计算机科学
COMPUTER SCIENCE

基于改进近端策略优化算法的智能渗透路径研究

王紫阳, 王佳, 熊明亮, 王文涛

引用本文

王紫阳, 王佳, 熊明亮, 王文涛. 基于改进近端策略优化算法的智能渗透路径研究[J]. 计算机科学, 2024, 51(11A): 231200165-6.

WANG Ziyang, WANG Jia, XIONG Mingliang, WANG Wentao. Intelligent Penetration Path Based on Improved PPO Algorithm [J]. Computer Science, 2024, 51(11A): 231200165-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200165

基于改进近端策略优化算法的智能渗透路径研究

王紫阳　王　佳　熊明亮　王文涛
新疆大学计算机科学与技术学院　乌鲁木齐 830000
新疆维吾尔自治区多语种信息技术重点实验室　乌鲁木齐 830000
(107552103703@stu.xju.edu.cn)

摘　要　渗透路径规划是渗透测试的首要步骤,对实现渗透测试的自动化有重大意义。现有渗透路径规划研究多将渗透测试建模为完全可观测的理想过程,难以准确反映部分可观测性的实际渗透测试过程。鉴于强化学习在渗透测试领域的广泛应用,将渗透测试过程建模为部分可观测的马尔可夫决策过程,从而更准确地模拟实际渗透测试过程。在此基础上,针对 PPO 算法使用全连接层拟合策略函数和价值函数无法提取部分可观测空间有效特征的问题,提出一种改进的 PPO 算法 RPPO,其中策略网络和评估网络均融合全连接层和 LSTM 网络结构以提升其在未知环境提取特征的能力。同时,给出一种新的目标函数更新方法,以增强算法的鲁棒性和收敛性。实验结果表明,在不同网络场景中,相较于现有 A2C,PPO 和 NDSPI-DQN 算法,RPPO 算法收敛轮次分别缩短了 21.21%,28.64%,22.85%,获得累计奖励分别提升了 66.01%,58.61%,132.64%,更适用于超过 50 台主机的较大规模网络环境。

关键词:渗透测试;渗透路径规划;强化学习;近端策略优化;长短期记忆网络
中图分类号　TP309

Intelligent Penetration Path Based on Improved PPO Algorithm

WANG Ziyang,WANG Jia,XIONG Mingliang and WANG Wentao
School of Computer Science and Technology,Xinjiang University,Urumqi 830000,China
Xinjiang Key Laboratory of Multilingual Information Technology,Urumqi 830000,China

Abstract　Penetration path planning is the first step of penetration testing,which is important for the intelligent penetration testing. Existing studies on penetration path planning always model penetration testing as a full observable process,which is difficult to describe the actual penetration testing with partial observability accurately. With the wide application of reinforcement learning in penetration testing,this paper models the penetration testing as a partially observable Markov decision process to simulate the practical penetration testing accurately. In general,the full connection of policy network and evaluation network in PPO cannot extract features effectively in penetration testing with partial observability. This paper proposes an improved PPO algorithm RPPO,which integrating of full connection and long short term memory(LSTM) in the policy network and evaluation network. In addition,a new objective function updating is designed to improve the robustness and convergence. Experimental results show that,the proposed RPPO converges faster than A2C,PPO and NDSPI-DQN algorithms. Especially,the convergence iterations is reduced by 21.21%,28.64% and 22.85% respectively. Meanwhile RPPO gains more cumulative reward about 66.01%,58.61% and 132.64%,which is more suitable for larger-scale network environments with more than fifty hosts.

Keywords　Penetration testing,Penetration path planning,Reinforcement learning,Proximal policy optimization,Long and short term memory networks

1　引言

渗透测试是一种通过模拟恶意黑客攻击来评估计算机系统、网络或应用程序安全性的重要方法[1]。传统渗透测试需要测试人员具备专业的安全知识和技能,并需要持续数天或数周来反复执行,费时费力。随着目标网络规模的持续增长,可能的渗透路径组合也呈指数型增加,为渗透测试带来巨大挑战。如何实现自动化渗透测试以减少专家依赖,节约时间和成本等,成为渗透测试的研究重点。渗透路径规划[2]作为渗透测试的首要步骤,是实现自动化渗透测试的关键技术之

