

引文格式:孙守道,卢毅,吴迪.基于改进深度强化学习的电力物联网自动渗透测试技术研究[J].太原理工大学学报,2026,57(2):243-252.

SUN Shoudao, LU Yi, WU Di. Research on automatic penetration testing technology for power internet of things based on improved deep reinforcement learning[J]. Journal of Taiyuan University of Technology, 2026, 57(2): 243-252.

基于改进深度强化学习的电力物联网自动渗透测试技术研究

孙守道^{1,2}, 卢毅¹, 吴迪¹

(1. 国网辽宁省电力有限公司沈阳供电公司, 辽宁 沈阳; 2. 华中科技大学人工智能与自动化学院, 湖北 武汉)

摘 要:【目的】随着泛在电力物联网的全面应用,一些不可控的因素和物理接触环境的变化使得电力物联网系统面临严峻的信息安全问题。渗透测试作为信息安全防护的重要手段,可以提前发现并修复安全漏洞,有效降低系统被入侵的风险。目前的渗透测试以人工测试为主,对人员技术、经验水平要求高,并且测试效率低。为此,提出一种改进深度强化学习的自动渗透测试方法。【方法】首先依据专家经验建立学习过程的状态空间,并引入注意力机制解决状态空间动态变化问题;然后基于深度强化学习模型进行攻击渗透路径自动探索。【结果】搭建实验仿真环境开展对比测试,结果表明所提出方法在不同网络规模下均表现出更快的收敛性能。

关键词:泛在电力物联网;渗透测试;深度强化学习;注意力机制;先验知识

中图分类号:TP393 **文献标识码:**A

DOI:10.16355/j.tyut.1007-9432.20250088

文章编号:1007-9432(2026)02-0243-10

Research on Automatic Penetration Testing Technology for Power Internet of Things Based on Improved Deep Reinforcement Learning

SUN Shoudao^{1,2}, LU Yi¹, WU Di¹

(1. Shenyang Power Supply Company of State Grid Liaoning Electric Power Co., Ltd., Shenyang, Liaoning, China; 2. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, China)

Abstract: 【Purposes】 With the comprehensive application of the ubiquitous power internet of things, uncontrollable factors and changes in the physical contact environment have made power internet of things systems face severe information security issues. Penetration testing, as an important means of information security protection, can discover and fix security vulnerabilities in advance, effectively reducing the risk of system intrusion. Current penetration testing mainly relies on manual testing, which requires high technical expertise and experience from personnel and has low testing efficiency. To address these, an automatic penetration testing method based on improved deep reinforcement learning is proposed. 【Methods】 First, the state space was established for the learning processes according to expert experience and an attention mechanism was introduced to solve the problem of dynamic changes in the state space. Then, a deep reinforcement learning model was used for automatic exploration of attack penetration paths. Finally, an experimental simulation environment was set up

收稿日期:2025-02-18;修回日期:2025-05-13

基金项目:国网辽宁省电力有限公司管理科技项目资助(2024YF-87)

通信作者:孙守道,博士生,高级工程师,主要从事电力系统网络安全及配电自动化技术等研究,(E-mail)leooh@163.com

for comparative testing. **【Results】**The results show that the proposed method exhibits improved convergence performance across different networks.

Keywords: ubiquitous power internet of things; penetration testing; deep reinforcement learning; attention mechanism; prior knowledge

随着大数据、人工智能和能源互联网等新一代信息技术的快速发展,电力物联网系统规模已进入爆发增长的阶段^[1]。在此背景下,2016年我国首次提出“国家泛在智能电网”概念,随后在2018年国家电网在信息通信工作会议上进一步明确要“打造全业务泛在电力物联网”,计划于2024年全面建成泛在电力物联网。作为能源互联网的核心载体,电力物联网通过终端感知设备、通信网络与云平台的深度融合,实现了电力系统“源-网-荷-储”全环节的智能协同^[2]。虽然电力物联网的高扩展性、高智能性和高效性极大地提高了电力系统的运行效率和智能水平,但系统中隐蔽性强的安全漏洞一旦被恶意攻击者利用,将引发严重的隐私数据泄露和系统破坏事件,不仅威胁电力物联网的安全稳定运行,更可能造成难以估量的经济损失^[3]。

近年来,因系统漏洞导致的数据泄露事件频发,例如2015年乌克兰电力系统遭到名为BlackEnergy的病毒攻击,造成当地多处停电事故;2017年中东地区数家能源公司采用的施耐德Triconex仪表控制系统被名为“TRITON”的恶意软件攻击,造成能源供应中断。对于电力物联网而言,攻击者可以通过勒索软件、拒绝服务、网络钓鱼、恶意软件、零日漏洞等方法破坏系统正常运行,进而导致停机、设备损坏、触发应急响应等安全事故。渗透测试是网络安全防护的重要手段,帮助发现并修复安全漏洞,保护信息和业务免受黑客攻击和数据泄露风险^[4]。由于漏洞的隐蔽性、依赖性等特点,渗透测试人员一方面需要对设备异构、协议多样、拓扑动态变化的电力物联网系统有所了解,还需要根据漏洞特征,并结合各种测试工具进行高效渗透。因此,渗透测试需要由开发经验丰富、技术知识全面的专家来负责。然而目前电力部门在渗透测试人员储备方面存在较大缺口,技能水平也参差不齐,难以满足电力物理网场景下的渗透测试需求。尤其是电力物联网的复杂性给人工发现隐蔽脆弱点、分析电力物联网攻击面、设计渗透测试方案带来了很大困难。为此,迫切需要研究自动渗透测试的辅助工具来降低对渗透测试人员技能水平的要求。

