基于AI的智能化渗透测试技术研究

Research on Intelligent Penetration Testing Technology Based on AI

张小梅[1]，郑　涛[1]，李长连[2]，刘　兵[3]，熊　琛[4]，王昭顺[4]（1. 中国联合网络通信集团有限公司，北京 100033；2. 中讯邮电咨询设计院有限公司，北京 100048；3. 北京墨云科技有限公司，北京 100094；4. 北京科技大学，北京 100083）

Zhang Xiaomei[1]，Zheng Tao[1]，Li Changlian[2]，Liu Bing[3]，Xiong Chen[4]，Wang Zhaoshun[4]（1. China United Network Communications Group Co.，Ltd.，Beijing 100033，China；2. China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China；3. Beijing Moyun Technology Co.，Ltd.，Beijing 100094，China；4. University of Science and Technology Beijing，Beijing 100083，China）

摘　要：

传统的渗透测试方式依赖测试人员的经验，而自动化测试通常基于已知的攻击模式和漏洞库，因此在面对复杂的网络场景时，难以实施灵活高效的渗透测试。针对上述问题，利用人工智能技术赋能自动化渗透测试，提出了基于强化认知决策的智能化渗透测试方案，通过拆解渗透攻击的各个阶段并提取攻击单元，设计迭代运行的系统架构，动态生成攻击行为，针对复杂的网络环境，利用强化学习实现攻击决策智能体的自进化学习，实现高效的智能化渗透测试。

关键词：

渗透测试；强化认知决策；攻击决策智能体；自进化学习；智能化渗透测试

doi：10.12045/j.issn.1007-3043.2024.08.001

文章编号：1007-3043（2024）08-0001-07

中图分类号：TP181

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

Traditional penetration testing relies on the expertise of engineers，while automatic testing based on known attack patterns and vulnerability databases lacks the flexibility and efficiency to address complex network scenarios. To address these challenges，it proposes an intelligent penetration testing approach empowered by artificial intelligence techniques，based on reinforcement cognition decision-making. By decomposing the penetration attack into various stages and extracting attack units，an iterative system architecture is designed to dynamically generate attack behaviors. To tackle complex network environments，a reinforcement learning-based approach is employed to enable self-evolution capabilities of the attack decision-making agent，achieving efficient intelligent penetration testing.

Keywords：

Penetration test；Reinforcement cognition and decision-making；Attack decision-making agent；Self-evolution learning；Intelligent penetration testing

引用格式：张小梅，郑涛，李长连，等. 基于AI的智能化渗透测试技术研究[J]. 邮电设计技术，2024（8）：1-7.

1 概述

渗透测试（Penetration Testing）是一种通过模拟恶意攻击者的行为来评估网络、系统或应用程序安全性的方法。随着信息技术的快速发展和网络攻击手段的日益复杂化，传统的渗透测试方法已经难以满足日益增长的安全需求。传统的渗透测试主要依赖于人工操作，要求渗透测试人员具备丰富的经验和专业知识，他们通过手动的方式进行信息收集、漏洞分析、攻击尝试等操作。这些方法虽然在一定程度上能够发现安全漏洞，但存在效率低下、容易受到测试人员技术水平限制等问题[1]。因此，智能化渗透测试技术应运而生，旨在提高渗透测试的效率和准确性[2]。

2 智能化渗透测试

智能化渗透测试技术利用人工智能、机器学习等先进的技术手段，实现渗透测试过程的自动化。这些先进的技术手段主要包括以下几个方面[3]。

a）自动化扫描工具：通过自动化工具进行大规模的漏洞扫描，快速识别潜在的安全风险。

收稿日期：2024-07-12

b）机器学习算法：利用机器学习算法对大量的安全数据进行分析，自动识别异常行为和潜在威胁。

c）智能决策系统：通过构建智能决策模型，模拟渗透测试专家的思维过程，自动制定测试策略和攻击路径。

d）自适应测试技术：根据测试对象的变化情况和测试过程中的反馈，动态调整测试策略和方法，以提高测试的针对性和有效性。

智能化渗透测试是一种结合了自动化技术和人工智能技术的网络安全评估方法，旨在模拟黑客的攻击行为，以发现并修复网络信息系统中可能被恶意利用的安全漏洞[2]。与传统的依赖人工操作的渗透测试相比，智能化渗透测试通过技术手段降低了对专业人员的依赖，提高了测试的效率和可扩展性[3-4]。智能化渗透测试的流程如图1所示，包括资产识别探测、漏洞检测利用、后渗透和系统智能决策等多个阶段，其中，资产识别探测阶段负责找到目标资产攻击面，为漏洞检测利用提供入口点；漏洞检测利用阶段负责检测及挖掘目标资产存在的漏洞并实施攻击利用，以定位目标脆弱点并为后渗透阶段做准备；后渗透阶段负责数据获取、横向移动及持久化驻留等，用以扩大智能化渗透战果；系统智能决策则围绕整个渗透流程进行攻击行为的决策，有效提升渗透的广度和深度。