基金项目:新一代人工智能国家科技重大专项(2022ZD0115803);新疆维吾尔自治区重点研发计划项目(2022B01008);国家自然科学基金项目(62363032);新疆维吾尔自治区自然科学基金项目;新疆维吾尔自治区教育厅项目(XJEDU2022P011);"天池博士"计划项目(202104120018)
This work was supported by the National Science and Technology Major Project(2022ZD0115803),Key Research and Development Program of Xinjiang Uygur Autonomous Region(2022B01008),National Natural Science Foundation of China(62363032),Natural Science Foundation of Xinjiang Uygur Autonomous Region(2023D01C20),Scientific Research Foundation of Higher Education(XJEDU2022P011) and "Heaven Lake Doctor" Project(202104120018).
通信作者:王佳(jw1024@xju.edu.cn)

一,已成为该领域的研究热点。

强化学习[3]是智能体通过与环境的多轮交互来实现目标的一种机器学习方法。智能体是拥有学习与决策的实例,主要包括感知、执行、学习 3 部分。其中,感知部分为智能体对环境进行观测,从而获得信息;执行部分为当前观测信息下应采取的动作;学习部分为智能体的学习算法。智能体与环境进行交互生成数据,并根据生成数据学习如何在环境中采取动作的最优策略。换言之,强化学习中智能体策略的改变会影响其与环境交互产生的数据分布。考虑有监督学习中数据分布固定不变的情况,强化学习更适用于实时交互环境,因而为自动化渗透路径的发现提供了新思路。

在实际场景的渗透测试中,渗透人员往往难以获得完整的网络状态信息,所能观测到的当前动作的相关信息(如主机 IP 信息、目标主机系统等),通常被称为部分可观测信息。为了有效描述渗透测试过程中的不确定性,特别是观测信息的部分可观测性,本文引入部分可观察马尔可夫决策过程[4](Partially Observable Markov Decision Process, POMDP)用以表征网络信息未知情况下的特征。将实际场景中的渗透测试建模为 POMDP 更有现实意义。

近年来,基于强化学习的渗透路径发现方法已成为智能化渗透测试领域的研究热点[5-7],但依然还存在着较多挑战。

(1)在实际场景的渗透测试中,渗透人员往往难以获得完整的网络信息,因此渗透测试过程无法满足马尔可夫性。由于 POMDP 可表达网络信息不完整情况下的特征,将实际场景中的渗透测试建模为 POMDP 更有现实意义。同时在 POMDP 模型中,由于智能体观测信息的不完整性,导致适用于 MDP 的强化学习算法在 POMDP 中通常难以进行求解[8]。

(2)随着网络规模的不断扩大,智能体在渗透过程中所能探测到的主机数量不断增多,从而导致智能体的状态空间呈指数型增长。由于智能体在获得敏感主机权限后才能得到奖励,因此随着状态空间不断增大,奖励稀疏问题亦随之而来。

针对实际渗透测试场景中存在的网络信息部分可观测问题,本文将渗透测试建模为 POMDP 问题,提出一种融合全连接层和 LSTM 网络结构的 RPPO 算法。主要贡献如下:

(1)考虑现有强化学习算法难以求解 POMDP 问题,在策略网络和评估网络中引入 LSTM 模型更好地提取未知状态动作空间中的特征,从而增强智能体学习时序样本数据的能力;

(2)在策略更新过程中采用 Clip-Annealing 方法,同时将 Tanh 函数作为激活函数来增强非线性表达能力,进而提高算法的鲁棒性和收敛性;

