutional neural networks,简称 DCNN)模型和 Ali 等人[36]提出的深度神经网络(deep neural networks,简称 DNN),同时结合遗传算法,在预测智能 APT 攻击中的网络钓鱼也有较好的表现,实现了更高的分类准确性和敏感性.

2.2 基于博弈论的行为预测

基于博弈论的攻击行为预测方法类似于前面讨论的图形模型检查方法,博弈游戏被用作攻击者和防御者之间交互的模型.与图形模型检查方法相反,博弈论方法旨在为“玩家”(参与者)找到最佳策略,而不是通过历史数据观察找到最频繁的攻击和预测未来可能发生的概率[60].因此,博弈论方法对于高级攻击者活动的预测具有较广泛的应用前景.

智能攻击一般具有复杂的攻击意图,和防御者之间存在一定的对抗性.基于博弈论的预测方法通过建立博弈论模型,可以根据攻击“动作”推理攻击意图,做出针对性防御“动作”,使防御者收益达到最大化.例如,在垃圾邮件和恶意软件检测中,攻击者利用随机化来混淆恶意数据并增加其在测试时逃避检测的几率,通常使用随机字符串或字节序列来混淆恶意软件代码以隐藏已知漏洞.对此,Bruckner 等人[37]将对抗性学习形式化为非零和博弈,提出一种游戏理论公式,模拟不同的规避攻击并相应地修改分类功能来学习安全的分类器,称为静态预测博弈.他们假设玩家同时行动,设计了满足该游戏独特 Nash 平衡的条件,并开发了用于学习相应鲁棒分类器——NashSVM 算法.

此外,为了研究复杂网络中的攻击和防御策略,Li 等人[38]提出一种具有完整信息的双人零和静态博弈模型,为研究复杂网络中的攻击和防御策略提供了新的理论框架.该模型同时考虑了攻击和防御策略,他们探索了攻击者-防御者双人零和游戏的纳什均衡,假设攻击者和防御者都有两种典型的策略:目标策略和随机策略,并证明当攻击者的攻击资源没有防御者的资源那么丰富时,模型网络和现实网络中都存在纯策略纳什均衡,防御者优先保护目标,而攻击者则倾向于随机选择目标;而当攻击资源远大于防御资源时,攻击者和防御者都在均衡中采用目标策略.

然而,上述工作中,分类函数和模拟数据操作均以确定性方式建模,而不考虑任何形式的随机化.因此,Bulò 等人[39]通过在玩家的策略中引入随机性来扩展上述方法,在 2016 年提出随机预测博弈(非合作博弈论公式)克服了这一局限,分类器(防御者)和攻击者根据在各自策略集上定义的某种概率分布来进行随机策略选择.该方法甚至可以针对所有不同于设计期间假设的攻击,并改善垃圾邮件和恶意软件的攻击检测与错误警报之间的权衡.此外,该策略还利用随机化以提高针对逃避攻击的学习算法的安全性,因为它可以向攻击者隐藏有关分类器的信息.Haopu 等人[40]提出一种基于动态贝叶斯博弈的预测模型,针对 APT 攻击出色的不可感知性和长期持久性的特点,提出了相应的定量方法来计算行为收益;然后,基于攻击过程建立动态贝叶斯模型,通过以上设计的解决方案计算博弈均衡,可以用来指导 APT 攻击行为的预测.

上述工作中,基于随机博弈的分析方法都采用了完全理性的假设,但在实际的网络安全对抗场景中,攻守双方都很难满足完全理性的高要求.为此,Zhang 等人[61]分析了有限理性对攻防随机游戏的影响,构建了一种随机博弈模型:针对网络节点数量增加时状态爆炸的问题,设计攻防图来压缩状态空间,提取网络状态和防御策略;在此基础上,引入智能学习算法 WoLF-PHC 进行策略学习和改进;然后设计了具有在线学习能力的防御决策算法,该算法有助于从候选策略集中选择收益最大的最优防御策略,所获得的策略优于先前的进化均衡策略,因为它不依赖于先前的数据;最后,通过引入资格跟踪来改善 WoLF-PHC,使学习速度和防御实时性得到显著提高.

2.3 基于攻击图的行为预测

攻击图是 Phillips 和 Swiler 在 1998 年引入的攻击场景的图形表示方法[62],然后迅速成为一种流行的形式化攻击表示方法,第 1 种攻击预测方法就是基于攻击图的方法.同时,攻击图还用作其他图网络模型预测方法的基础,例如贝叶斯网络和马尔可夫模型等方法.

从 DDoS 攻击到恶意软件和垃圾邮件分发等,僵尸网络具有多种攻击形式,同时,大部分破坏都是在对其遏制的过程中完成的,而智能僵尸网络的规避攻击、自主传播进一步带来更大的威胁,提高检测的准确性或速度已力所不及.2016 年,Abaid 等人[63]首先利用攻击图的方法采取了新的方向——在僵尸网络的攻击发生之前,预测即将发生的攻击,向网络管理员提供预警,然后通过隔离主机、切断网络等实现针对性防御.他们将僵尸网络感染序列建模为马尔可夫链,识别可能导致攻击的行为.然而,面对具有变体特征的物联网僵尸网络的攻击,攻击模式通常难以构建,当前的方法无法通过分析注册表信息来识别此类行为.Yassin 等人[41]通过使用依赖关系图组成不同僵尸网络的相似和相异模式,以促进恶意软件变体特征识别的过程,这样既构建了精确的攻击模式,也能够将其视为未来的僵尸网络预测.

攻击图一般仅能够提供有关漏洞利用可能性的静态信息,这对于预测未来并不可靠;此外,攻击图未考虑概率的不确定性.针对上述问题,GhasemiGol 等人[42]定义了一种不确定性感知(uncertainty-aware)攻击图,通过考虑攻击概率的不确定性来评估当前网络安全状态;然后使用不确定性感知攻击图和依赖图信息定义预测攻击图,预测未来可能发生的攻击.该方法可应对攻击发生的不确定性,更精确地预测未来的网络攻击并动态适应环境的变化;同时,在预测过程中使用其他信息,例如入侵警报、主动响应和依赖图等,使得该预测攻击图的大小和复杂性在大规模网络中也能较好地适用.

尽管现有文献提供了丰富的攻击图生成算法,仍需要更多方法用来帮助对攻击图进行分析.Nandi 等人[43]提出一种结合了攻击图和博弈论的双层防御者-攻击者模型,对具有攻击图资源限制的两人和顺序的防御者-攻击者游戏进行建模.此外,考虑安全预算因素,他们在攻击图上找到最优可承受的拦截计划的方法,保护该子集免受攻击,最大程度地减少了由于安全漏洞而造成的损失.

为了通过 API 调用序列预测恶意软件,同时解决为其建立攻击图的工程量和计算复杂度过高的问题,Amer 等人[44]研究了如何生成表征恶意软件的简单行为图.他们首先使用词嵌入来识别恶意软件调用序列中 API 函数之间存在的上下文关系,然后基于恶意软件和良性软件调用序列之间存在的显著区别,对两种 API 调用序列的行为进行马尔可夫链建模,并生成一个描述了 API 函数之间实际关系的语义转换矩阵,实现了从初始 API 调用功能对 API 调用序列是否为恶意代码的预测.

2.4 基于数据挖掘的行为预测

与前面 3 种预测方法相比,数据挖掘对数据深层的隐藏特征和内部模式具有更强的表征能力,但通常作为其过程中一种技术手段.具体来说,基于数据挖掘的预测方法通过对海量攻击警报、检测结果等先验知识进行统计分析、规则关联及分类归纳等,挖掘出攻击信息之间的规律,对未来攻击进行分类和预测,或结合攻击图、博弈论等算法建模预测[64].