2.1 智能化资产识别

通常，攻击渗透的目标内部资产数量庞大、种类繁多、应用场景以及应用方式复杂多样，传统的资产识别方式基于规则匹配，识别速度慢，覆盖种类不全面，无法准确、高效地识别资产。因此，需要研制一种智能化的资产识别方式，以提升资产识别的准确率及效率[5]。

2.1.1 资产数据降噪技术

建立精确稳定的资产识别模型需要对攻击过程

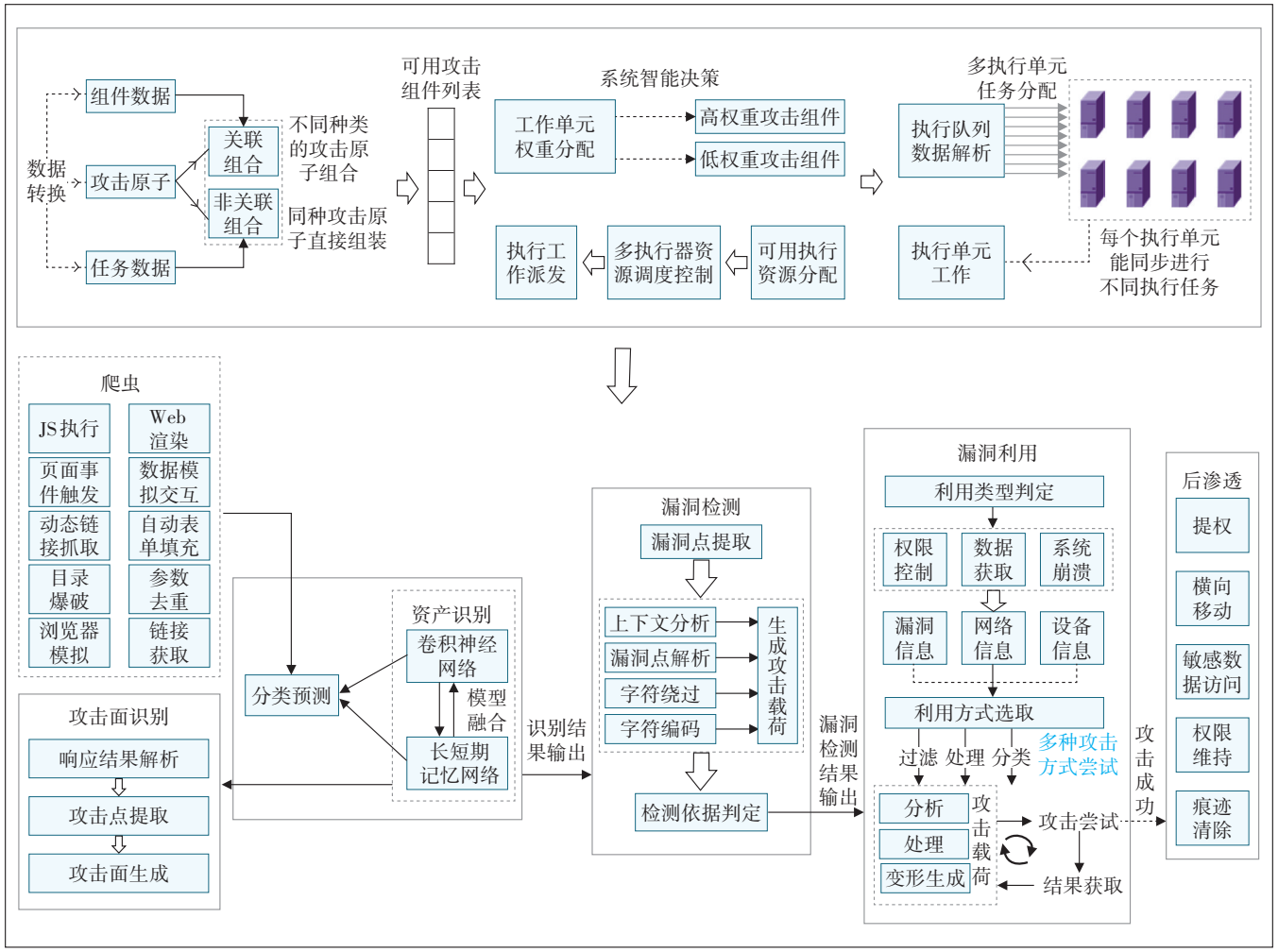

图1　智能化渗透测试流程

中产生的攻击数据进行有效的概括和表示，然而这些数据具有数据量大、内容嘈杂、格式多样、维度过高、且噪声分布广等特点，难以被有效地概括和表示。

堆叠去噪自编码器（Stacked Denoising Autoencoders，SDAE）模型是一种重要的深度学习方法，已经广泛应用于图像分类、行为识别、自然语言处理等领域。SDAE模型是由单层的去噪自编码（Denoising AutoEncoder，DAE）堆叠而成[6]，可对资产数据进行深层次的去噪和特征提取。SDAE的核心思想是利用多个DAE模块的级联效应，以递进的方式提炼出数据中的关键信息，同时抑制噪声的影响[7]，SDAE网络结构如图2所示。