早期的自动渗透测试采用基于规则的方法,抽象出特定攻击策略所需的一系列操作,然后集成到渗透工具中,常用的工具包括Nmap和Nessus^{[5][6]}。根据规则的类型,自动渗透测试可以分为单规则和多规则两大类。单规则方法侧重于针对特定的渗透战术进行动作步骤组合,以帮助渗透测试人员自动化完成渗透任务,如DCShadow和Atbroker^[7]等,但该类方法严重依赖于参数配置,无法适应复杂场景下多种战术组合利用的需求。多规则方法通过集成多种攻击策略,可完成渗透测试各阶段的不同任务需求,如信息收集、漏洞利用等,提高了测试过程的自动化程度,代表性工具包括Kaboom、Metasploit、Cobalt Strike等^[8]。然而,该类方法中多种攻击策略之间的协调性相对较低,渗过过程仍然依赖于规则配置。

随着人工智能技术的快速发展,利用人工智能模拟黑客行为已成为当前研究的主流方法,特别是基于模型的自动渗透测试方法。这些方法可以根据目标环境自动选择渗透策略和战术,同时使用机器学习或深度学习算法不断学习和调整渗透测试过程。该方法的主要组成部分包括:1)构建渗透知识库,包括漏洞数据库,用于识别CVE、NVD等系统弱点;用于执行渗透测试任务的战术知识库,如ATT&CK和攻击知识图^[9];2)理解渗透场景,将目标环境的物理信息(IP地址、端口、服务等)转换为数字信息,满足机器理解渗透测试对象的要求^[10];3)规划渗透路径,根据当前目标环境选择有效的攻击策略,以发现渗透路径^[11];4)执行渗透计划,即根据环境选择不同的策略,在渗透过程中使用深度学习等模型进行自主决策,实现对目标系统的渗透测试。另外,为了理解渗透场景,攻击图通常被用于分析目标环境中的漏洞与由此产生的潜在威胁之间的关系^[12]。而对于渗透过程中的自主决策,主要采用基于强化学习的方法。目前基于强化学习框架,开发出了DeepExploit^[13]、AutoPentest-DRL^[14]、OpenVAS^[15]等工具实现自动渗透测试。

综上所述,传统基于规则的方法过度依赖测试人员的专业知识,并且仍需要花费高昂的劳动成本,而智能化渗透测试是解决这些问题的一个有效

途径^[16]。虽然针对自动渗透测试已开展了许多研究,特别是基于强化学习的方法,但相关领域研究仍处于快速发展阶段,还需要进一步深入研究相关问题:1)学习过程中状态空间爆炸和奖励稀疏等问题。随着目标系统复杂性增加,基于强化学习的渗透测试方法也会遇到瓶颈,动作空间和状态空间的复杂性导致策略学习存在困难,并且存在低学习效率导致大量无效的探索问题等;2)由于缺乏针对电力物联网系统的安全信息的有效积累和整合,目前针对电力物联网的智能渗透测试工具相对较少。因此如何将漏洞数据库、威胁情报数据库等安全知识融入渗透测试工具中需要进一步解决。

为此,本文提出一种基于改进深度强化学习(Deep Q-Learning Network, DQN)的自动渗透测试方法。首先依据电力物联网系统拓扑、资产、组件、漏洞等相关先验知识,构建了描述渗透测试过程系统状态的先验知识模型,形成系统状态空间矩阵;在DQN中引入注意力机制,探索面向电力物联网系统的渗透测试的最优路径决策方法。最后以配网自动化系统作为原型,搭建仿真实验靶场,对提出方法进行了测试验证与对比分析。本文主要创新点总结如下:

1)将系统先验知识融入DQN训练过程,减少了智能体的迭代次数,提升了渗透测试的执行效率。具体而言,在传统的DQN训练过程中,智能体需要探索整个状态空间,在复杂环境中可能非常庞大且耗时。通过构建先验知识状态矩阵表征主机节点之间的连接关系,以压缩状态空间,降低无效路径探索次数;另外,在传统贪婪策略中引入了专家经验动作,使得智能体在面临未知状态时能够以一定概率选择专家经验动作,有助于避免产生大量错误决策,加快收敛速度。

2)采用注意力机制解决神经网络输入状态空间变化问题,使得智能体能够在系统状态未知的情况下执行渗透测试任务。具体而言,将智能体获取的环境状态信息通过编码器编码成对应的特征表示,然后引入注意力机制将每个状态的特征表示为键值对,利用Softmax函数计算各个特征的权重因子,并通过计算特征的加权和得到供DQN训练的固定维度的状态矩阵输入。因此智能体在面对未知环境时,不会因获取到新信息(主机节点、主机资产、漏洞等)导致输入状态维度变化而无法匹配神经网络输入维度问题。

1 预备知识

强化学习是机器学习的一种变体,它允许智能体在特定场景中通过不断感知环境状态与动作奖励来调整自身行为,从而找到该场景中的最优行动策略。对于一个标准的强化学习框架,智能体首先在每个观察时间 t 从环境中获取当前状态 $s_t \in \mathcal{S}$,并执行对应的动作 $a_t \in \mathcal{A}(s_t)$,然后接收来自环境的奖励信号 $r_t \in \mathcal{R}$,以及下一个转移状态 s_{t+1} 。通过循环迭代上述过程,智能体逐步学习以找到最大化累计奖励的最优策略 π^* :

$$\pi^* = \arg \max_{\pi \in \Pi} E_{\tau \sim \mathcal{D}_\pi} \left[\sum_{t=0}^T \gamma^t r_t \right]$$

其中 $\gamma \in [0, 1]$ 表示确定长期奖励重要性的折扣因子, Π 表示策略集合, τ 表示轨迹 $(s_0, a_0, s_1, a_1, \dots, s_T)$, $\mathcal{D}_\pi$ 表示策略集合 π 下 τ 的分布。Q-Learning算法是一种常用于寻找最优策略的技术。对于Q-learning算法,Q-function $Q(s_t, a_t)$ 表示智能体从状态 s_t 进行动作 a_t 的期望回报,其更新方程为:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \epsilon \left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right).$$