数据挖掘在智能攻击的行为预测,特别是对钓鱼网站的预测有较好的表现.网络钓鱼通过模仿合法网站来欺骗在线用户,以窃取其敏感信息.由于反网络钓鱼解决方案旨在准确地预测网站类别,与数据挖掘分类技术的目标完全匹配,因此,网络钓鱼可以看作是数据挖掘中的典型分类问题,分类器由大量网站的特征构成.Mohammad 等人[65]在 2014 年评估了基于规则的数据挖掘分类技术在预测网络钓鱼网站方面的良好表现,并通过实验证明了不同分类技术的可靠性.Al-diabat 等人[66]研究了根据分类性能确定有效特征集的特征选择方法,他们比较了两种已知的特征选择方法,以便确定数据挖掘进行网络钓鱼预测的最少特征集.

智能钓鱼攻击大部分集中于社交网络,攻击者利用机器学习分析电子邮件和媒体资料,向特定用户发布网络钓鱼帖,或者模仿用户编写电子邮件窃取信息等.Hernández 等人[45]提出了一种对 Twitter 内容的情感分析方法,以预测未来对 Web 的攻击.该方法基于每天收集的两组用户的推文,通过数据挖掘进行关联规则和统计分析,然后构建 SVM 模型以预测是否可能发生攻击.Banerjee 等人[46]基于从蜜网和社交网络获取的网络流量数据部署本地蜜网,对于给定的持续时间和位置,探索社交网络并提取事件、活动和新闻等详细信息;从蜜网获得的数据集通过相关性和相似性用于僵尸网络的检测,而关联规则挖掘技术实现对僵尸网络攻击的预测.

相似性度量是基于数据挖掘的预测方法中常用的技术,一般应用于恶意代码比对、系统调用序列比对等[67].针对恶意代码趋于智能化,逃避当前大部分基于签名的检测系统的问题,Lim 等人[47]提出了一种应用相似序列比对算法对恶意代码进行分类的技术.他们通过提取恶意代码的网络行为模式并比较 DNA 序列的相似性,排除了代码的其他特性,从而只对恶意代码的网络行为进行分类,同时还能够预测攻击行为序列.Vaishnavi 等人[48]基于从日志文件包含的信息中提取的属性,针对恶意可执行文件的代码伪装逃逸检测,提出了随机森林、随机树和 Rep 树等 3 种基于数据挖掘的算法,并比较了这些算法的预测准确率、计算复杂度等特性.

推荐系统一般用于预测用户的行为习惯和偏好,通过系统响应用户浏览行为,部署机器学习算法进行个性化计算,发现用户的兴趣点,实现针对性推荐[68].受推荐系统的启发,Zamani 等人[69]构建了类似的推荐系统来预测僵尸网络攻击.他们基于遥测僵尸网络流量收集的数据,定期对推荐系统进行来自网络恶意威胁消除和补救平台系统的威胁数据的训练,特别是攻击的地理位置和时间戳;同时,结合基于 *K*-Means 和 DBSCAN 集群的机器学习算法,根据来自给定地理位置坐标的等级,和针对某些僵尸网络类型的高密度位置的警报,推荐最有可能发生的攻击.

2.5 其他预测方法

除了上述 4 种主流分类的预测方法之外,研究人员也使用其他的方法进行了尝试,例如支持向量机(support vector machine,简称 SVM)、推荐系统、相似性度量等实现智能攻击行为的预测,并取得了一定的成果,本节挑选部分列举如下.

僵尸网络的智能化,使得超过 70%的恶意软件使用多种逃避技术来避免检测,其中,基于静态阈值的僵尸网络检测模型面临着概念漂移的挑战.Zhi 等人[49]引入 Venn-Abers 算法,通过构建 Venn-Abers 预测器,缓解由恶意软件逃逸产生的概念漂移问题.他们认为:即使预测结果正确,预测质量的下降也是概念漂移的标志.因此选择

KNN 和 KDE 作为评分分类器,每个预测都有一个由 Venn-Abers 预测器输出的概率间隔,可以准确地指示预测的质量.

Mursleen 等人[50]提出了一种基于 SVM 和水波优化(water wave optimization,简称 WWO)结合的新方法,用于检测和预测变态恶意软件.WWO 能够在搜索空间内进行运动搜索,寻找到最优解,用于确定 SVM 的参数,提高 SVM 模型性能.同时,使用 ClustalW 和 T-Coffee 两种序列比对软件在主要配对期间进行签名对齐和次要多重对齐,以避免出现代码长度可变的问题.

Roseline 等人[51]认为:新一代可逃逸的恶意软件是普通恶意软件的变体,可以使用基于视觉的方法分析攻击模式并形象化地呈现出来,表征其恶意行为.他们使用多个连续层中的集成森林来交替生成类向量和特征向量,以对恶意软件进行分类,提出了一种比深度学习模型更健壮和有效的混合堆叠多层组合的预测方法.与深度神经网络相比,该方法使用的超参数更少,并由于具有自动设置参数(连续级别数)的自适应特性,计算效率更高,可适用于不同规模的数据集.

Kou 等人[52]分析了现有的网络安全状况评估方法,对攻击意图和网络配置信息之间的关联性进行了深入分析,提出了一种基于攻击意图识别的网络安全态势评估和入侵预测方法.与传统方法不同,该方法基于入侵者意图,首先对攻击事件进行因果分析,发现并简化入侵路径以识别每个攻击阶段,然后根据攻击阶段进行态势评估,最后识别出攻击意图,并根据已达到的攻击阶段,结合漏洞和网络连接预测下一个攻击阶段.该方法能更准确地反映攻击的真实性,而且该方法不需要对历史序列进行训练,因此在入侵预测中更加有效.

3 智能攻击的行为预测方法分析

本节依据第 1 节中提出的智能攻击、行为预测的定义和分类,对第 2 节中概述的代表性工作分别从方法原理、方法特征和适用范围这 3 个方面进行分析,从不同角度对几种主流的智能攻击行为的预测方法进行讨论、对比和总结.

3.1 方法原理分析

3.1.1 基于神经网络的预测方法

基于神经网络的预测是目前最流行的攻击预测方法之一,神经网络预测模型具有良好的拟合性、对目标样本的自学习和自记忆功能,还具有并行处理、高度容错和极强的逼近能力等特性,可以获取复杂非线性数据的特征模式.总体而言,基于神经网络的预测方法以一些攻击行为的输入输出数据作为训练样本,通过神经网络的自学习能力调整权值,构建预测模型;然后运用模型,实现从输入状态到输出状态空间的非线性映射.

通过第 2.2 节中的研究工作概述可以看出,其中,循环神经网络占据了主要地位.循环神经网络具有记忆性、图灵完备(Turing completeness)以及参数共享等特点,因此,在对序列的非线性特征进行学习时具有一定优势;智能攻击行为一般由多个事件构成,具有一定的时序和逻辑关系,利用循环神经网络可以很好地对其建模和预测.

以 LSTM 为例,传统 RNN 存在“梯度消失”的主要问题,由于误差随时间膨胀或衰减而无法访问远程上下文,只能对短时间序列进行建模.为了解决这个问题,LSTM 引入了 3 个门保持状态:接受上一时刻的输出结果、当前时刻的系统状态和当前系统输入,然后通过输入门、遗忘门和输出门更新系统状态,并将最终的结果进行输出.如图 2 所示,3 个门分别为输入门 $\boldsymbol{i}_t$、遗忘门 $\boldsymbol{f}_t$ 和输出门 $\boldsymbol{o}_t$.其中,$\boldsymbol{i}_t$ 和 $\boldsymbol{o}_t$ 控制信息的流入和流出网络,$\boldsymbol{f}_t$ 控制先前序列的影响.具体公式如下:

$$\left\{\begin{array}{l}\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{ci}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i)\\ \boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{cf}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{co}\boldsymbol{c}_{t-1}+\boldsymbol{b}_o)\\ \boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c)\\ \boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh\boldsymbol{c}_t\end{array}\right\} \tag{1}$$

其中,$\boldsymbol{c}_t$ 表示 t 时刻的存储单元,$\boldsymbol{h}_t$ 表示隐藏层的输出,$\boldsymbol{b}$ 表示偏置,$\boldsymbol{W}=\{\boldsymbol{W}_{xi},\boldsymbol{W}_{xo},\boldsymbol{W}_{xf},\boldsymbol{W}_{ci},\boldsymbol{W}_{co},\boldsymbol{W}_{cf},\boldsymbol{W}_{hi},\boldsymbol{W}_{ho},\boldsymbol{W}_{hf}\}$表示加权参数,并且通过反向传播共同学习时间序列.