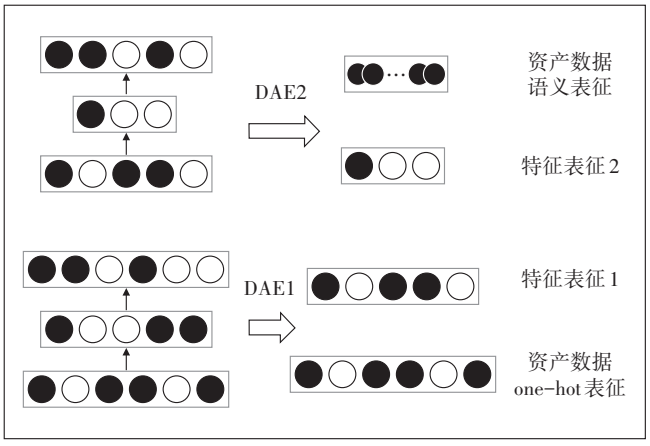

图2　SDAE网络结构

SDAE模型的层叠结构使得网络能够逐渐学习到数据的更高层次的特征，在每一层中，DAE都会对输入数据添加噪声，然后尝试重构出无噪声的数据，这种层叠的方式不仅增强了模型对噪声的鲁棒性，而且还有助于捕捉数据中的复杂模式和结构。在资产数据去噪的应用场景中，SDAE模型能够有效地处理大规模、高维度且含有噪声的数据。通过逐层去噪和特征提取，SDAE能够为资产指纹识别提供更加准确和鲁棒的数据表示。这对于提高网络安全防御系统的资产识别能力具有重要意义，尤其是面对复杂网络攻击和大量异构数据场景。

2.1.2 资产数据分类识别技术

渗透攻击的目标网络一般存在资产数量庞大、种类繁杂、应用场景以及应用方式复杂的问题。传统的资产识别方法存在很大局限性，它需要获取大量目标响应数据并覆盖大量匹配规则去判定目标资产类型，这导致其匹配效率低下且覆盖率低。深度学习在文本分类问题方面有着良好的应用，利用深度学习技术能够提升资产识别的效率及准确度[8]。

资产数据往往具备语义时序性以及特征分布不均匀的特点，长短时记忆-卷积神经网络（Long short-term memory&Convolutional neural networks，LSTM-CNN）模型能够输入降噪后的资产数据，快速输出其对应的类型。当目标资产特征点发生改变时，传统规则难以全面覆盖，而LSTM-CNN模型可以提取目标相应数据中的特征点并对数据上下文进行解析，以实现对资产的判定[9]，具体网络结构如图3所示。