其中 $\epsilon \in [0, 1]$ 表示学习速率。然而传统的Q-Learning算法在训练过程中使用表格存储 s_t 和 a_t ,这种方法在实际中存在较大的限制,不能解决大状态空间问题。为此,研究者们引入了神经网络最优状态逼近器生成状态-动作集合,称之为深度强化学习算法DQN^[17]。DQN算法训练过程如图1所示,智能体从环境中获取当前系统状态,并根据评估网络计算出在当前状态下采取各个动作的Q值,选择使 $Q(s, a; \theta)$ 最大的动作 a_t ;在动作执行后环境的状态转移到一个新状态 s_{t+1} ,并得到当前动作奖励 r_t ;每次交互产生的状态-动作集 (s, a, r, s') 存储到经验回放空间中。在智能体训练过程中,从经验回放空间中随机抽取一批样本 (s, a) 进行学习;评估网络和目标网络由多层神经网络组成,评估网络用于计算当前策略下的Q值 $Q(s, a; \theta)$,而目标网络用于计算目标Q值 $\max_a Q(s', a; \theta')$,其中 θ 和 θ' 分别表示两个网络的参数;然后通过比较评估网络预测的Q值和目标网络计算的目标Q值,可以计算损失函数 $\mathcal{L}$,衡量评估网络预测的Q值与目标Q值之间的差异;根据损失函数,使用反向传播算法调整评估网络的参数 θ ,使其预测的Q值更接近目标Q值;当评估网络

的参数更新后,经过一定的迭代训练次数,将其更新至目标网络,即 $\theta \rightarrow \theta'$ 。整个过程持续迭代,智能体通过与环境的交互不断学习和优化其行为策略,最终达到最大化累积奖励的目的。DQN 损失函数的计算公式为:

$$\mathcal{L}_i(\theta_i) = E_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right].$$

式中, θ_i 是神经网络参数在第 i 次迭代时的更新, θ^- 是 Target 网络的参数集合, $r + \gamma \max_{a'} Q(s', a'; \theta^-)$ 表示 Target 网络在第 i 次迭代时的 Q 值。

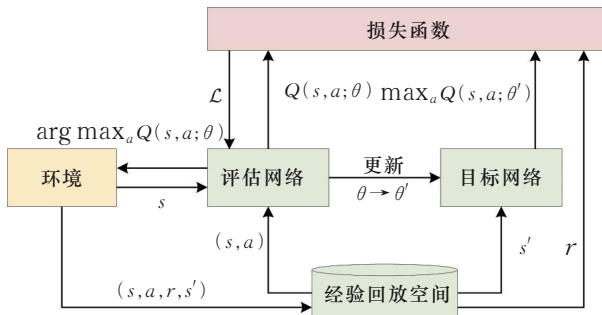


图1 DQN训练过程

Fig.1 DQN training processes

为了缓解DQN中过高的 Q 值估计问题,VAN H 等人^[18]提出通过使用两个网络(目标网络和行为网络)分别估算最大 Q 值和选择动作。Double DQN的目标网络的 Q 值计算方式如下:

$$y_i = r + \gamma Q(s', \arg \max_a Q(s', a; \theta_i); \theta^-).$$

其中 $\arg \max_a Q(s', a; \theta_i)$ 表示使用行为网络 θ_i 选择下一个状态的最佳动作, $Q(s', \cdot; \theta^-)$ 表示使用目标网络 θ^- 来评估选定动作的价值。

另外,在某些情况下动作的选择对于结果的影响较小,而状态本身的价值更为重要,为此 WANG 等人^[19]提出 Dueling DQN 的改进方法,将 Q 值分解为价值流 $V(s)$ 和优势流 $A(s, a)$,以便更好地分离状态价值和动作价值。 Q 值的计算为:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + A(s, a; \theta, \alpha) - \frac{1}{|A|} \sum_{a'} A(s, a'; \theta, \alpha).$$

2 方法论

2.1 问题描述

渗透测试是一个典型的序贯决策问题^[20],其目

标是在每个时间状态下通过观测被测系统状态,并执行攻击动作。在此过程中,智能体通过环境反馈的奖励来衡量当前攻击动作的优劣,进而调整攻击策略,找到最优攻击方案。而马尔科夫决策过程是一个用于解决有限状态、动作下路径生成问题的框架,能够描述一个智能体在一个随机环境中采取的决策情况^[8,21]。因此,基于DQN的自动渗透测试问题可以建模成一个马尔科夫决策过程。

具体而言:渗透测试环境中主机控制状态、配置、端口开放等情况可以表示为一个状态矩阵,对应马尔科夫过程的状态空间,即一个状态转移到另一个状态的随机过程。资产信息收集、漏洞扫描、漏洞利用等操作的结果存在不确定性,因此可以利用马尔科夫决策的转移概率量化这种不确定性。而在渗透测试过程中,对于当前状态 s_t 和动作 a_t ,马尔科夫决策过程根据奖励函数和状态转移函数得到下一个状态 s_{t+1} 和即时奖励 r_t 反馈给智能体。智能体根据当前状态 s_t 从动作集合中选择一个动作,旨在最大化长期累积奖励,即马尔科夫决策过程中的策略。马尔科夫决策过程通常由元组 $\langle S, A, R, F, \gamma \rangle$ 表示,其中 S 表示环境的状态空间, A 为动作空间, $R: S \times A \rightarrow R$ 表示奖励函数, $F: S \times A \times S \rightarrow [0, 1]$ 为状态转移概率函数,即攻击者执行攻击 $a_t \in A$ 后,状态 s_t 转移到 s_{t+1} 的概率; γ 为折扣因子,被用于确定长期奖励的重要性。

1) 状态空间:智能体在渗透过程中通过扫描环境获取到的信息可表述为一个状态矩阵,状态矩阵的每一行元素表示网络中每个主机节点的状态以及与其他主机的连接关系,定义如下:

$$s^{(i)} = [\{M_{i,j}\}_{1 \times N}, \text{Ass}(M_i), \text{Attr}(M_i)].$$

其中, $\{M_{i,j}\}_{1 \times N}$ 表示主机节点 i 与网络其他节点的连接关系, N 为网络规模; $\text{Ass}(M_i)$ 和 $\text{Attr}(M_i)$ 分别表示主机节点 i 的资产信息以及漏洞信息。因此整个测试环境的状态空间为 $\mathcal{S} = [s^{(1)}, s^{(2)}, \dots, s^{(N)}]^T$ 。状态空间 $\mathcal{S}$ 中每一行表述为一个特定的主机节点的状态。因此存在一个需要注意的情况,即:随着探索的主机节点的增多,以及暴露的漏洞的增加,将使得状态空间的维度不断增多。