(3)采用开源网络攻击模拟器 NASim[9]的部分可观测模式进行训练,并与现有方法进行比较,验证所提算法的有效性。

2 相关工作

根据观测信息是否为部分可观测,现有渗透测试路径研究基本分为基于马尔可夫决策的渗透测试路径和基于部分可观测马尔可夫决策的渗透测试路径。

基于马尔可夫决策的渗透测试路径研究通常假定渗透测试过程具备马尔可夫性,以模拟实际场景中的确定性。针对动作空间庞大和奖励稀疏问题,Zennaro 等[10]将渗透测试过程简化为网络安全夺旗赛,采用 Table-Q 学习算法验证强化学习方法进行渗透测试的可行性。Zhou 等[11]提出 NDSPI 算法来适应大规模网络场景,从而得到更好的收敛性、鲁棒性和可扩展性。Zhang 等[12]提出一种带有渗透动作选择模块的改进型 PPO 算法 IPPOPAS,且随着主机中漏洞数量的增加,该算法的收敛速度加快。Chen 等[13]利用专家知识对深度学习模型进行预训练,同时采用 GAIL 判别器进行训练。所提 GAIL-PT 具有良好的稳定性和更低的成本,但 GAIL-PT 的网络结构依然较复杂。Zhou 等[14]提出一种基于网络信息增益的路径发现算法 NIG-AP,采用网络信息熵来获取激励并指导智能体选择最佳行动。NIG-AP 可在无先验知识的情况下自动发现攻击路径。Cody 等[15]将真实网络的信息抽象成攻击图,提出一种基于 MDP 的分层参考模型,用以帮助后续智能化渗透测试系统的研究开展。日本立命馆大学学者[16]提出一种基于 Wolpertinger 架构的 A2C 算法,在处理多达 1000 台主机的大型网络环境时,采用 A2C 算法的 DAA 模型远远优于其他基于 DQN 的模型。然而上述研究均将渗透测试建模为马尔可夫决策过程,其中目标网络配置的完整信息(如网络节点上运行的服务和网络拓扑等)均可获取,与实际渗透测试环境还有一定差距。

考虑实际渗透测试场景中存在的部分可观测空间问题,部分学者将渗透测试过程建模为 POMDP。Sarraute 等[17]将攻击者的不完整信息引入渗透测试的模拟过程,同时将原有较大网络场景划分为较小网络场景逐一进行求解。Zhang 等[18]将黑盒测试建模为 POMDP,提出 ND^{3}RQN 算法来应对主机数量低于 40 的渗透测试场景。Ghanem 等[19]提出一个智能渗透测试系统 IAPTS,将渗透测试环境和任务建模为 POMDP,并使用 POMDP 求解器进行求解。考虑 POMDP 难以求解和网络规模导致的问题复杂化特性,上述方法无法扩展到较大规模网络场景。Ghanem 等[20]提出一种基于分层网络模型的自动化渗透测试方法,针对大型网络规模导致的复杂 POMDP 问题,将网络场景划分为多个安全集群并使用外部 POMDP 求解器分别求解。由于该方法无法进行端到端的训练,求解效率受到限制。Koroniotis 等[21]提出基于深度学习的渗透测试框架,用以解决物联网环境中的零日漏洞问题,其检测攻击准确率显著优于其他技术,但该研究主要目的是漏洞发现而不是渗透测试过程。Schwartz 等[22]将博弈论与 POMDP 模型进行融合,引入信息衰减因子以模拟网络防御的动态性,使模型对实际渗透测试建模更加真实。

由于渗透测试过程建模为 POMDP 的特殊性,针对现有方法求解繁琐、结构复杂且无法适用于大规模网络环境,本文提出一种端到端的融合全连接层和 LSTM 网络结构的 RPPO(Recurrent PPO)算法。

3 RPPO 算法

考虑到部分可观测渗透测试数据的时序依赖性和模型的收敛速率问题,本文提出一种端到端的 RPPO 算法作为智能体的学习算法。RPPO 算法框架如图 1 所示。设 θ 和 μ 分别是基于时序依赖的策略网络和基于时序依赖的评估网络的权重参数。所有权重参数进行随机初始化。假定 I 和 N 分别为总迭代次数和轨迹回放池中的采样阈值,其值均根据网络环境中的主机数量、子网数量等环境参数动态变化。在任意的第 i 次迭代过程中,主要包括有轨迹数据产生和渗透路径