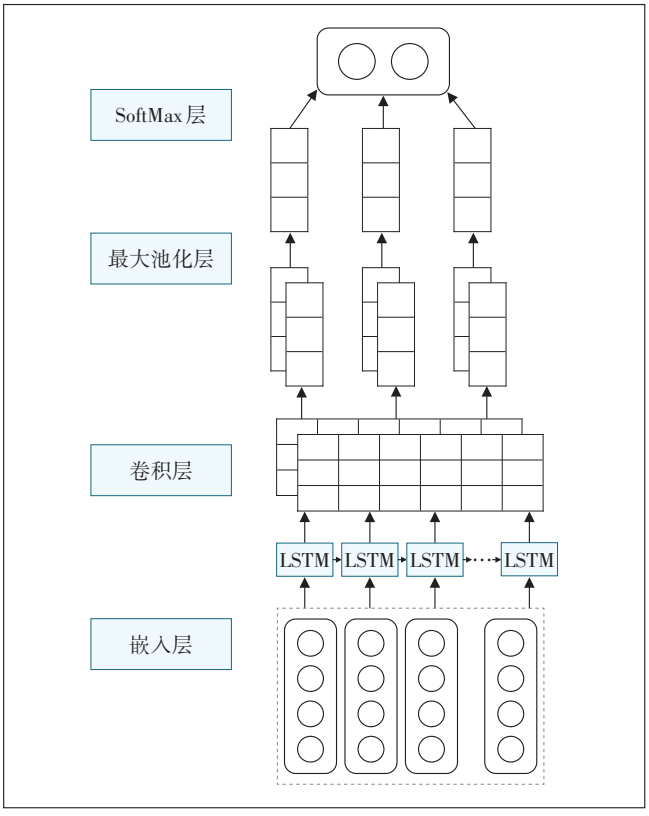

图3　LSTM-CNN模型结构

CNN模型使用卷积滤波器来捕获相邻元素之间的本地依赖关系，受滤波器长度的限制，CNN模型很难了解整个句子的整体依赖关系，因此引入LSTM网络。在LSTM模型中，LSTM构造的特征向量携带了整个句子的整体依赖关系[9]，在给定输入条件下，将文本序列映射成初始特征向量，将这种向量表示输入LSTM网络，并将LSTM网络的输出作为CNN网络的输入，进一步提取序列的特征向量，最后通过Softmax层，输出模型预测的资源类别。

2.2 智能化攻击载荷生成

漏洞检测的能力与检测时对目标发送的攻击载荷有着直接关系。在通用型漏洞检测中，采用单一攻击载荷与固化攻击载荷往往难以覆盖广泛的检测场景，检测能力低下。因此，如何根据目标场景动态生成攻击载荷，提升检测的覆盖度，是漏洞检测的关键点[10]。

攻击载荷通常由多个部分组成，每个部分需要根据对方的漏洞输入点及防护程度，执行标签闭合、逻辑通过、字符转义、混淆绕过等操作。因此，单个攻击载荷的每个部分都存在一些可用攻击代码来选取填充，以此构造整个渗透载荷。在进行攻击检测时，利用机器学习的分类算法，统计当前目标响应下攻击成功概率最高的攻击代码，从而智能化地生成攻击载荷，其整体逻辑如图4所示。

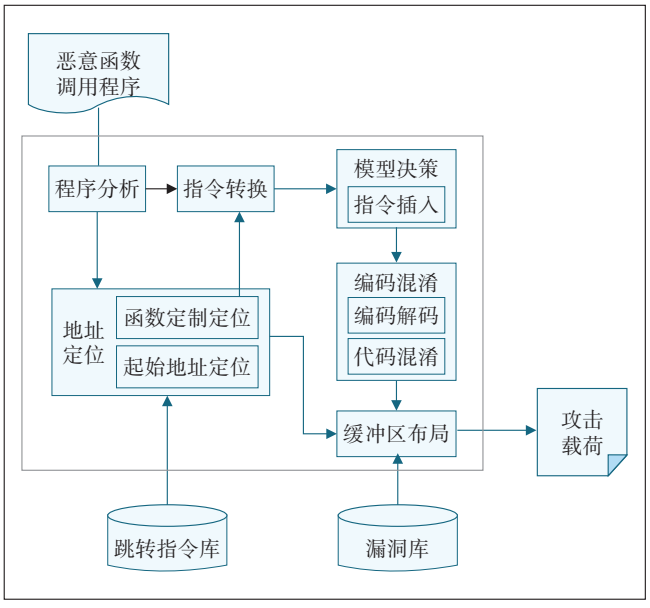

图4　智能化攻击载荷生成逻辑

2.3 攻击路径动态规划

在复杂网络环境中，由于组网方式及目标脆弱程度的多样性，穷举式的攻击方式在工程上不可行，而固定规则式的攻击行为的攻击次数多，攻击时间长，造成攻击行为易暴露。因此，需要尽可能减少攻击尝试，规划合理的攻击路径，提高攻击行为命中率，缩短攻击时间。

如图5所示，网络渗透攻击可分为目标探测/挖掘、攻击面拓展、漏洞检测、攻击利用等4个主要阶段。在每个阶段内都存在多种攻击方法，通过对这些攻击方法进行分析，可以将其中的共性部分拆分出来，形成比攻击方法粒度更细的攻击基础单元，即攻击原子[11]。