2) 动作空间:智能体在渗透测试过程中与目标环境交互可以执行的一组攻击行为,包括资产信息收集(如端口扫描、服务扫描、操作系统检测、系统软硬件识别等)、漏洞扫描(如操作系统漏洞扫描、服务漏洞扫描、软硬件漏洞探测等)、漏洞利用(根

据已发现的漏洞采取的特定服务)。

3)奖励函数:智能体的目标是收集有关目标主机的额外信息,以最小的攻击成本,提升攻击利用漏洞对主机进行攻击的成功率。因此,对环境中的任意主机节点*i*而言,其奖励函数 $r^{(i)}$ 可表示为:

$$r^{(i)} = \text{Value}(i) + \text{Score}(v_i) - \sum_{a \in \mathcal{A}} \text{Cost}(a)$$

其中 $\text{Value}(i)$ 表示主机节点*i*的价值; $\text{Score}(v_i)$ 为主机节点*i*漏洞评分,它可通过通用漏洞评分系统 CVSS 给出的公式以及主机节点*i*的漏洞信息计算;而 $\text{Cost}(a)$ 表示执行攻击*a*所花费的成本,可以依据攻击执行时间、资源消耗等指标综合计算。

4)优化目标:神经网络的优化目标是通过执行策略 π^* 最大化累计奖励。因此,攻击行为价值函数 Q-function 可以表示为:

$$Q^*(s, a) = \max_{\pi \in \Pi} E_{\tau \sim \mathbb{D}_\pi} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) | s_t = s, a_t = a \right]$$

因此,针对状态*s*可以得到最优的攻击行为 a^* :

$$a^* = \arg \max Q^*(s, a)$$

根据以上建模方式,可以对渗透测试过程进行形式化描述,但在实际应用中 also 面临一些问题:1)在规模较为复杂的被测环境中,系统整个状态空间庞大,并且存在大量的无效探索行为,从而造成算法难以收敛;2)当探索到未知主机和漏洞时,会新增一些系统状态信息,导致状态空间的维度增加,

进而导致状态空间维度与 DQN 网络输入维度不匹配问题。为此,本文提出一种改进深度强化学习方法,通过引入先验知识以及注意力机制来解决上述问题。

2.2 知识嵌入

相较于让智能体盲目探索,将先验知识融入 DQN 中,有助于提升智能体的学习能力。因此,本文提出了一种将系统先验知识和安全先验知识嵌入到 DQN 学习过程的方法,具体如图 2 所示。首先,根据电力物联网的拓扑结构(本文以配网自动化系统为例),建立 Node 连接模型,并根据 Node 的连接关系,进行向量化,得到关系矩阵 $M = \{M_{i,j}\}_{N \times N}$,其中 $M_{i,j}$ 表示节点*i*和节点*j*是否存在连接关系,1 表示连接,-1 表示不连接,*N* 表示网络规模。然后根据知识库实例化测试环境中每个节点。

知识库包括节点的资产属性和漏洞信息。其中节点的资产属性包括 Hardware、OS、Software、Port、Service 等,而漏洞信息根据 CVE 漏洞库和 CVSS 漏洞评分系统获得漏洞的复杂性、可利用性及相关的安全风险数据^[22]。对节点的资产属性和漏洞属性进行向量化,分别得到资产属性向量 $\text{Ass}(M_i) = \{d_1, d_2, \dots, d_m\}$ 和漏洞属性向量 $\text{Attr}(M_i) = \{c_1, c_2, \dots, c_n\}$ 。最后按照图 2 所示方式得到状态矩阵 *S*。

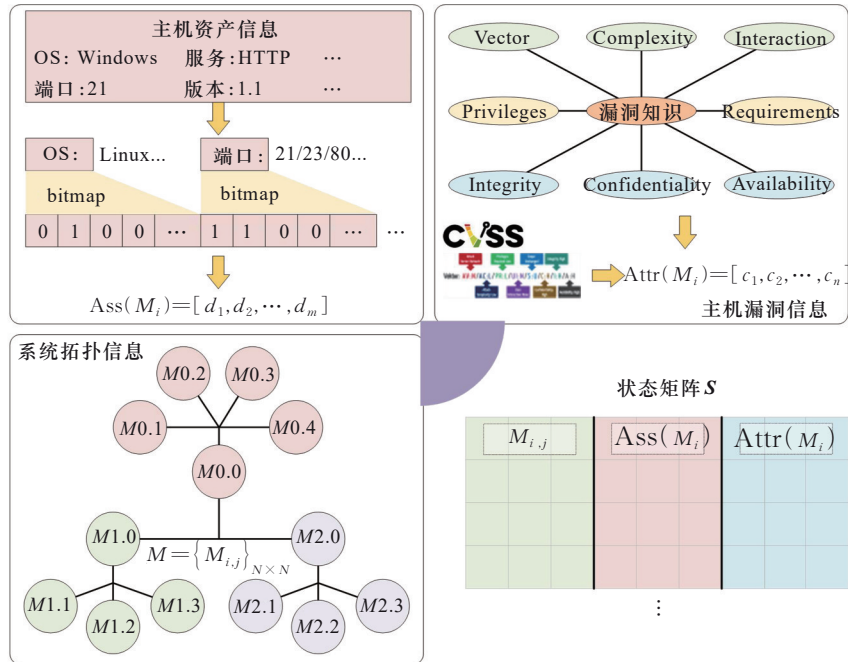


图2 基于先验知识的状态矩阵构建

Fig.2 State matrix construction based on prior knowledge

通过先验知识构建关系矩阵后,有助于智能体选择下一个目标主机节点进行渗透攻击,降低无效

路径探索次数,从而减少训练时间。另外,随着测试环境规模复杂度增加,智能体可能存在使用大量

的攻击动作的情况,为此本文在传统贪婪策略基础上,引入专家经验知识,形成带有经验知识的贪婪策略。令 ϵ 表示随机探索的概率, ϵ_e 表示选择专家经验动作的概率,则贪婪策略可以表示为

$$\pi = \begin{cases} a_e, & \text{以概率 } \epsilon_e \text{ 选择专家经验动作,} \\ a^*, & \text{以概率 } 1 - \epsilon - \epsilon_e \text{ 选择 DQN 估计动作,} \\ a, & \text{以概率 } \epsilon \text{ 进行随机探索。} \end{cases}$$