发现模型。在轨迹数据产生阶段，任意的第 t 个时间步，由当前 θ 所得的策略 π_θ 与环境不断交互产生轨迹数据 $\{o_t^i, a_t^i, r_t^i, o_{t+1}^i\}$ 并存放于轨迹回放池中。其中 o_t^i, a_t^i, r_t^i，与 o_{t+1}^i 分别代表在第 i 次迭代中第 t 个时间步所得到的观测信息、动作、奖励以及下一个时间步的观测信息，当轨迹数据量阈值超过 N 时，则进入渗透路径发现模型。假定 M 为模型训练轮次。在第 m 轮训练中，随机从轨迹回放池中抽样出 S 条轨迹数据，分别用于更新基于时序依赖的策略网络和基于时序依赖的评估网络中的 θ 和 μ。当 M 轮次的模型训练结束后，得到更新后的 θ 和 μ，清空轨迹回放池，根据新产生的策略 π_θ 和环境交互开始下一次迭代。

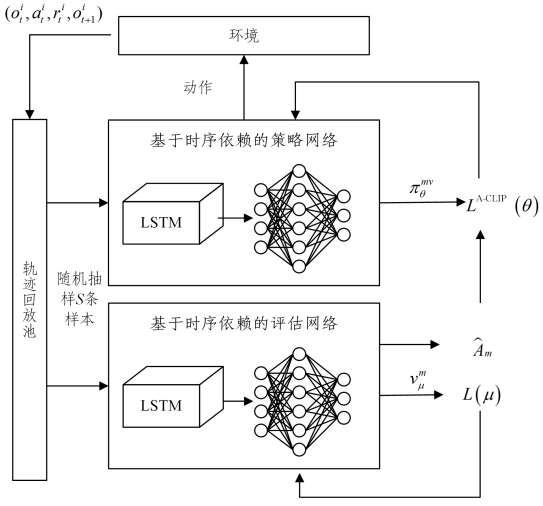


图 1　RPPO 算法结构

Fig. 1　Framework of RPPO algorithm

3.1　基于时序依赖的策略网络

本文将渗透测试建模为 POMDP 后，原有 PPO 算法中的全连接层（Full Connected Layer，FC）无法从未知环境中有效提取特征。同时考虑 LSTM 网络对时序特征提取的优越性能，本文提出基于时序依赖的策略网络以提升智能体在部分可观测环境中对时序样本的学习能力。

在第 m 轮训练中，采样轨迹回放池中的 S 条 POMDP 环境中的轨迹数据作为基于时序依赖的策略网络输入数据输入策略网络结构。其网络结构如图 2 所示，其中输入层神经元数量为 128 个，LSTM 层数为 1，全连接层层数为 2，神经元数量均为 128，采用 Relu 函数作为激活函数。相较于传统的外部 POMDP 求解器方法，将 LSTM 网络融入策略网络对 POMDP 模型进行求解，可以利用 LSTM 中隐状态来替代 POMDP 中信念状态，因而无需人工设计信念状态表示[7]，更易于实现与应用。基于时序依赖的策略网络的输出为当前轮次得到的策略值 π_θ^{mv}，从而得到新旧策略比 $r_\theta = \pi_\theta^{mv} / \pi_\theta^{v}$。其中 π_θ^v 是策略 π_θ 经过当前以 θ 为权重参数的策略网络得到的策略值。

强化学习算法训练前期增加策略更新幅度能够使策略尽快逼近优化方向；同时在模型训练后期，为增加稳定性并提升收敛速度，则应逐渐减小策略更新的幅度，故本文对现有 PPO-CLIP 方法的截断范围进行优化。由于 $clip(x, l, r) = \max(\min(x, r), l)$ 是将 x 限制在 $[l, r]$ 内，为保证策略 r_θ 在小范围内线性衰减以提高模型的收敛效率，提出动态截断区间的目标函数计算方法。根据基于时序依赖的评估网络所得到的优势函数 $\hat{A}_m$ 和 r_θ，目标函数计算如式(1)所示。

$$L^{A\text{-}CLIP}(\theta) = \hat{E}_t[\min(r_\theta \hat{A}_m, clip(r_\theta, 1-\varepsilon, 1+\varepsilon)\hat{A}_m)] \quad (1)$$

其中，ε 是动态截断参数，用于辅助设置策略更新的范围，设置为 $[0.2, 0.3]$。根据目标函数 $L^{A\text{-}CLIP}(\theta)$ 对策略网络参数进行更新，从而得到 π_θ^v 以便于下一轮次的参数更新。