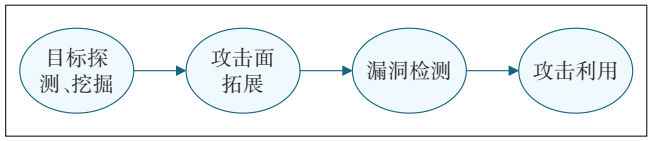

图5　网络渗透主要阶段

本文研究并实现一套用于描述攻击原子的策略语法，该策略语法将各种攻击原子进行关联和组合，形成多种多样的攻击方式和深度不限的攻击路径。攻击原子策略组合如图6所示。

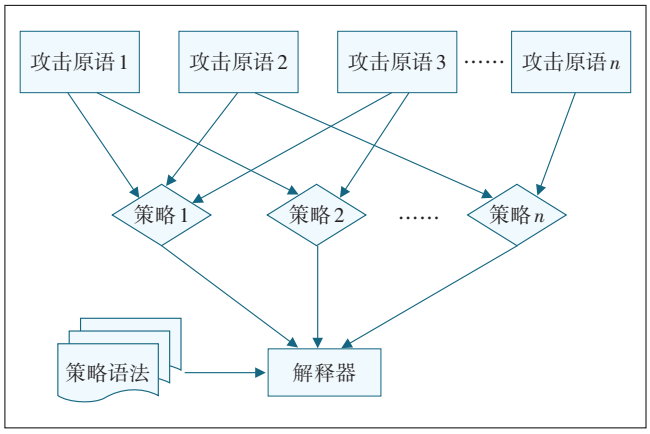

图6　攻击原子策略组合

统筹漏洞检测技术以及新兴的一些自动化渗透攻击技术，都是将常见的漏洞检测流程或攻击过程固化成模块，设定目标后，一旦对目标展开检测和攻击的过程，在模块的生命周期内，就会按预先固化的流程执行，因此，缺乏对实时状况的反馈处理能力和攻击路径的动态规划能力。

在网络渗透攻击过程中，将每一个攻击原子所产生的实际状态和输出数据推送给智能决策系统进行评估，智能决策系统做出反馈，并指导下一步的攻击行为。下一步攻击行为被执行后，再次将产出的新的输出数据交给决策系统进行评估，如此循环反复，最终实现攻击过程的动态判定与规划，绘制出一条当前攻击状态下的最优攻击路径[12]。

攻击路径规划采用动态规划算法作为路径规划基础，将目标当前的攻击状态作为动态规划过程中的系统状态，将智能决策结果作为状态转移策略，将攻击组件集作为攻击范围，将关键目标情报的获取作为攻击边界条件。这样，即使只进行少量攻击行为即可

规划最优攻击路径。同时,随着攻击过程的不断推进,关键节点的产出会探测出新资产以及新知识,从而提升了攻击的广度和深度。

2.4 攻击过程智能决策

在渗透攻击过程中,如何针对目标环境,从海量的攻击方法中选取有效的攻击方法,是智能化攻击决策的关键,这就需要基于安全专家的经验常识及渗透经验数据集形成策略模型。在非规划的情况下,该策略模型能够预测生成成功率高的攻击路径和攻击行为,并根据每次攻击成功与否及产出,形成价值判断,从而优化并指导下一步的攻击行为。

智能化网络攻击的重要一环是自动化渗透攻击工具能够智能化地根据网络信息状态进行自主决策,判定需要执行的攻击行为,并实施“由浅入深”的攻击方法。因此,智能决策是系统内网攻击重要的一环。

如图7所示,攻击决策模型应用在自动化渗透攻击系统中的攻击行为选择阶段,在每次攻击迭代过程中,当需要攻击决策模型做出决策时,决策模型会获取当前状态下目标的漏洞结果信息、攻击利用结果信息以及资产结果信息,并根据这些数据统筹决策来判断当前环境下可用攻击行为的成功概率。

2.4.1 面向攻击事件的统筹决策技术

在内网攻击过程中,生成攻击行为之后,如果使用全量化的攻击手段,往往会导致攻击暴露风险大、对目标造成极强的损伤,常规的解决办法是利用规则匹配的方式,只允许规则覆盖之内的攻击组件在其定义的环境下运行。然而,在面对新的攻击目标时,规则匹配往往无法覆盖,导致无法进一步完成攻击。此外,内网环境复杂,无法利用有限的规则去覆盖多种内网场景,导致攻击方法匮乏,因此,需要采用智能化的攻击决策技术来应对各种攻击场景。