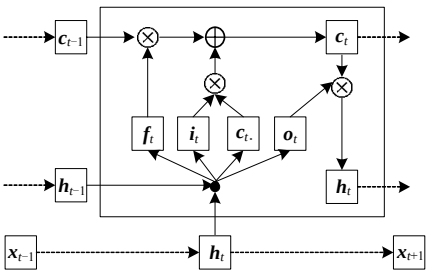

Fig.2 Structure diagram of LSTM[56]

图 2 LSTM 结构图[56]

基于 LSTM 预测智能攻击行为的示例如图 3 所示(“LSTM 记忆单元”展开如图 2 所示).

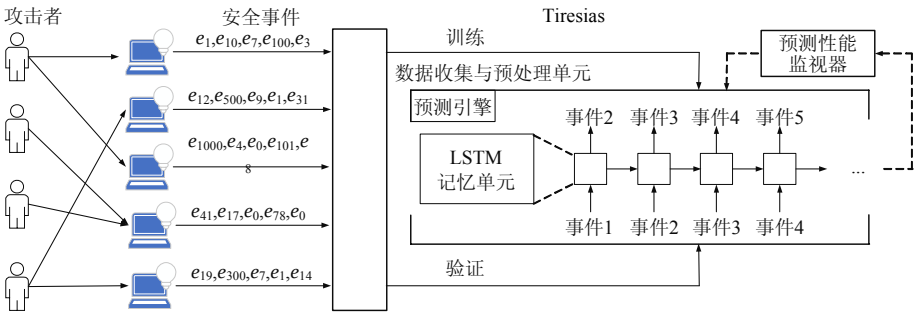

Fig.3 An example of how LSTM model predict intelligent attack behavior[33]

图 3 LSTM 模型预测智能攻击行为示例[33]

Shen 等人提出一种名为 Tiresias 的预测系统,以 27 天的时间内通过商业入侵防御系统 74 万台计算机中收集的 34 亿个安全事件作为数据集,其中,$e\{e_1,e_2,\ldots,e_n\}$代表系统日志记录的所有事件,$e\{e_1,e_2,\ldots,e_{13}\}$为协同攻击事件.Tiresias 的操作包括 4 个阶段:数据收集和预处理、模型训练和验证、安全事件预测、预测性能监视.本质上,Tiresias 通过对隐含层应用仿射变换,然后加上 *softmax*,在给定历史观察事件 $e\{e_1,e_2,\ldots,e_w\}$的情况下,指定 e_{w+1} 个可能事件的概率分布,其中,w 表示回滚窗口大小.

$$P_r(e_{w+1}\mid e_{1:w})=\frac{\exp(h^w\cdot p^j+q^j)}{\sum_{j'\in E}\exp(h^w\cdot p^{j'}+q^{j'})} \tag{2}$$

其中,p^j 是输出嵌入 $P\in R^{m\times|V|}$的第 j 列,q^j 为偏差项.因此,在给定训练数据 D_T的情况下,Tiresias 的训练目标是使所有事件序列的负对数似然度ζ最小:

$$\zeta=-\sum_{t=1}^{|D_T|}P_r(e_t\mid e_{1:t-1}:\theta) \tag{3}$$

模型训练完成后,Tiresias 将历史事件 $e\{e_1,e_2,\ldots,e_i\}$作为初始输入(即与真实世界场景内联的可变长度输入序列),并预测 e_{i+1} 的概率分布.给定 E 为 $P_r[e_{i+1}|e_{0:i}]=\{e_1{:}p_1,e_2{:}p_2,\ldots,e_{|E|}{:}p_{|E|}\}$,对 $P_r[e_{i+1}|e_{0:i}]$进行排序,然后选择概率最高的事件作为预测结果.最后,为了保持预测准确性,预测性能监视器会跟踪并报告 Tiresias 是否预测了正确的事件,如果预测精度下降到某个阈值以下,系统将自动重新训练模型.

3.1.2 基于博弈论的预测方法

博弈论旨在分析对象与经常冲突的对象之间的相互作用,其基本假设是博弈的“玩家”,即参与者是理性的,他们追求各自的目标,并考虑其他玩家的知识或期望而进行战略性推理,为最大化自己的收益而选择最优的行动.博弈游戏是博弈论战略互动的模型,该游戏包括:(1) 有限的N个玩家集合(通常是系统中的攻击者和防御者/管理员);(2) 每个玩家的非空动作序列 $A_i,i\in N$;(3) 每个玩家的收益函数 $u_i,i\in N$,即每个结果给玩家 i 数量为

$a \in \times_{j \in N} A_j$ 的收益.

玩家策略是一种功能,在玩家做出决定时提供不同的动作,分为纯粹策略和混合策略:纯粹策略针对每种情况提供单一的措施,混合策略则为每种情况分配一组玩家行动的概率分布.博弈论中解决方案的概念并不明确,最常用的是纳什均衡[70].在纳什均衡中,两个玩家都选择了同样的策略,但是他们都不偏离策略;通常找到博弈的纳什均衡比较难以计算,但是某些类型的博弈可以使用近似于纳什均衡,计算复杂度较低的算法.

有多种类型的博弈模型可用于攻击预测,根据表现形式,可以分为一般博弈和战略博弈:战略博弈中,每个玩家选择且仅选择一次动作,并且所有玩家的动作同时进行;相反,在一般形式的博弈中,玩家依次或同时选择动作(可能无限多次),并且可以将所有可用信息用于决定中.另外,根据玩家获得的彼此历史动作信息的完善度,可以分为完善信息与不完善信息的博弈模型[71].

早期网络攻击预测框架的博弈理论形式化如图 4 所示,防御者和攻击者拥有共同的知识、各自的私有信息、策略空间及收益,通过正常访问的信息进行重复博弈,预测未来的攻击者意图.智能攻击的攻击策略更加复杂,例如,现有的基于随机博弈的网络攻防分析方法都采用了完全理性的假设,但在实际的网络攻防中,攻守双方都很难满足完全理性的高要求.因此,研究人员以此为基础,提出了不同的博弈模型进行应对:Zhang 等人[61]在分析了有限理性对攻防随机游戏影响的基础上,构建了一个随机博弈模型,针对网络节点数量增加时状态爆炸的问题,设计了博弈攻防图来压缩状态空间,提取网络状态和防御策略.攻防对抗的过程和策略选择如图 5 所示.

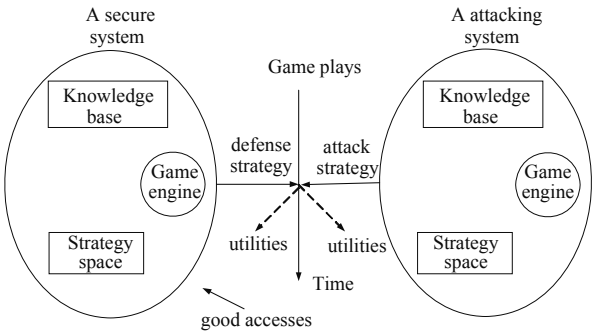

Fig.4 Framework of game theory predicting attack behavior[72]

图 4 博弈论攻击行为预测框架[72]

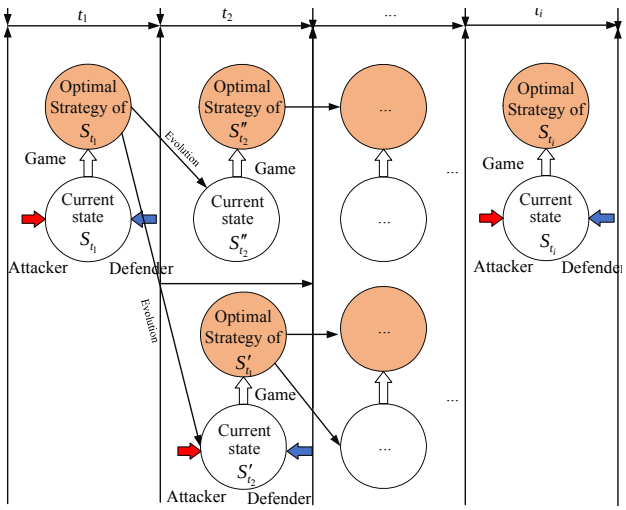

Fig.5 Game process and strategy selection in attack and defense[39]