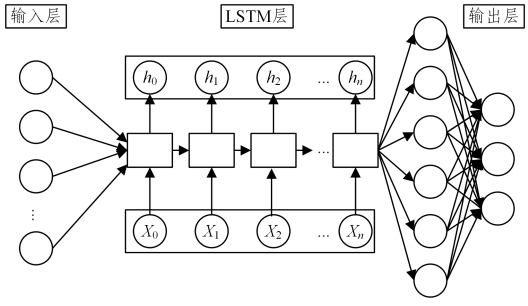


图 2　基于时序依赖的策略网络结构

Fig. 2　Policy network based on sequential dependency

3.2　基于时序依赖的评估网络

考虑 POMDP 问题的部分可观测特性，基于时序依赖的评估网络由 LSTM 层和全连接层组成，与基于时序依赖的策略网络结构相同。在第 m 轮训练中，基于时序依赖的策略网络中使用的 S 条 POMDP 环境中的轨迹数据作为输入数据进入基于时序依赖的评估网络，其输出为状态价值函数 v_μ^m 和优势函数 $\hat{A}_m$。该网络的目标函数计算如式(2)所示。

$$L(\mu) = \frac{1}{2}(r_t + \gamma v_\mu^m(o_{t+1}) - v_\mu^m(o_t))^2 \quad (2)$$

其中，γ 为折扣因子。根据 $L(\mu)$ 对评估网络的权重参数 μ 进行更新。

4　实验分析

4.1　实验配置

网络攻击模拟器（Network Attack Simulator，NASim）[9]是一款使用 Python 语言编写的轻量级计算机网络仿真环境。在抽象出真实网络环境特征的同时，保留渗透测试的核心。NASim 支持网络动态变化，能够真实地反映实际网络环境，使研究人员能够专注于网络渗透测试路径规划性能的评估。鉴于 NASim 在渗透测试问题中仿真测试的普遍性[9,11-13,16,18,23]，本文采用 NASim 的部分可观测模式进行实验。通常来说，随着网络规模增大，算法观测信息维度进一步增加，导致算法从高维观测信息中提取有效特征更为困难，从而影响算法的效率和性能。为验证所提算法的性能，文章在 5 种不同规模的场景下分别进行了实验，各网络场景配置如表 1 所列。

表 1　不同规模的网络场景配置

Table 1　Different settings of network scenarios

场景	主机	敏感主机	子网数	进程数
场景 1	10	2	4	3
场景 2	30	2	8	5
场景 3	50	2	12	6
场景 4	75	2	17	6
场景 5	100	2	22	6

文中所比较算法均采用 Pytorch 框架、Python3.8 语言

来编写。硬件配置为 Intel(R) Xeon(R) Platinum 8255C CPU，NVIDIA GeForce RTX2080TiGPU，操作系统为 Linux。不同规模网络场景实验中所用的超参数设置如表 2 所列。

表 2 实验超参数设置

Table 2 Hyperparameters setting

超参数	含义	值
S	轨迹回放池的抽样数据量	128
Learning_rate	学习率	0.0003
γ	折扣因子	0.99
M	模型训练轮次	10

网络场景规模逐步增大，导致智能体的观测空间维度以及所需要的操作同步增长，增加轨迹回放池的数据量阈值 N 可以帮助智能体更好地探索环境。经过多次训练，场景 1－5 中的数据量阈值 N 分别设置为 128，256，512，1 024，2 048。同时，采用与文献[12,16,23]中相同的环境设置，即敏感主机价值 $value(host)$为 100，普通主机价值为 1。

假设 t 为当前时间步，动作代价 $cost(a)$为动作 a 所产生的代价，则执行动作后的奖励r_t的计算如(3)所示：

$$r_t = value(host) - cost(a) \tag{3}$$

在规定迭代次数内，每次迭代所获得的累计奖励值收敛越快，则模型性能越好，其计算式如式(4)所示：

$$R = \sum_{t=1}^{N} r_t \tag{4}$$

其中，r_t为当前轮次轨迹回放池中第 t 个时间步采样数据对应的奖励值。

4.2 对比实验结果分析

以累计奖励 R 为评估指标，对比 A2C 算法[12]、PPO 算法[16]、DNSPI-DQN 算法[23]与所提算法在不同场景下的实验结果。实验过程中，并不是每一次动作都能够获得主机价值。随着网络场景规模的增加，智能体会进行大量的探索，从而带来巨大的动作代价。根据式(3)中各时间步奖励r_t的计算方式，此时所得到的奖励r_t通常为负数。