为了能够针对复杂的网络环境做出最优的攻击决策,需要对当前的网络环境进行攻击单元的挖掘,将攻击事件拆解成最细粒度的执行单元,如图8所示。

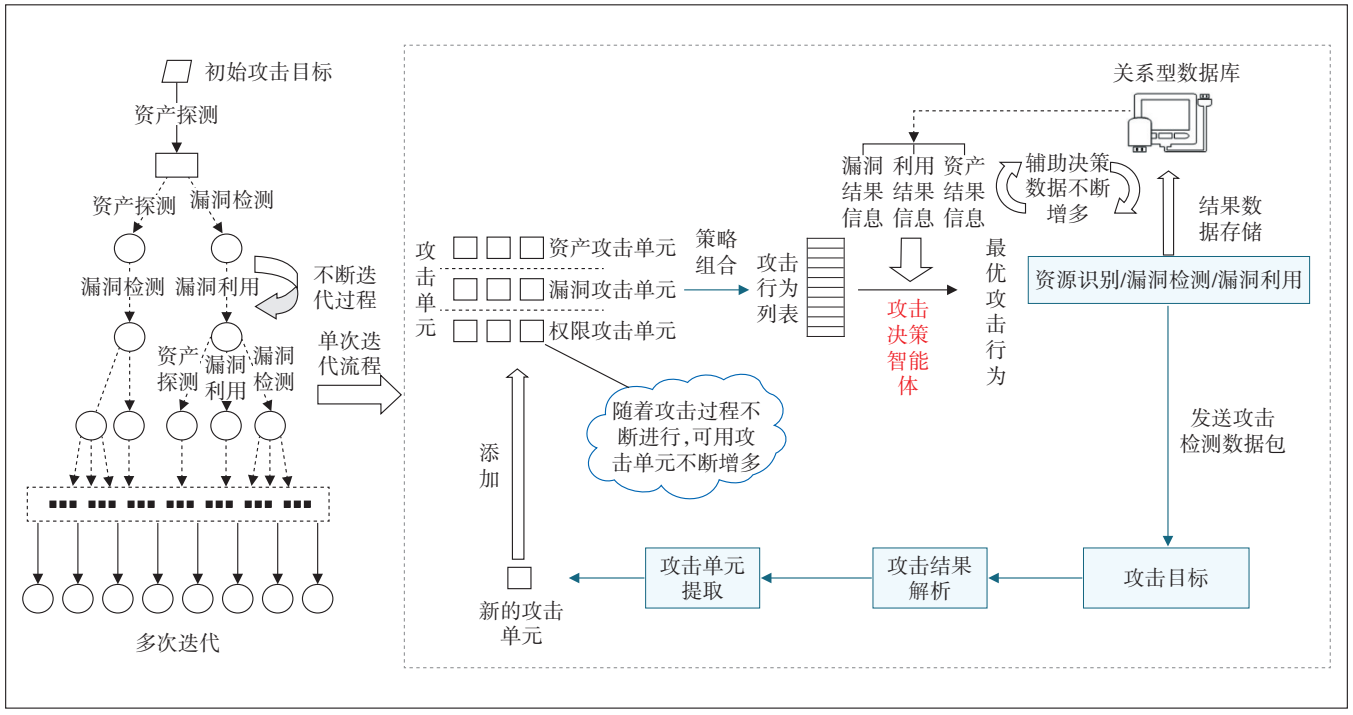

图7　攻击过程智能决策

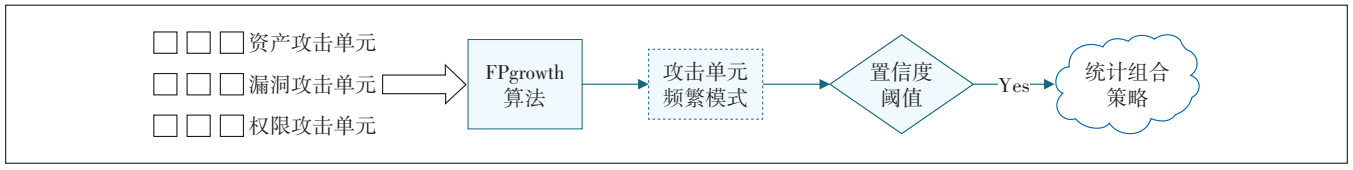

图8　攻击单元模式挖掘流程

通过对大量数据的模式挖掘，能够获取众多攻击单元的组合方法。在自动化渗透攻击系统中，随着攻击的不断深入和资产的不断清晰，能够动态地组合大量的攻击单元。在单次迭代攻击过程中，首先获取上次迭代攻击动态产出的结果数据，从数据中提取攻击单元，并将现有攻击单元与新产生的攻击单元结合，动态生成新的攻击行为。然后，通过图9所示的攻击决策模型对组合生成的大量攻击行为进行优选并执行[13]，执行成功后，将会获取新的攻击结果数据，进入下次迭代攻击。

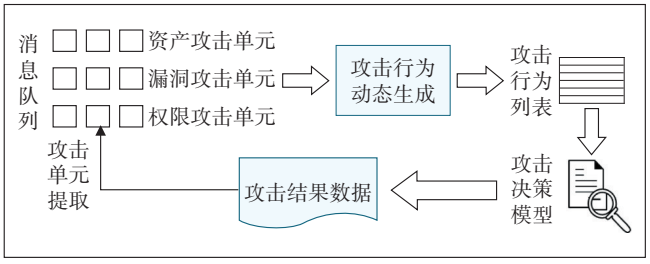

图9　迭代攻击流程

将网络安全攻击事件抽离为攻击原子后，以攻击原子为数据建模基础，以攻击事件为模型数据集，构建一张深度卷积神经网络作为攻击决策网络，攻击决策网络能够将决策结果映射为攻击组件的概率分布。在实际攻击过程中，通过输入可用的攻击行为与当前执行环节的状态结果，攻击决策网络能够根据深度神经网络的权重矩阵生成新的攻击概率矩阵，以指导下一步的攻击。