通过这种方式,当智能体面临未知状态时,以概率 ϵ_e 选择专家经验动作可以提供较为稳健的决策方式,避免重复产生大量错误决策;而以概率 $1 - \epsilon - \epsilon_e$ 选择 DQN 估计的最优动作,确保了智能体能够根据当前所学知识作出最优决策,从而在长期训练中得到更高的累计奖励。

2.3 模型构建

图 2 所示的方法中,状态矩阵 S 将会随着渗透测试过程探索的主机节点数量增加而增加。具体而言,状态矩阵 S 的每一行包括 3 个部分:表示主机节点连接关系的向量 M_i ,表示主机节点资产属性的向量 $\text{Ass}(M_i)$,以及表示主机节点漏洞属性的向量 $\text{Attr}(M_i)$ 。由于随着渗透测试的持续进行,被测系

统的主机信息、资产信息以及漏洞信息会不断被发现,从而导致状态矩阵的维度不断增加。因此对于 DQN 网络来说,其输入维数会随之变化。然而一般 DQN 采用全连接层神经网络表示,这种网络拓扑要求各层神经元数量是固定的。但由于基于 DQN 的渗透测试过程中,智能体只能在其动作空间内选择动作。如果对于新发现的主机资产信息和漏洞信息没有可供选择的动作,智能体也无法利用这些信息。因此,可以根据智能体的动作空间预先将资产信息和漏洞信息的长度定义为固定长度(内容可暂时为空值)。通过这种处理方式,只需要处理新发现主机节点造成的状态矩阵维度变化问题。

为了解决上述问题,本文引入了一种注意力机制,以适应状态矩阵 S 动态变化情况,具体如图 3 所示。首先将状态矩阵拆分成多个具有时间属性的 H 向量,即: $H_{1,t} = M_1, H_{2,t} = \text{Ass}(M_1), H_{3,t} = \text{Attr}(M_1), H_{4,t} = M_2$ 等。图 3 所示注意力机制部分, H 向量的数量是可增加的,通过计算各个关系的加权和,以获得维度保持不变表示方法:

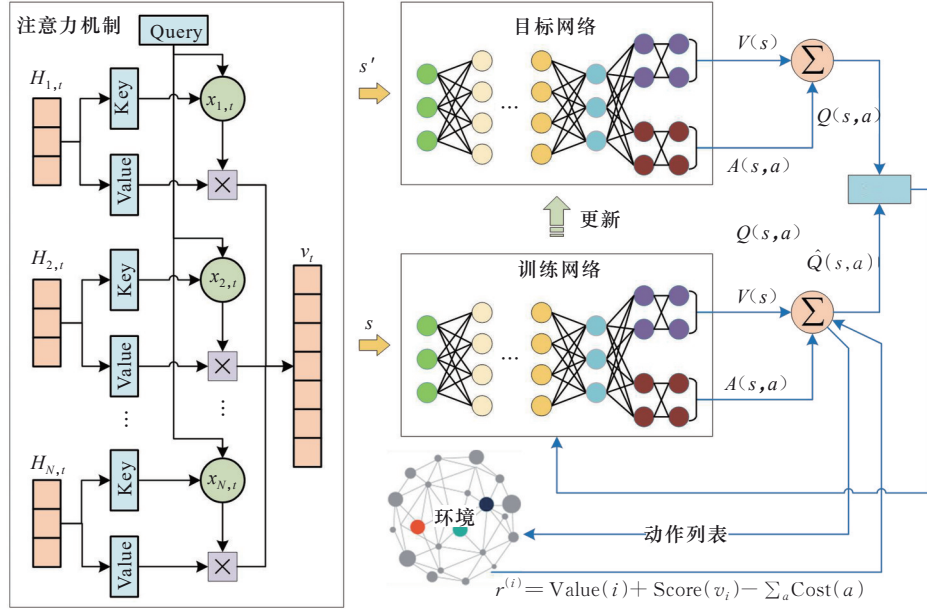


图 3 基于改进 DQN 的渗透测试训练框架

Fig.3 Penetration testing training framework based on improved DQN

$$v_i(H_1, H_2, \dots, H_N) = \sum_{i=1}^N \chi_{i,t} \cdot x.$$

其中 x 表示输入序列中每个元素的内容; $\chi_{i,t}$ 表示注意力权重。权重因子的计算方法可采用 softmax 函数表示:

$$\chi_{i,t} = \text{softmax}(l_{i,t}) = \frac{\exp(l_{i,t})}{\sum \exp(l_{i,t})},$$

$$l_{i,t} = x_i W_k^T W_q x_i$$

其中 W_k 和 W_q 为自动学习的参数, W_k 映射为网络的 Key 值,而 W_q 映射为 query。因此,通过上述处理

方式,无论智能体发现的主机节点数量为多少,DQN网络需要学习的参数都可以保持不变。一方面可以简化训练过程,另一方面还有助于增加智能体对环境动态变化的适应性。

图3所示右半部分为DQN的实现框架。首先从环境中捕获状态信息 s ,并构造矩阵 Z ;然后结合注意力机制提取融合的状态特征,输入到训练网络 Θ_{train} 中;随后将噪声网络应用到神经网络中得到各层连接的权重 ω_i ,并使用dueling结构计算价值流 $V(s)$ 和优势流 $A(s,a)$;应用double DQN矫正 Q 值 $\hat{Q}(s,a)$,并选择使得 Q 值最大的动作 a^* 应用到目标环境中,存储 (s,a^*,r,s') 到buffer B 中。重复上述步骤直到存储的buffer满足训练要求。