图 4 为场景 1 中每次迭代累计奖励随迭代次数的变化。从图 4 中可以发现，在较小的实验场景中，不同算法的累计奖励差距并不大，但在收敛速度方面，RPPO 算法显著优于其他算法，RPPO 算法收敛轮次至少减少了 24.57%，获得累计奖励提升了 15.34%。在前 200 次迭代中，RPPO 算法已经完成收敛，这充分体现了 RPPO 模型借助 LSTM 网络能够在 POMDP 环境中进行更好的决策，这是因为 LSTM 的隐状态可以提取样本时序数据中的长期依赖关系，从而辅助模型更好地观测当前的环境状态信息。

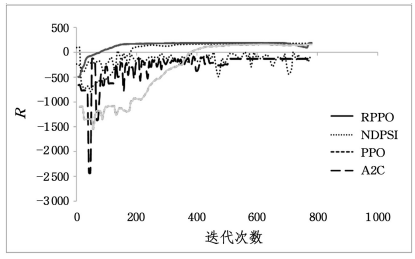
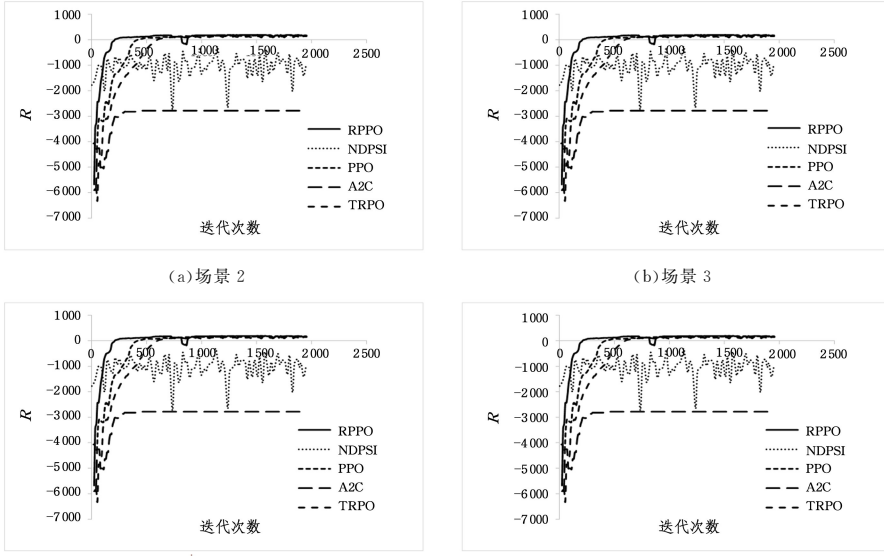


图 4 累计奖励随训练轮次的变化(场景 1)

Fig. 4 Cumulative reward varies withtraining rounds in scenario 1

场景 2－场景 5 的实验结果如图 5 所示。与场景 1 相比，场景 2 进一步增加了主机数量与搭载的服务数量，文章所提 RPPO 算法并未受到网络规模增加所带来的影响，但 A2C 算法则由于其策略更新梯度不受限制而陷入了局部最优，无法获得更多奖励。同时 NDSPI-DQN 算法已经无法收敛，故在后续场景 3－场景 5 实验中并未与其进行对比。为进一步验证 RPPO 算法在较大网络规模下的性能，场景 3－场景 5 将主机数量提升至 50，75 与 100。实验结果表明，随着网络场景增加，算法所需要的迭代次数骤增，此时其他对比算法已陷入局部最优，而 RPPO 算法的累计奖励值最高，验证了在 PPO 算法中加入 LSTM 结构能有效提升算法解决渗透测试中的部分可观测问题。