如图10所示，攻击决策模型采用DCN模型作为架构基础，将目标当前状态下的漏洞结果信息、利用结果信息、资产指纹信息、上次迭代攻击行为及本次攻击行为转化为向量数据并进行拼接组合后输入至DCN模型，模型接收这些拼接向量，并输出该攻击行为的攻击概率。对于攻击概率大于阈值的攻击行为，

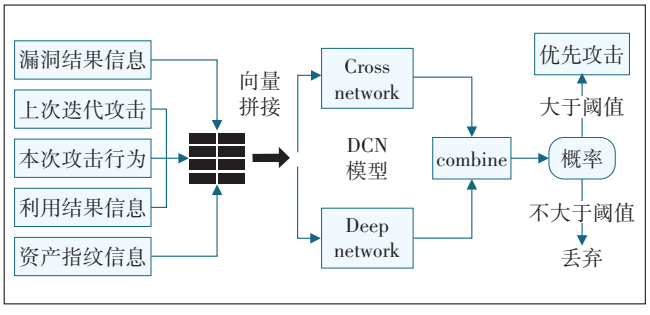

图10　攻击决策模型

则会在价值评估后进行优先攻击；对于攻击概率低于阈值的攻击行为，认定该攻击无法成功，则舍弃该攻击任务[14]。

2.4.2　价值度量技术

攻击决策网络生成的概率分布往往无法直接利用，而是需要在其基础上进行价值判定，抽取其中价值最高的攻击组件。回归模型是一种预测建模技术的方法，它能够体现因变量和自变量之间的关系，常常被用于寻找变量之间的因果关系。因此，笔者设计一种基于回归模型的攻击价值度量网络与一套价值计算公式。通过将价值计算公式与攻击事件样本相结合，计算出历史攻击事件中每个执行环节的价值评分，以此作为训练数据，训练攻击价值度量网络。攻击价值度量网络能够对当前的执行组件进行进一步的估值评分，获取最高分为最优攻击行为。

如图11所示，基于攻击价值度量网络的价值度量技术，内网攻击能够摒除低价值攻击行为，生成最优的攻击组件，引导控制系统向最优方向进行攻击选择，结合攻击决策网络，共同实现智能化的攻击决策。

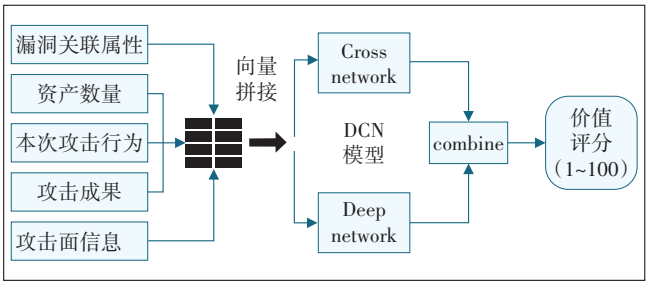

图11　攻击价值评估模型

2.4.3　攻击智能体迭代进化

一旦自动化攻击渗透智能体构建完成，如果其不能根据不同的攻击环境不断学习，那么随着网络环境的进一步改变与发展，其攻击效能将会持续下降。因此，需要设计攻击智能体的自学习方法，使其能够通过各种环境不断攻击尝试的结果进行迭代进化，不断增强智能体的攻击能力。

强化学习目前在机器自我学习进化方向上有很多结合应用，它包含4个主要元素：智能体代理、环境状态、行动及奖励。攻击渗透智能体在学习历史经验数据后，能够在一定环境下做出攻击过程的智能决策。然而，随着目标攻击环境的不断变化、攻击方法的不断丰富，攻击智能体也需要具备根据不同的环境进行自我学习进化的能力[15]。智能体学习反馈如图

12所示。

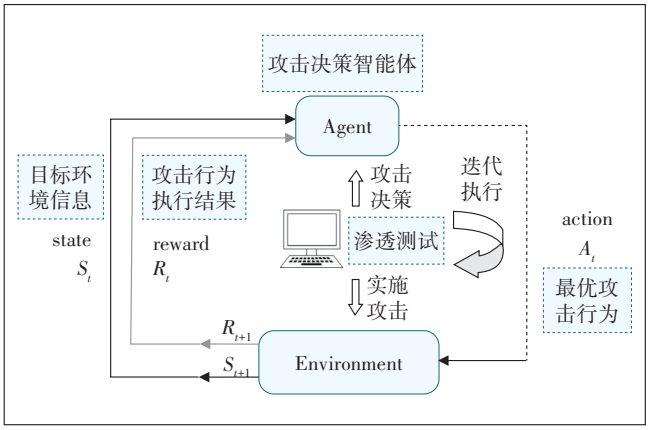

图12　智能体学习反馈

网络靶场通常用来仿真真实网络环境，攻击渗透智能体能够根据不断地攻击网络靶场以及不同的真实攻击环境来提升自身的能力。在这个过程中，以目标环境的系统资产信息、漏洞信息、风险信息作为强化学习智能体的环境状态信息，以当前攻击系统中的可用攻击行为作为行动信息，以对目标是否攻击成功作为智能体的奖励信息。每当攻击成功时，智能体接收到正向奖励，并更新当前状态下的智能体的攻击决策网络。通过不断地变化环境，迭代式地实施攻击尝试，实现攻击智能体的进化。