3 案例分析

本章节主要通过构建实验场景验证所提方法的有效性,主要回答两个研究问题:

1)所提出的方法是否能够减少智能体的迭代次数以实现快速收敛。

2)所提出的方法是否适应不同规模的网络拓扑。

3.1 实验环境设计

为了回答上述研究问题,在Windows 10环境下基于GNS3(version 2.2.52)和VMware Workstation Pro 17搭建了如图4所示的配电自动化系统渗透测试环境。包括1个主站和2个子站,共10台主机、3台子网交换机和1台核心交换机;每个站内部署一个敏感主机,即渗透测试的目标主机(操作员站:M0.1,FTU:M1.2和M2.2)。目标环境中主机的配置信息如表1所示,其中敏感主机的价值为100,而非敏感主机的价值为0。主站的四台主机采用Windows操作系统,配置HTTP、SSH、FTP、SAMBAs等服务,子站的主机采用Linux操作系统,配置FTP、SSH等服务;交换机采用Cisco image直接加载到GNS3中。攻击主机即为渗透测试主机,连接通过网络桥接映射到主站的物理以太网接口,其目标是发现网络中的三个敏感主机,并完成渗透测试。

另外,为了验证方法的可扩展性,进一步构建了不同网络规模的测试环境,分别采用50、100台主机进行搭建,命名为场景2和场景3,并且仍然维持3个敏感主机的数量不变。因此随着网络规模的扩大,奖励变得更加稀疏。

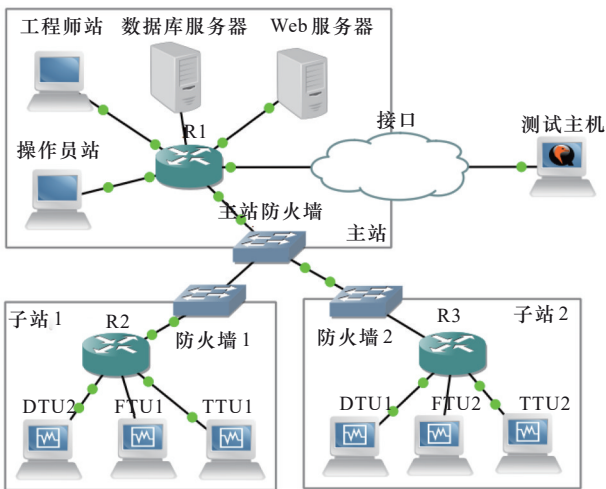


图4 渗透测试实验场景

Fig.4 Experimental scenario for penetration testing

表 1 实验场景配置

Table 1 Configurations in the experimental scenario

主机	操作系统	服务	进程	价值
M0.1	Windows	Http	Schtask	100
M0.2	Windows	Http, Ssh	Schtask	0
M0.3 M0.4	Windows	Http FTP Samba	Daclsvc	0
M1.1 M1.3 M2.1 M2.3	Linux	Ssh	Tomcat	0
M1.2 M2.2	Windows	FTP	Tomcat	100
M0 M0.0 M1.0 M2.0	Cisco	FTP Ssh	—	100

选择经常被攻击者利用的服务和进程作为被测环境漏洞,例如智能体可以利用FTP漏洞来获取访问权限,或者利用Tomcat漏洞从受感染的主机上提升用户权限级别。表2给出了本实验环境中智能体可用的动作,以及执行动作的代价、成功概率以及成功后可获得的权限。动作列表包括四种特定服务的漏洞利用操作、三种特定进程的提权操作、三种扫描操作。漏洞利用操作的成功概率根据CVSS的高、中、低难度级别分别设定为0.2、0.5和0.8,而其他操作的成功概率设定为1。

表 2 智能体的动作列表

Table 2 Action list for the agent

动作	策略	代价	概率	权限
Http-Exp	利用	2	0.8	用户
Ssh-Exp	利用	3	0.8	用户
Ftp-Exp	利用	1	0.5	管理员
Samba-Exp	利用	2	0.5	管理员
Schtask-Pr	提权	1	1	管理员
Daclsvc-Pr	提权	1	1	管理员
Tomcat-Pr	提权	1	1	管理员
Service-Scan	扫描	1	1	—
OS-Scan	扫描	1	1	—
Process-Scan	扫描	1	1	—

攻击主机的渗透测试工具在基于 AMD R9 处理器、A4000 显卡、64G 内存的 Linux 环境中使用 Python 编写,采用 PyTorch 作为算法的使用框架。实验的部分超参数如表 3 所示。

表 3 模型训练的部分超参数
Table 3 A part of hyperparameters for model training

参数名称	参数值	参数描述
Max iteration	10 000	最大迭代步数
Episodes	2 000	训练回合数
Batch size	128	训练批次
Memory size	20 000	回放空间大小
Discount size	0.9	折扣因子
Learning rate	0.002 5	学习率

3.2 结果讨论

3.2.1 不同方法的性能比较(针对问题 1)

智能体的学习目标是优化其行为策略,以便使用尽可能少的步数获取目标环境中所有敏感主机的权限,从而最大化奖励值。这个奖励机制作为评估智能体策略效果的重要指标,直接反映了智能体在环境中的表现优劣。针对图 4 所示实验环境,分别采用传统 DQN、Double DQN、Dueling DQN 以及本文所提的方法进行实验,实验结果如图 5 所示。图 5(a)为每回合累计奖励值随训练回合数的变化趋势,而图 5(b)则为每回合所需的步数随训练过程的变化情况。