图 5 攻防中的博弈过程和策略选择[39]

他们使用 Soloris 平台的 Sadmind 漏洞的 DDoS 攻击为例,该攻击通过多个步骤实施,包括 IP 扫描、Sadmind ping、Sadmind 攻击、安装 DDoS 软件和实施 DDoS 攻击,每个攻击步骤都可能导致网络安全状态的改变.第 1 步,初始网络状态表示为 $S_0(H_1,none)$,这意味着攻击者 Alice 没有主机 H_1 的任何特权;然后,攻击者 Alice 通过开放端口 445 对 H_1 进行了 IP 扫描攻击,并获得了 H_1 的用户特权,此网络状态表示为 $S_1(H_1,User)$;此后,如果防御者 Bob 从候选策略集中{重新安装侦听器程序,安装补丁,关闭未使用的端口}中选择并实施了防御策略,则网络状态将被转移回 S_0;否则,网络可能会继续发展到另一个更危险的状态 S_3.

3.1.3 基于攻击图的预测方法

基于攻击图的预测方法是一种经典的网络攻击预测方法,攻击图模型主要利用图网络的结构表示攻击者以及其攻击手段之间的联系,一般以攻击者或者防御者的身份作为实体或者节点,以攻击行为或手段作为图网络的边,表示"实体"之间的逻辑关系,通常表示为元组 $G=(S,r,S_0,S_s)$.其中,S 是一组状态,$r\subseteq S\times S$ 是一个过渡关系,$S_0\subseteq S$ 是一个初始集状态,$S_s\subseteq S$ 是一组成功状态.初始状态表示攻击开始之前的状态,过渡关系表示攻击者可能采取的行动,通常通过攻击者选择动作的概率来对它们进行加权.如果攻击者采取了所有从初始状态过渡到任何成功状态的操作,则攻击成功,成功状态表示系统受到损害.

攻击图一般由手动或自动构建,例如使用数据挖掘来生成攻击图,基础的攻击图示例如图 6 所示.在节点中可以看到构成攻击的可能事件,边缘值表示与端节点关联的事件将发生的概率,边缘值称为可预测性.基于攻击图的预测通过遍历图搜索成功的攻击路径,或基于图中边缘的概率值.假设模型当前处于攻击的特定状态,则将该节点标记为初始状态,从初始状态遍历所有可能的路径,例如使用广度优先搜索,并选择成功导致系统妥协的路径作为可能的攻击路径,权重可用于预测最可能的路径,也可以在每个节点中考虑攻击者的最可能动作.

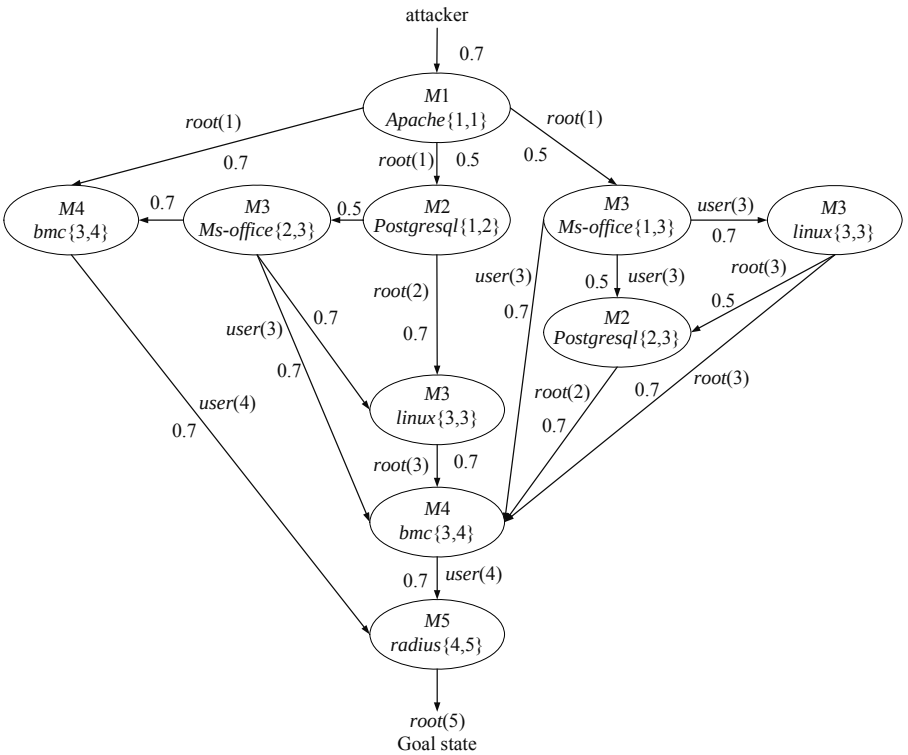

Fig.6 An example of how attack graph predict attack behavior[73]

图 6 攻击图攻击行为预测举例[73]

不同种类的多步攻击需要马尔可夫链来表示每种攻击的状态,通过选择与每个入侵的相似阶段相对应的

HMM 状态数来构建这些链,HMM 模型中的 S 参数由这些状态组成.图 7 显示了分别具有 4 个、3 个和 5 个状态的 3 种多步攻击(入侵 A、入侵 B 和入侵 N)的示例,每个入侵都有不同的 HMM 模型(λ_A,λ_B,λ_C).在该图的右侧,Holgado 等人将所有马尔可夫链表示为攻击图,从管理员界面显示.在该示例中,*Step*-A_1 和 *Step*-B_1、*Step*-B_3 和 *Step*-N_4 以及 *Step*-A_4 和 *Step*-N_5 处于相同状态.

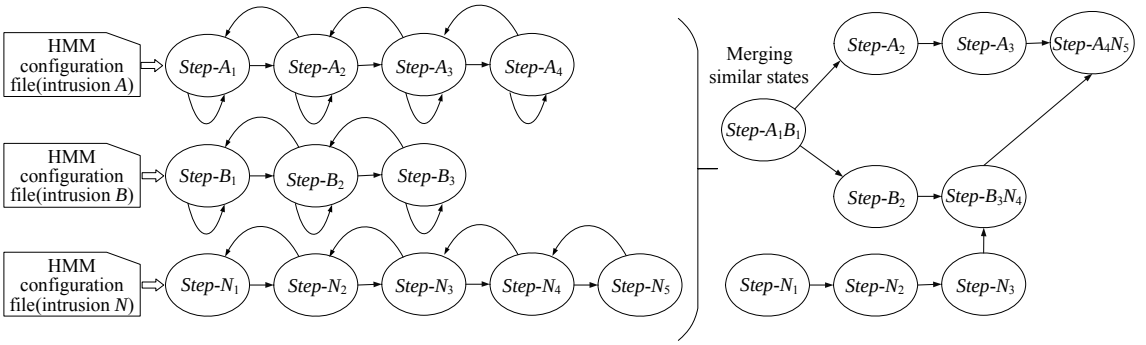

Fig.7 An example of how hidden Markov model predict intelligent attack behavior[74]

图 7 隐马尔可夫模型预测智能攻击行为示例[74]

3.1.4 基于数据挖掘的预测方法

数据挖掘的预测建模分为分类建模和预测建模两种方式,其中,分类主要用于对离散数据进行预测,对未来数据进行分类;预测建模通过构造、使用模型来对某个样本的值进行估计,例如预测某个未知值或者缺失值,主要用于对连续或有序的数据进行预测.