3 智能化渗透测试的总结和展望

传统渗透测试需要渗透测试人员具备各方面的专业知识、熟悉漏洞机理、熟练运用各种安全测试工具。因此，要摆脱渗透测试对于人工的依赖，需要不断推进智能化渗透测试技术的发展。

随着人工智能技术的发展，将会有更多更成熟的人工智能算法应用到渗透测试的各个阶段中。基于机器学习和深度学习的指纹识别，智能识别测试目标的端口服务、中间件、主机操作系统等指纹信息，从而有效提高渗透的效率。在渗透攻击阶段，通过知识推理，根据目标的网络环境，智能化选择攻击目标，优先攻击具备高渗透价值的目标，智能化选择最合适的攻击载荷，减少渗透尝试的次数，提高渗透测试的效率。针对整个渗透测试过程，通过优先级调度算法对多线程渗透任务的各个线程进行智能网络资源分配，进一步提高渗透效率。未来随着人工智能技术的发展，渗透测试的成功率、自动化程度将会变得更高。

参考文献：

[1] 严俊龙. 基于Metasploit框架自动化渗透测试研究[J]. 信息网络安全，2013(2)：53-56.
[2] 李匀. 网络渗透测试：保护网络安全的技术、工具和过程[M]. 北京：电子工业出版社，2007.
[3] 武杰. 智能化网络渗透测试系统的设计与研究[D]. 长春：长春工业大学，2018.
[4] 黄承彬. Web安全渗透测试[J]. 网络安全技术与应用，2018(7)：21-22.
[5] 陈健，钱星桥. 网络资产识别与安全风险探测软件开发[J]. 技术与市场，2021，28(3)：87-88，91.
[6] 孙艳青，潘广贞，王凤. 结合SDAE网络和ODB学习策略的多目标视觉跟踪[J]. 小型微型计算机系统，2018，39(1)：189-192.
[7] CHEN Y B，LIU Y C，JIANG D S，et al. SdAE：self-distillated masked autoencoder[C]//Computer Vision – ECCV 2022. Cham：Springer，2022：108-124.
[8] 孙澄，胡浩，杨英杰，等. 基于网络防御知识图谱的0day攻击路径预测方法[J]. 网络与信息安全学报，2022，8(1)：151-166.
[9] XIA K，HUANG J G，WANG H Y. LSTM-CNN architecture for human activity recognition[J]. IEEE Access，2020(8)：56855-56866.
[10] BRUMLEY D，POOSANKAM P，SONG D，et al. Automatic patch-based exploit generation is possible：techniques and implications[C]//2008 IEEE Symposium on Security and Privacy (sp 2008). Piscataway：IEEE，2008：143-157.
[11] LIU X G. A network attack path prediction method using attack graph[J/OL]. [2024-01-29]. https://doi. org/10.1007/s12652-020-02206-5.
[12] RADEMAKER T J，BENGIO E，FRANÇOIS P. Attack and defense in cellular decision-making：lessons from machine learning[J]. Physical Review X，2019，9(3)：031012.
[13] BORGELT C. An implementation of the FP-growth algorithm[C]// Proceedings of the 1st international workshop on open source data mining：frequent pattern mining implementations. New York：Association for Computing Machinery，2005：1-5.
[14] GAO K，SU J P，JIANG Z B，et al. Dual-branch combination network (DCN)：towards accurate diagnosis and lesion segmentation of COVID-19 using CT images[J]. Medical image analysis，2021(67)：101836.
[15] 杨飞，周晗，曹京卫，等. 自动化渗透测试技术思考与展望[J]. 邮电设计技术，2022(9)：5-8.

作者简介：
张小梅，硕士，主要从事网络安全技术研究及管理工作；郑涛，硕士，主要从事网络安全技术研究及管理工作；李长连，硕士，主要从事网络安全技术研究及管理工作；刘兵，硕士，主要从事网络安全技术研究及管理工作；熊琛，硕士，主要从事网络智能运维技术研发及管理工作；王昭顺，博士生导师，博士，主要研究方向为信息安全、软件工程、计算机系统结构。