从图 5(a)的数据中可以看出,在训练初期,由于智能体对环境的理解有限,因此它每回合所能获得的奖励值相对较小,每回合累计奖励集中在-50~-250 之间。但随着训练次数的增加和智能体对环境理解的加深,其每回合能够获取的奖励值也逐渐上升。传统 DQN、Double DQN 和 Dueling DQN 均需要超过 800 回合才能收敛,而本文提出的改进算法大约在 500 个回合就能快速收敛至最大奖励值。此外,观察图 5(b)中关于每回合所需步数的变化,可以发现本文提出的方法虽然在训练初期每回合需要较多的迭代步数,最大需要超过 10 000 步,而其他对比方法约 7 000 步以内;但本文所提出的方法经过约 200 回合后每回合需要的迭代步数少于 2 500 步,并且在 500 回合后步数达到最低(约 200 步),并且可以一直稳定维持这个最低步数。对比而言,其它方法至少需要 800 回合才能达到最低步数,其中传统 DQN 算法最不稳定,在经过 1 500 训练回合后,迭代步数偶尔仍然需要超 2 000 步。通过图 5 所示结果说明,所提出的方法不仅在收敛速

度上领先,而且在整个训练过程中展现出了更高的稳定性,表明该算法能够在更短的时间内找到最优路径,减少不必要的探索步骤,从而提高了整体的学习效率。

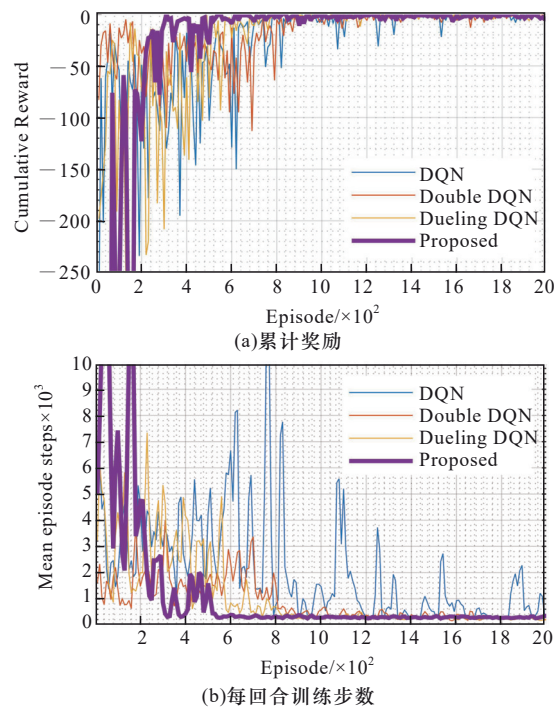


图 5 渗透测试训练结果(图 4 所示场景)

Fig.5 Training results for penetration testing (in the scenario shown in Fig.4)

3.2.2 不同网络规模下性能比较(针对问题 2)

图 6 给出了另外 2 个实验场景(50 个主机和 100 个主机)下不同算法的训练曲线。可以看出,随着网络中主机规模的增加,训练收敛需要更多的回合。尤其是当网络主机规模增加到 100 时,由于奖励变得越发稀疏(敏感主机数量占比只有 2%),不同模型表现出的性能差异更大。在本文构建的实验场景下,传统 DQN 的性能表现相对较差,分别需要约 1 500 回合(场景 2)和 1 800 回合(场景 3)才能收敛;而 Double DQN 和 Dueling DQN 的性能有所改良,但也均需要约 1 000 回合(场景 2)和 1 400 回合(场景 3)才能收敛。本文所提出的方法在 2 个场景下分别在 600 回合(场景 2)、800 回合(场景 3)左右收敛,相较于其他算法,收敛速度更快。

4 结语

渗透测试作为信息安全防护的重要手段,对提高电力物联网系统的信息安全防护能力具有重要意义。本文针对电力物联网自动渗透测试需求,提出一种基于改进深度强化学习的自动渗透测试方

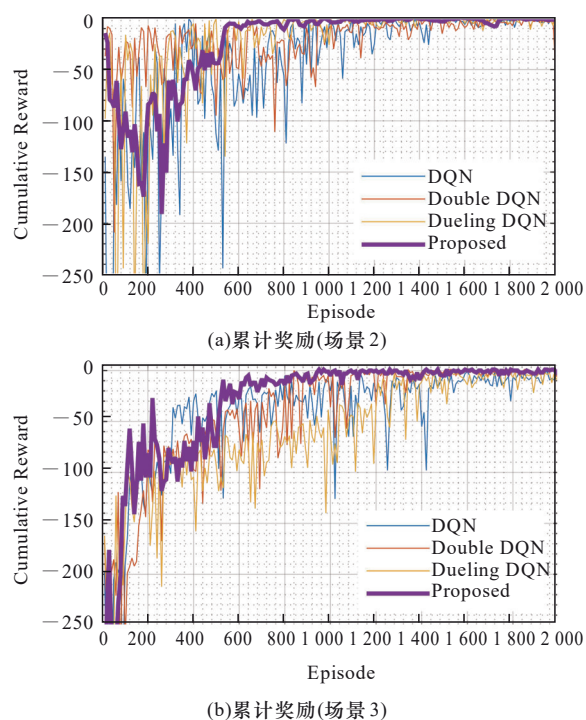


图6 不同场景下的渗透测试训练结果

Fig.6 Training results for penetration testing in the different scenarios

法。通过将电力物联网系统的拓扑结构、资产属性及漏洞信息等先验知识嵌入深度强化学习框架,并引入注意力机制动态适应状态空间变化,有效解决传统方法面临的状态空间爆炸、奖励稀疏、行业适应性问题。最后通过搭建三个不同规模的试验环境,验证了所提出方法的有效性。实验结果表明,所提方法在对应的试验环境下分别可以在500回合、600回合以及800回合收敛,相较于传统DQN、Double DQN及Dueling DQN等算法,其测试效率有大幅度提升。

然而,目前的工作仍基于模拟的实验环境,对真实场景的因素考虑较为欠缺,例如目标环境的漏洞仅注入少数漏洞,可能与真实情况存在差异;未来可以考虑集成传统工具提供的模块库作为智能体动作空间的基础,并通过搭建半实物或者实物模拟环境进行测试验证,推动自动渗透测试技术在电力物联网行业的应用落地。

参考文献:

- [1] 张彦杰,王辉,李延,等. 电力物联网下基于SM9的CP-ABE访问控制方案[J]. 太原理工大学学报,2025,56(3):453-464.
ZHANG Y J, WANG H, LI Y, et al. SM9-based CP-ABE access control scheme under Power IoT[J]. Journal of Taiyuan University of Technology, 2025, 56(3): 453-464.
- [2] 王金浩,樊瑞,肖莹,等. 基于低碳需求响应的新型电力系统中源-网-荷-储协调的优化调度[J]. 太原理工大学学报,2024,55(1):46-56.
WANG J H, FAN R, XIAO Y, et al. Optimal dispatch of coordinated source-grid-load-storage in new power system based on low-carbon demand response[J]. Journal of Taiyuan University of Technology, 2024, 55(1): 46-56.
- [3] 谷良,狄婷,杜锋. 电力系统中基于微分博弈的拟态防御调度策略[J]. 太原理工大学学报,2025,56(5):410-418.
GU L, DI T, DU F. Differential game based mimicry defense scheduling strategy in power systems[J]. Journal of Taiyuan University of Technology, 2025, 56(5): 410-418.
- [4] 王晓凡,周天阳,臧艺超,等. 大规模网络渗透测试路径规划方法研究[J]. 计算机应用与软件,2023,40(5):324-330.
WANG X F, ZHOU T Y, ZANG Y C, et al. Path planning method of large-scale network penetration test[J]. Computer Applications and Software, 2023, 40(5): 324-330.
- [5] RAHALKAR S, RAHALKAR K, KARKA. Quick start guide to penetration testing[M]. USA: BerKerly Apress, 2019.
- [6] ALKHURAYYIF Y, ALMARSHDY Y S. Adopting automated penetration testing tools: A cost-effective approach to enhancing cybersecurity in small organizations[J]. Journal of Information Security and Cybercrimes Research, 2024, 7(1): 51-66.
- [7] 陈可,鲁辉,方滨兴,等. 自动化渗透测试技术研究综述[J]. 软件学报,2024,35(5):2268-2288.
CHEN K, LU H, FANG B X, et al. Survey on automated penetration testing technology research[J]. Journal of Software, 2024, 35(5): 2268-2288.
- [8] 董卫宇,刘鹏坤,刘春玲,等. 基于深度强化学习 Noisy Net-A3C 算法的自动化渗透测试方法[J/OL]. 郑州大学学报(工学版), <https://doi.org/10.13705/j.issn.1671-6833>. 2024. 02. 011.
DONG W Y, LIU P K, LIU C L, et al. Automated penetration testing method based on deep reinforcement learning Noisy Net-A3C algorithm[J/OL]. Journal of Zhengzhou University (Engineering Science), <https://doi.org/10.13705/j.issn.1671-6833>. 2024. 02. 011.
- [9] MEYERS B S, ALMASSARI S F, KELLER B N, et al. Examining penetration tester behavior in the collegiate penetration testing competition[J]. ACM Transactions on Software Engineering and Methodology, 2022, 31(3): 1-25.
- [10] LI Q, WANG R, LI D, et al. DynPen: Automated penetration testing in dynamic network scenarios using deep reinforcement

- learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19:8966-8981.
- [11] LI Q,ZHANG M,SHEN Y, et al. A hierarchical deep reinforcement learning model with expert prior knowledge for intelligent penetration testing[J]. Computers & Security, 2023, 132:103358.
- [12] 刘磊,许静,朱静雯,等. 基于攻击反馈模型的SQL注入漏洞渗透测试方法[J]. 计算机应用与软件, 2023, 40(6):323-329.
LIU L,XU J,ZHU J W, et al. SQL injection vulnerability penetration test approach based on attacking feedback model[J]. Computer Applications and Software, 2023, 40(6):323-329.
- [13] MAEDA R,MIMURA M. Automating post-exploitation with deep reinforcement learning[J]. Computers & Security, 2021, 100:102108.
- [14] BEURAN R,HU Z,ZENG Y, et al. Artificial intelligence for cybersecurity education and training Artificial Intelligence and Cybersecurity: Theory and Applications[M]. Cham: Springer International Publishing, 2022:103-123.
- [15] 李成恩,朱东君,贺杰彦,等. 基于强化学习的智能化渗透路径规划与求解优化[J]. 计算机科学, 2024, 51(11):329-339.
LI C G,ZHU D J,HE J Y, et al. Intelligent penetration path planning and solution optimization based on reinforcement learning [J]. Computer Applications and Software, 2024, 51(11):329-339.
- [16] RAILKAR D N,JOSHI S. Penetration testing framework using the Q learning ensemble deep CNN discriminator framework [J]. International Journal of Advanced Computer Science & Applications, 2024, 15(3):845-856.
- [17] 史腾飞,王莉,臧嵘. 深度强化学习局部策略迁移方法[J]. 太原理工大学学报, 2024, 55(4):705-711.
SHI T F,WANG L,ZANG R. Deep reinforcement learning local policy transfer method[J]. Journal of Taiyuan University of Technology, 2024, 55(4):705-711.
- [18] VAN H H,GUEZ A,SILVER D. Deep reinforcement learning with double Q-learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1):2094-2100.
- [19] WANG Z,SCHAUL T,HESSLE M, et al. Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2016:1995-2003.
- [20] 曾庆伟,张国敏,邢长友,等. 基于分层强化学习的智能化攻击路径发现方法[J]. 计算机科学, 2023, 50(7):308-316.
ZENG Q W,ZHANG G M,XING C Y, et al. Intelligent attack path discovery based on hierarchical reinforcement learning [J]. Computer Science, 2023, 50(7):308-316.
- [21] 占力戈,沙乐天,肖甫,等. 基于强化学习的自动化Windows域渗透方法[J]. 网络与信息安全学报, 2023, 9(4):104-120.
ZHAN L G,SHA L T,XIAO P, et al. Automated windows domain penetration method based on reinforcement learning[J]. Chinese Journal of Network and Information Security, 2023, 9(4):104-120.
- [22] LACHKOV P,TAWALBEH L,BHATT S. Vulnerability assessment for applications security through penetration simulation and testing[J]. Journal of Web Engineering, 2022, 21(7):2187-2208.

(编辑:贾丽红)