数据挖掘可挖掘攻击场景的警报信息,对先验知识进行统计和关联.Husák 等人[75]提出一种基于信息交换和数据挖掘的方法预测网络入侵行为,他们收集了在 SABU 平台内共享的网络安全警报,其中,每天共享大约 220 000 个来自不同地理位置的传感器(入侵检测系统和蜜罐)的警报,然后使用顺序规则挖掘的方法来识别常见的攻击模式,并导出用于预测攻击的规则.表 2 展示了顺序规则挖掘提取到的排名前 10 的规则,包括输入输出、支持度、置信度、最小和平均时差等.

Table 2 Top 10 rules extracted by sequential rule mining[75]

表 2 顺序规则挖掘提取到的排名前 10 的规则[75]

规则	输入	输出	支持度	置信度	Min.Δt	Avg.Δt
1	Org A.tarpit:Recon.Scanning:2323, Org_A.nemea.hoststats:Recon.Scanning::None	Org_A.tarpit: Recon.Scanning:23	0.004 38	0.883 86	12	1 530
2	Org A.nemea.bruteforce:Attempt.Login:23	Org_A.tarpit: Recon.Scanning:23	0.008 24	0.534 65	121	7 539
3	Org_A.nemea.hoststats:Recon.Scanning::None	Org_A.hoststats: Recon.Scanning:None	0.019 87	0.682 14	1	401
4	Org A.tarpit:Recon.Scanning:2323	Org_A.tarpit: Recon.Scanning:23	0.066 55	0.700 99	901	5 882
5	Org A.tarpit:Recon.Scanning:2222	Org_A.tarpit: Recon.Scanning:22	0.008 34	0.581 55	914	7 041
6	Org A.tarpit:Recon.Scanning:2323, Org A.hoststats:Recon.Scanning:None	Org_A.tarpit: Recon.Scanning:23	0.004 87	0.890 71	21	2 019
7	Org_A.nemea.hoststats:Recon.Scanning::None, Org_B.nemea.hoststats:Recon.Scanning::None	Org_A.hoststats: Recon.Scanning:None	0.005 44	0.800 88	4	735
8	Org_A.hoststats:Recon.Scanning::None, Org A.tarpit:Recon.Scanning:443	Org_A.tarpit: Recon.Scanning:80	0.002 89	0.900 00	35	22 754
9	Org_A.hoststats:Recon.Scanning::None, Org_B.nemea.hoststats:Recon.Scanning::None	Org_A.nemea.hoststats: Recon.Scanning:None	0.004 11	0.602 84	1	2 698
10	Org A.tarpit:Recon.Scanning:2323, Org_A.hoststats:Recon.Scanning::None, Org_A.nemea.hoststats:Recon.Scanning::None	Org_A.tarpit: Recon.Scanning:23	0.002 66	0.839 62	12	1 528

如表 2 所示,大多数规则包含仅来自两个组织的少量入侵检测系统生成的网络扫描警报,这表明两个组织正在运行带有大量分布式入侵检测系统的大型骨干网;与其他类型的警报(例如暴力破解)相比,网络扫描警报的数量大,因此包含网络扫描警报的规则排在前 10 名;由于顺序规则挖掘不随时间变化,因此规则之间的最小和平均时间差取自与规则匹配的事件,表明了可以缓解预测出的攻击的剩余时间.

3.2 方法特征分析

由第 3.1 节中的论述可以看出:不同种类的智能攻击行为预测方法在原理上存在一定的差异,算法的侧重点、属性及优劣势等各方面的表现也不完全相同.因此,本节从实时性、计算复杂度、训练样本规模、假设和先验知识等方面对 4 类主流的预测方法进行对比分析,提取和总结不同方法的特征,比较各种方法之间的算法优势、局限性等,结果见表 3.

Table 3 Comparison of methods for behavior-prediction of intelligent attack

表 3 智能攻击行为的预测方法对比

方法类型	实时性	计算复杂度	训练样本规模	假设&先验知识	算法优势	算法劣势
神经网络	◐	●	●	✘	良好的拟合性和数据表征能力,准确率较高	对数据样本依赖性较强,训练代价较大
博弈论	○	●	◐	✔	考虑收益型战略推理,更深刻地理解复杂攻击意图和多模式攻击行为	参数设置缺乏标准,具有一定的主观性
攻击图	●	○	○	✔	小规模数据表现较好	需要一定的先验知识
数据挖掘	○	◐	●	✘	基于海量样本挖掘,分类和归纳更准确,深层数据特征描述	对数据样本依赖性较强,训练代价较大

注:✔=满足,✘=不满足;○=低,◐ =中,●=高

如表 3 所示,基于神经网络的预测方法以人工神经网络算法作为基础,在对智能攻击事件序列的非线性特征进行学习时具有绝对的优势,并且具有良好的拟合性、对目标样本的自学习和自记忆等特性,可以获取到智能攻击事件中复杂非线性数据的特征模式,代表性的工作有 Tiresias[33]、BRNN-LSTM[57]、ALEAP[58]等.基于神经网络的预测方法基于大规模样本训练,对智能攻击事件之间的逻辑关系、规律的挖掘准确率较高,但对数据样本质量依赖性较强,训练耗时久,代价较高,且容易陷入局部最小点,易出现过度拟合而使得泛化能力较差,网络拓扑结构的确定没有成熟的理论指导,其解不具有稀疏性且难以解释等缺陷.

基于博弈论的预测方法通常针对具有攻防博弈的对抗环景,根据攻击者和防御者掌握对手信息的完整性,而建立不同的博弈游戏模型,代表工作有 NashSVM 算法[37]、双人零和静态博弈[38]、随机预测博弈[39]、基于动态贝叶斯博弈的预测模型[40]等.基于博弈论的方法考虑收益型战略推理,可以更深刻地理解攻击者的意图,包括攻击目标、攻击来源、攻击行为之间的联系等,并描述行为之间的逻辑关系,以此和攻击者进行博弈和对抗,做出更具针对性的决策.

基于攻击图的预测方法以图网络结构构建模型,例如有向攻击图、马尔可夫链、贝叶斯网络图等,代表性的工作有僵尸网络依赖关系图[41]、不确定性感知攻击图[42]以及结合了攻击图和博弈论的双层攻防模型[43]等.此类算法通常以身份作为节点,攻击手段作为图网络的边,表示“实体”之间的不同联系,在小规模的数据场景下表现较好,但需要一定的先验知识作为基础.

与前面 3 种预测方法相比,数据挖掘对数据深层的隐藏特征和内部模式具有较强的表征能力,但通常作为其过程中的一种技术手段,代表性的工作有情感分析方法[45]、相似性序列比对[47,48]以及构建推荐系统[69].基于数据挖掘的预测方法通过对海量攻击警报、检测结果等先验知识进行统计分析、规则关联及分类归纳等,挖掘出攻击信息之间的规律,对未来攻击进行分类和预测;或结合攻击图、博弈论等算法建模预测,对钓鱼网站、社交网络攻击的预测具有良好的表现.

3.3 适用范围分析

通过第 3.1 节对几类主流预测方法的原理的详细阐述,结合第 3.2 节对不同方法的实时性、计算复杂度、

训练样本规模等特征的对比分析,可以看出:各种方法在不同的领域均有其适用性或局限性,也因此适用于不同的应用场景.基于以上表述,本节对基于神经网络、博弈论、攻击图以及数据挖掘这 4 种主流智能攻击行为的预测方法的算法特点和适用范围进行讨论,结果展示见表 4.