综上所述，与 A2C 算法、PPO 算法和 NDSPI-DQN 算法相比，文章所提算法可学习到更优的策略，每次迭代获取的累计奖励更高，收敛速度也更快。这主要是由于 RPPO 算法使用 LSTM 结合全连接层改进了原有策略网络、评估网络结构，并使用线性衰减的方式改进目标函数的截断范围，有效提升了算法的性能。



(a)场景 2 (b)场景 3

(c)场景 4 (d)场景 5

图 5 场景 2－场景 5 下累计奖励随训练轮次的变化

Fig. 5 Cumulative reward varies with training iterations in scenarios 2, 3, 4, and 5

4.3 消融实验结果分析

本文提出的 RPPO 算法是对 PPO 算法的改进，在策略与评估网络中融合 LSTM 网络，并引入了改进后的目标函数。为进一步验证所提算法的有效性，分别对采用不同策略和评估网络以及目标函数得到的 RPPO，RPPO-LSTM，RPPO-CLIP 和 PPO，PPO-RNN 算法进行消融实验，各算法如表 3 所列。如 RPPO-CLIP 算法表示策略和评估网络采用 LSTM 与 DNN 网络，目标函数采用原 PPO 算法目标函数$L^{\mathrm{CLIP}}(\theta)$构成的算法模型。

表 3 消融实验算法

Table 3 Ablation experiment algorithms

算法名称	策略、评估网络结构	改进后的目标函数
RPPO	LSTM+DNN	$L^{\text{A-CLIP}}(\theta)$
RPPO-LSTM	DNN	$L^{\text{A-CLIP}}(\theta)$
RPPO-CLIP	LSTM+DNN	$L^{\mathrm{CLIP}}(\theta)$[12]
PPO	DNN	$L^{\mathrm{CLIP}}(\theta)$[12]
PPO-RNN	RNN+DNN	$L^{\text{A-CLIP}}(\theta)$

图 6 给出不同场景下各算法收敛时的累计奖励值。实验结果表明，RPPO 算法表现最佳，凸显了融合 LSTM 结构后的策略与评估网络以及新目标函数的有效性。RPPO-LSTM 算法与 RPOO-CLIP 算法移除 LSTM 结构与改进后的目标函数后，二者整体收敛速度和累计奖励上均有所下降，说明 LSTM 结构在捕捉状态信息时序依赖性方面的关键作用以及改进后目标函数在模型训练上的优势。PPO-RNN 算法在场景 1—场景 3 中表现较为良好，收敛时累计奖励略逊于 RPPO 算法，但随着网络场景规模增大，POMDP 模型中观测信息维度也进一步增加，由于 RNN 本身在处理长序列时难以捕捉长期依赖关系，导致 PPO-RNN 算法性能骤然下降。PPO 算法在场景 1 与 2 中表现较为良好，当网络场景规模进一步增大时，PPO 算法便很难从网络环境中采样到正值奖励样本，大量无效负值奖励样本导致其累计奖励较低。

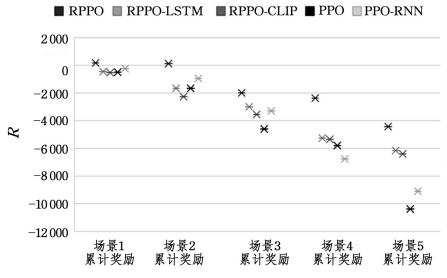


图 6 不同场景下各算法收敛时的累计奖励值

Fig. 6 Cumulative reward when each algorithm converges in different scenarios

综上所述，RPPO 算法相较于其他变体算法能够获得更高的累计奖励，这进一步证实了 LSTM 网络融合策略、评估网络以及策略网络目标函数的改进对提高算法在 POMDP 环境中性能的重要性。在不同网络场景中，相较于现有 A2C，PPO 和 NDSPI-DQN 算法，RPPO 算法收敛轮次分别缩短了 21.21%，28.64%，22.85%，获得累计奖励分别提升了 66.01%，58.61%，132.64%。在 100 台主机以内的网络场景下，RPPO 算法能够获得较多的累计奖励。