Table 4 Application scope of methods for behavior-prediction of intelligent attack

表 4 智能攻击行为的预测方法适用范围

方法类型	算法特点	适用范围
神经网络	基于大规模样本训练,良好的拟合性和数据表征能力	密集型攻击、大规模攻击场景
博弈论	考虑收益的战略性推理,复杂攻击意图识别	复杂意图攻击、多模式攻击
攻击图	小规模数据表现较好,实时性高	小规模数据场景
数据挖掘	海量数据的挖掘、分类和归纳,深层数据特征描述	密集型攻击、大规模攻击场景,如钓鱼网站

如表 4 所示,基于图网络的预测方法对小规模的数据表现较好,但需要一定的先验知识,因此更适用于小规模数据场景;基于神经网络的预测方法基于大量样本训练,准确率较高,但是对数据样本依赖性较强,模型训练耗时久,适用于主机入侵、网络攻击等,大量的威胁情报或者系统日志可以为神经网络模型提供充足的训练数据,有效提高模型精度和鲁棒性.与图形模型检查方法相反,基于博弈论的推理式预测方法旨在为“玩家”(防御者)找到最佳策略,而不是通过历史数据观察和找到最频繁的攻击,此类方法对于高级攻击者活动的预测有较广泛的应用前景.基于数据挖掘的预测方法在对钓鱼网站的预测有较好的表现,网络钓鱼通过模仿合法网站来欺骗在线用户以窃取其敏感信息.由于反网络钓鱼解决方案旨在准确地预测网站类别,与数据挖掘分类技术的目标完全匹配,网络钓鱼可以看作是数据挖掘中的典型分类问题,其中,分类器由大量网站的特征构成.

在实际场景中,用户也可以根据不同智能攻击的特点选择合适的预测方法.例如:网络攻击通常频繁、密集,数据量较大,可以选择基于神经网络的预测方法训练深度模型,能够对未来网络态势有较好的理解和估计;而对物联网中无人车、无人机等攻击,通常具备一定的竞争和对抗性,则可以考虑基于博弈论的预测方法,甚至选择多种预测方法相结合,以期达到最优的效果.

4 研究展望

日益增长和复杂化的智能攻击正逐渐发展为人工智能领域中的隐患,虽然针对智能攻击的预测取得了一定的研究成果,但由于智能攻击随着 AI 技术的发展不断演变和进化,为该领域提出了更多新的挑战.通过前文对当前智能攻击行为预测相关的研究热点和主要工作的分析,本节对该领域的现存问题和未来研究趋势加以阐述.

4.1 复杂攻击场景建模与技术迁移

从表 1 的相关工作统计中可以看出,当前对智能攻击行为预测的研究工作多数集中于网络攻击领域:一方面,网络智能攻击通常具备多事件攻击场景,且多个攻击事件具有一定的逻辑关系,容易实现建模和分析;另一方面,网络攻击易于产生大量完备信息的数据,如威胁情报、主机入侵日志等,为预测模型的训练数据提供支撑.

然而,某些系统攻防的场景下,例如对抗攻击中代表性的基于有限内存的 BFGS(limited-memory broyden-fletcher-goldfarb-shan-no,简称 L-BFGS)、快速梯度符号法(fast gradient sign method,简称 FGSM)等白盒攻击,其原理是通过访问模型的结构和权重,计算真实的模型梯度或近似梯度,根据防御方法和参数调整其攻击方法,但根本性质为一次性的静态攻击;又如物联网智能设备的某些攻击,多次攻击行为之间可能无法具备直接的关联,或是行为特征的提取较为困难,例如已加密的路由器进出站流量、深度数据包等.此外,随着 AI 技术的发展,新的攻击媒介和安全范式不断出现,如何进行建模和预测,以及把网络攻击的预测技术迁移或拓展到上述领域中,将是未来需要重点研究的方向之一.

4.2 可信数据与隐私保护

Yahoo 公司 30 亿账户数据泄密而崩溃,Heartland 支付系统 1.34 亿信用卡账号、Equifax 超过 1.455 亿用户数据被窃取[76],海量的用户数据是构建预测模型的重要支撑,但同时也是智能攻击的主要目标与数据来源.构建

可信、可靠以及隐私保护的数据处理技术体系,是保障机器学习模型安全的基石.可信数据作为可靠的数据来源,可以使防御模型的预测更加精准;反之,利用可信数据与隐私保护也能够减弱智能攻击模型的数据来源的可靠性,从而降低其攻击效率.因此,对社交网络、主机日志等隐私数据的保护也是应对智能攻击的根本途径之一.

目前,数据可信处理与隐私保护的第 1 个挑战是:如何有效增强机器学习模型训练数据的质量,以保证数据的可靠性和安全性[77].在智能攻击中,主要体现在:对于 AI 模型的对抗攻击中,攻击者通常人为制造噪声,改变数据分布,或是生成恶意的对抗样本,即“投毒攻击”,干扰模型的分类过程.未来研究中,需要建立一定的数据评估体系,同时采用重复消除、缺失处理、逻辑错误检测等方式增强数据质量,提高数据可靠性.

另一个挑战是如何突破敏感数据隐私化处理技术,以保证训练数据隐私甚至是训练模型机密性.在智能攻击中,主要体现在智能 APT 攻击、恶意软件逃逸等,攻击者通过对社交网络、主机日志等隐私数据的入侵,进而利用这些数据训练攻击模型,根据用户习惯、系统漏洞发起针对性的攻击.因此,如何增强隐私数据传输安全性,建立完善的数据隐私性保护体系,也将成为智能攻击预测研究发展的重要趋势.

4.3 基于AI技术的扩展和应用

智能攻击基于人工智能技术,而从人工智能技术本身入手,使用 AI 技术来防御 AI 攻击,也是最有效的方式之一.从第 3 节的对比分析中可以看出,Tiresias[33],BRNN-LSTM[57],ALEAP[58]等基于神经网络的预测方法逐渐成为主流的研究趋势.神经网络对非线性时间序列数据具有很强的逼近和拟合能力,并具有自学习能力、中短期预测精度较高以及半监督或无监督等优点.但基于神经网络的预测算法仍存在一定的局限,如普遍的泛化能力弱、易陷入局部极小值等问题.目前,应用广泛、表现较好的仍是对序列的非线性特征学习具有优势的循环神经网络.

此外,数据深层的隐藏特征和内部模式更能代表攻击行为之间的逻辑关联以及攻击者复杂的攻击意图,通过利用数据挖掘和机器学习算法对这些特征的挖掘和分析,分析智能攻击行为规律,能够更加精准和有效地判断攻击者意图,进而预测未来攻击行为.如第 2.4 节所述的 Hernández 等人[45]的情感分析方法、Lim 等人[47]和 Vaishnavi 等人[67]的相似性序列比对、以及 Zamani 等人[69]构建推荐系统来预测智能攻击行为,均取得了显著的成效.

因此,未来的研究中,除了基于应用较为广泛的 SVM,LSTM 等模型进行扩展的同时,可以采用更多的神经网络进行尝试,以及利用多个分类器系统的方式增加规避难度,优化特征选择来制作特征平均分配等;同时,在智能攻击行为的预测中进行数据挖掘和其他机器学习算法的进一步改进和在实践中的利用,也将成为提升预测能力关键方法,例如对网络数据和警报从批处理到流数据处理的过渡、在协作环境(协作入侵检测系统或警报共享平台)中对攻击预测的研究等.

4.4 对抗环境与攻防博弈

智能攻击一般为动态攻击,会随着数据训练不断演化和趋于自动化、智能化.然而,目前很多防御性方法均是基于强假设的被动静态经验性防御,即使是基于 AI 技术的预测模型,也是由历史攻击数据训练而来,不能保证完全覆盖未来发生的智能攻击特征;并且在实际应用场景中,这些假设条件很难满足,通常攻击者和防御者之间的模型内部结构信息均是未知,多数情况下只能执行黑盒测试.

对抗防御与攻防博弈是一种多维度、持续性防御过程,是在网络安全的任务层和战略层的防御策略,从对智能攻击行为的预测,已经体现出主动防御和对抗防御的特点,而依赖静态经验的被动防御已经完全不能适应不断学习、动态演化的智能攻击.因此,在面向动态、对抗的攻防博弈环境下,如何能够预测复杂的攻击者意图,建立动态自适应的防御体系,将成为未来研究的重要方向;基于博弈论或者概率模型推理攻击策略,也将成为未来重点的研究对象.

5 总 结

基于 AI 技术的智能攻击对网络通信、智能设备、社交媒体等不同领域进行入侵,对网络安全、政府办公、

国防军事等造成了极大的威胁.行为预测根据历史报警信息预测未来即将发生的攻击动作,并建立机器的自动感知和自学习机制,能够有效预防智能攻击,提高系统的安全性.本文对面向智能攻击的行为预测方法进行了全面的调查和论述,界定了智能攻击行为预测的问题域,对其相关的研究领域进行了概述;梳理了面向智能攻击的行为预测的研究方法和相关工作,并进行分类和详细介绍;分别阐述了不同种类预测方法的原理机制,从特征及适应范围等角度做进一步对比和分析;展望了智能攻击行为预测的挑战和未来研究方向,为之后对智能攻击行为预测的研究提供了新的思路,对于今后该领域的继续和深入研究具有一定的参考意义.

References:

[1] Acemoglu D, Restrepo P. Artificial Intelligence, Automation and Work. Social Science Electronic Publishing, 2018.

[2] Falco G, Viswanathan A, Caldera C, *et al*. A master attack methodology for an ai-based automated attack planner for smart cities. IEEE Access, 2018,6:48360−48373.

[3] Meng Y, Tu S, Yu J, *et al*. Intelligent attack defense scheme based on DQL algorithm in mobile fog computing. Journal of Visual Communication and Image Representation, 2019,65:Article No.102656.

[4] Ranjbar MH, Kheradmandi M, Pirayesh A. Assigning operating reserves in power systems under imminent intelligent attack threat. IEEE Trans. on Power Systems, 2019,34(4):2768−2777.

[5] Zhang R, Chen X, Wen S, *et al*. Using AI to attack VA: A stealthy spyware against voice assistances in smart phones. IEEE Access, 2019,7:153542−153554.

[6] Nazer TH, Xue G, Ji Y, *et al*. Intelligent disaster response via social media analysis a survey. ACM SIGKDD Explorations Newsletter, 2017,19(1):46−59.

[7] Qiu S, Liu Q, Zhou S, *et al*. Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences, 2019, 9(5):Article No.909.

[8] Xu H, Ma Y, Liu H, *et al*. Adversarial attacks and defenses in images, graphs and text: A review. arXiv preprint arXiv:1909.08072, 2019.

[9] Husák M, Komárková J, Bou-Harb E, *et al*. Survey of attack projection, prediction, and forecasting in cyber security. IEEE Communications Surveys & Tutorials, 2018,21(1):640−660.

[10] Jia YJ, Lu Y, Shen J, *et al*. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In: Proc. of the Int'l Conf. on Learning Representations (ICLR 2020). 2020.

[11] Ren K, Zheng T, Qin Z, *et al*. Adversarial attacks and defenses in deep learning. Engineering, 2020,6(3):346−360.

[12] Li C, Zhou W, Yu K, *et al*. Enhanced secure transmission against intelligent attacks. IEEE Access, 2019,7:53596−53602.

[13] Zhang C, Benz P, Lin C, *et al*. A survey on universal adversarial attack. arXiv preprint arXiv:2103.01498, 2021.

[14] Subasi A, Molah E, Almkallawi F, *et al*. Intelligent phishing website detection using random forest classifier. In: Proc. of the 2017 Int'l Conf. on Electrical and Computing Technologies and Applications (ICECTA). IEEE, 2017. 1−5.

[15] Jiang F, Fu Y, Gupta BB, *et al*. Deep learning based multi-channel intelligent attack detection for data security. IEEE Trans. on Sustainable Computing, 2018,5(2):204−212.

[16] Conti M, Dargahi T, Dehghantanha A. Cyber Threat Intelligence: Challenges and Opportunities. In: Cyber Threat Intelligence. Cham: Springer-Verlag, 2018. 1−6.

[17] Yi P, Wang KD, Huang C, *et al*. Adversarial attacks in artificial intelligence: A survey. Journal of Shanghai Jiaotong University, 2018,52(10):172−180 (in Chinese with English abstract).

[18] Khalid F, Hanif MA, Rehman S, *et al*. FAdeML: Understanding the impact of pre-processing noise filtering on adversarial machine learning. In: Proc. of the 2019 Design, Automation & Test in Europe Conf. & Exhibition (DATE). IEEE, 2019. 902−907.

[19] Johnson J. Artificial intelligence, drone swarming and escalation risks in future warfare. The RUSI Journal, 2020,165(2):26−36.

[20] Chaudhary P, Gupta BB, Gupta S. A framework for preserving the privacy of online users against XSS worms on online social network. Int'l Journal of Information Technology and Web Engineering (IJITWE), 2019,14(1):85−111.

[21] Ren Z, Chen G, Lu W. Malware visualization methods based on deep convolution neural networks. Multimedia Tools and Applications, 2020,79:10975−10993.

[22] Kolosnjaji B, Demontis A, Biggio B, *et al*. Adversarial malware binaries: Evading deep learning for malware detection in executables. In: Proc. of the 2018 26th European Signal Processing Conf. (EUSIPCO). IEEE, 2018. 533−537.

[23] Feily M, Shahrestani A, Ramadass S. A survey of botnet and botnet detection. In: Proc. of the 2009 3rd Int'l Conf. on Emerging Security Information, Systems and Technologies. IEEE, 2009. 268−273.

[24] Danziger M, Henriques MAA. Attacking and defending with intelligent botnets. In: Proc. of the XXXV Brazilian Symp. on Telecommunications and Signal Processing-SBrT. 2017. 457−461.

[25] Mohurle S, Patil M. A brief study of wannacry threat: Ransomware attack 2017. Int'l Journal of Advanced Research in Computer Science, 2017,8(5):1938−1940.

[26] Virvilis N, Gritzalis D. The big four-what we did wrong in advanced persistent threat detection? In: Proc. of the 2013 Int'l Conf. on Availability, Reliability and Security. IEEE, 2013. 248−254.

[27] https://vicariousinc.tumblr.com/post/65316134613/vicarious-ai-passes-first-turing-test-captcha

[28] Seymour J, Tully P. Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. Black Hat USA, 2016,37:1−39.

[29] Anagnostopoulos T, Anagnostopoulos C, Hadjiefthymiades S. Enabling attack behavior prediction in ubiquitous environments. In: Proc. of the Int'l Conf. on Pervasive Services (ICPS 2005). IEEE, 2005. 425−428.

[30] Kou G, Wang S, Tang G. Research on key technologies of network security situational awareness for attack tracking prediction. Chinese Journal of Electronics, 2019,28(1):162−171.

[31] Meng J. Research on key techniques in network security situation assessment and prediction [Ph.D. Thesis]. Nanjing: Nanjing University of Science and Technology, 2012 (in Chinese with English abstract).

[32] Chen C, Yan BP. Network attack forecast algorithm for multi-step attack. Computer Engineering, 2011,37(5):172−174,178 (in Chinese with English abstract).

[33] Shen Y, Mariconti E, Vervier PA, *et al*. Tiresias: Predicting security events through deep learning. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. 2018. 592−605.

[34] Kishioka K, Hongyo K, Kimura T, *et al*. Prediction method of infection spreading with CNN for self-evolving botnets. In: Proc. of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC). IEEE, 2018. 1810−1815.

[35] Lu S, Ying L, Lin W, *et al*. New era of deeplearning-based malware intrusion detection: The malware detection and prediction based on deep learning. arXiv preprint arXiv:1907.08356, 2019.

[36] Ali W, Ahmed AA. Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. IET Information Security, 2019,13(6):659−669.

[37] Bru¨ckner M, Kanzow C, Scheffer T. Static prediction games for adversarial learning problems. Journal of Machine Learning Research, 2012,13(1):2617−2654.

[38] Li Y, Deng Y, Xiao Y, *et al*. Attack and defense strategies in complex networks based on game theory. Journal of Systems Science and Complexity, 2019,32(6):1630−1640.

[39] Bulò SR, Biggio B, Pillai I, *et al*. Randomized prediction games for adversarial machine learning. IEEE Trans. on Neural Networks and Learning Systems, 2016,28(11):2466−2478.

[40] Haopu Y. Method for behavior-prediction of APT attack based on dynamic Bayesian game. In: Proc. of the 2016 IEEE Int'l Conf. on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2016. 177−182.

[41] Yassin W, Abdullah R, Abdollah MF, *et al*. An IoT botnet prediction model using frequency based dependency graph: Proof-of-concept. In: Proc. of the 2019 7th Int'l Conf. on Information Technology: IoT and Smart City. 2019. 344−352.

[42] GhasemiGol M, Ghaemi-Bafghi A, Takabi H. A comprehensive approach for network attack forecasting. Computers & Security, 2016,58:83−105.

[43] Nandi AK, Medal HR, Vadlamani S. Interdicting attack graphs to protect organizations from cyber attacks: A bi-level defender-attacker model. Computers & Operations Research, 2016,75:118−131.

[44] Amer E, Zelinka I. A dynamic Windows malware detection and prediction method based on contextual understanding of API call sequence. Computers & Security, 2020,92:Article No.101760.

[45] Hernández A, Sanchez V, Sánchez G, *et al*. Security attack prediction based on user sentiment analysis of Twitter data. In: Proc. of the 2016 IEEE Int'l Conf. on Industrial Technology (ICIT). IEEE, 2016. 610−617.

[46] Banerjee M, Agarwal B, Samantaray SD. An integrated approach for botnet detection and prediction using honeynet and socialnet data. In: Proc. of the Int'l Conf. on Intelligent Computing and Smart Communication 2019. Singapore: Springer-Verlag, 2020. 423−431.

[47] Lim H, Kim W, Noh H, *et al*. Research on malware classification with network activity for classification and attack prediction of attack groups. Journal of the Korean Institute of Communication Sciences, 2017,42(1):193−204.

[48] Vaishnavi N, Thiyagarajan K. A study on prediction of malicious program using classification based approches. Int'l Journal of Computer Science and Mobile Computing, 2018,7(5):38−46.

[49] Wang Z, Gao HZ, Zhang YM, *et al*. Fortifying botnet classification based on venn-abers prediction. In: Proc. of the 2017 2nd Int'l Conf. on Computer Science and Technology (CST 2017). 2017.

[50] Mursleen M, Bist AS, Kishore J. A support vector machine water wave optimization algorithm based prediction model for metamorphic malware detection. Int'l Journal of Recent Technology and Engineering, 2019,7:1−8.

[51] Roseline SA, Sasisri AD, Geetha S, *et al*. Towards efficient malware detection and classification using multilayered random forest ensemble technique. In: Proc. of the 2019 Int'l Carnahan Conf. on Security Technology (ICCST). IEEE, 2019. 1−6.

[52] Kou G, Wang S, Tang G. Research on key technologies of network security situational awareness for attack tracking prediction. Chinese Journal of Electronics, 2019,28(1):162−171.

[53] Hopfield JJ. Artificial neural networks. IEEE Circuits and Devices Magazine, 1988,4(5):3−10.

[54] Khashei M, Bijari M. An artificial neural network (p,d,q) model for timeseries forecasting. Expert Systems with applications, 2010,37(1):479−489.

[55] Rhode M, Burnap P, Jones K. Early-stage malware prediction using recurrent neural networks. Computers & Security, 2018,77: 578−594.

[56] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Proc. of the Interspeech. 2012. 601−608.

[57] Fang X, Xu M, Xu S, *et al*. A deep learning framework for predicting cyber attacks rates. EURASIP Journal on Information Security, 2019,2019(1):Article No.5. [doi: 10.1186/s13635-019-0090-6]

[58] Fan S, Wu S, Wang Z, *et al*. ALEAP: Attention-based LSTM with event embedding for attack projection. In: Proc. of the 2019 IEEE 38th Int'l Performance Computing and Communications Conf. (IPCCC). IEEE, 2019. 1−8.

[59] Hasan KMZ, Hasan MZ, Zahan N. Automated prediction of phishing websites using deep convolutional neural network. In: Proc. of the 2019 Int'l Conf. on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). IEEE, 2019. 1−4.

[60] Roughgarden T. Algorithmic game theory. Communications of the ACM, 2010,53(7):78−86.

[61] Zhang Y, Liu J. Optimal decision-making approach for cyber security defense using game theory and intelligent learning. Security and Communication Networks, 2019,2019(2):1−16.

[62] Phillips C, Swiler LP. A graph-based system for network-vulnerability analysis. In: Proc. of the '98 Workshop on New Security Paradigms (NSPW'98). New York: Association for Computing Machinery, 1998. 71−79.

[63] Abaid Z, Sarkar D, Kaafar MA, *et al*. The early bird gets the botnet: A Markov chain based early warning system for botnet attacks. In: Proc. of the 2016 IEEE 41st Conf. on Local Computer Networks (LCN). IEEE, 2016. 61−68.

[64] Han J, Kamber M, Pei J. Data mining: Concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 2011,5(4):83−124.

[65] Mohammad RM, Thabtah F, McCluskey L. Intelligent rule-based phishing websites classification. IET Information Security, 2014, 8(3):153−160.

[66] Al-diabat M. Detection and prediction of phishing websites using classification mining techniques. Int'l Journal of Computer Applications, 2016,147(5):5−11.

[67] Lin YS, Jiang JY, Lee SJ. A similarity measure for text classification and clustering. IEEE Trans. on Knowledge and Data Engineering, 2013,26(7):1575−1590.

[68] Moreno MN, Segrera S, López VF, *et al*. Web mining based framework for solving usual problems in recommender systems. A case study for movies' recommendation. Neurocomputing, 2016,176:72−80.

[69] Zamani NA, Ariffin AFM, Abdullah SNHS. Recommender system based on empirical study of geolocated clustering and prediction services for botnets cyber-intelligence in Malaysia. Int'l Journal of Advanced Computer Science and Applications, 2018,9(12): 473−478.

[70] Nash JF. Non-cooperative games. Annals of Mathematics, 1951,54:286−295.

[71] Roy S, Ellis C, Shiva S, *et al*. A survey of game theory as applied to network security. In: Proc. of the 2010 43rd Hawaii Int'l Conf. on System Sciences. IEEE, 2010. 1−10.

[72] Liu P. A game theoretic approach to cyber attack prediction. Technology Report, No.DOE/ER/25527, University Park: Pennsylvania State University, 2005.

[73] Hu H, Zhang H, Liu Y, *et al*. Quantitative method for network security situation based on attack prediction. Security and Communication Networks, 2017,2017:1−19.

[74] Holgado P, Villagrá VA, Vazquez L. Real-time multistep attack prediction based on hidden Markov models. IEEE Trans. on Dependable and Secure Computing, 2017,17(1):134−147. [doi: 10.1109/TDSC.2017.2751478]

[75] Husák M, Kašpar J. Towards predicting cyber attacks using information exchange and data mining. In: Proc. of the 2018 14th Int'l Wireless Communications & Mobile Computing Conf. (IWCMC). IEEE, 2018. 536−541.

[76] Alneyadi S, Sithirasenan E, Muthukkumarasamy V. A survey on data leakage prevention systems. Journal of Network and Computer Applications, 2016,62:137−152.

[77] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021,32(1):41−67 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

附中文参考文献:

[17] 易平,王科迪,黄程,等.人工智能对抗攻击研究综述.上海交通大学学报,2018,52(10):172−180.

[31] 孟锦.网络安全态势评估与预测关键技术研究[博士学位论文].南京:南京理工大学,2012.

[32] 陈灿,阎保平.针对复合攻击的网络攻击预测算法.计算机工程,2011,37(5):172−174,178.

[77] 纪守领,杜天宇,李进锋,沈超,李博.机器学习模型安全与隐私研究综述.软件学报,2021,32(1):41−67. http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

马钰锡(1995－),男,博士生,主要研究领域为人工智能,强化学习,存储保护.

张全新(1974－),男,博士,副教授,主要研究领域为人工智能,深度学习,信息安全.

谭毓安(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为深度学习及对抗,物联网与嵌入式系统,数据存储安全,Android 安全.

沈蒙(1988－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为数据安全与隐私保护,人工智能安全.