结束语 考虑实际渗透测试过程中存在的部分可观测特性，文章将渗透测试过程建模为 POMDP。同时，现有 PPO 算法无法学习未知网络环境中存在的隐含信息，采用 LSTM 提取时序特征以提升算法解决渗透测试中部分可观测问题的能力。此外，文章给出一种新的目标函数更新方法以增强算法的鲁棒性和收敛性。实验结果表明，与 A2C，PPO 和 NDSPI-DQN 算法相比，RPPO 算法收敛速度更快，获得累计奖励更高，且在较大的网络规模中表现良好。

当前基于强化学习的渗透测试依旧停留在模拟阶段，暂时无法在真实环境下进行直接应用；现有的仿真环境过于理想化，并未考虑到现有的防御手段，例如拟态防御、移动目标防御、网络安全态势感知等，因此如何尽可能将模拟结果应用到真实网络环境进行理论分析以及如何在拥有防御手段的网络中进行智能化渗透测试路径是下一步工作计划。

参 考 文 献

[1] ARKIN B, STENDER S, MCGRAW G. Software Penetration Testing[J]. IEEE Security & Privacy, 2005, 3(1): 84-87.

[2] SARRAUTE C, RICHARTE G, LUCÁNGELI OBES J. An Algorithm to Find Optimal Attack Paths in Nondeterministic Scenarios[C]// Proc. of the 4th ACM Workshop on Security and Artificial Intelligence. Chicago, US, 2011: 71-80.

[3] SILVER D, HUANG A, MADDISON C J, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search[J]. Nature, 2016, 529(7587): 484-489.

[4] WARRINGTON A, LAVINGTON J W, SCIBIOR A, et al. Robust Asymmetric Learning in Pomdps[C]// Proc. of the 38th International Conference on Machine Learning (PMLR). New York, US, 2021: 11013-11023.

[5] VAN OTTERLO M, WIERING M. Reinforcement Learning and Markov Decision Processes[M]// Reinforcement learning: State-of-the-art. Berlin, Heidelberg: Springer, 2012: 3-42.

[6] MCKINNEL D R, DARGAHI T, DEHGHANTANHA A, et al. A Systematic Literature Review and Meta-analysis on Artificial Intelligence in Penetration Testing and Vulnerability Assessment[J]. Computers & Electrical Engineering, 2019, 75: 175-188.

[7] MAEDA R, MIMURA M. Automating Post-exploitation with Deep Reinforcement Learning[J]. Computers & Security, 2021, 100: 102-108.

[8] LADOSZ P, BEN-IWHIWHU E, DICK J, et al. Deep reinforcement learning with modulated hebbian plus Q-network architecture[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(5): 2045-2056.

[9] SCHWARTZ J, KURNIAWATI H. Autonomous Penetration Testing Using Reinforcement Learning[J]. arXiv: 1905.05965, 2019.

[10] ZENNARO F M, ERDŐDI L. Modelling Penetration Testing with Reinforcement Learning Using Capture-the-flag Challenges: Trade-offs between Model-free Learning and A Priori knowledge[J]. IET Information Security, 2023, 17(3): 441-457.

[11] ZHOU S, LIU J, HOU D, et al. Autonomous penetration testing based on improved deep q-network[J]. Applied Sciences, 2021, 11(19): 8823.

[12] ZHANG G M, ZHANG S Y, ZHANG J W. Attack Path Discovery and Optimization Method Based on PPO Algorithm[J]. Information Network Security, 2023, 23(9): 47-57.

[13] CHEN J, HU S, ZHENG H, et al. GAIL-PT: An Intelligent

Penetration Testing Framework with Generative Adversarial Imitation Learning[J]. Computers & Security, 2023, 126: 103055.

[14] ZHOU T, ZANG Y, ZHU J, et al. NIG-AP: A New Method For Automated Penetration Testing[J]. Frontiers of Information Technology & Electronic Engineering, 2019, 20(9): 1277-1288.

[15] CODY T. A Layered Reference Model for Penetration Testing with Reinforcement Learning and Attack Graphs[C]// Proc. of 2022 IEEE 29th Annual Software Technology Conference (STC). Gaithersburg, MD, USA, IEEE, 2022: 41-50.

[16] NGUYEN H V, TEERAKANOK S, INOMATA A, et al. The Proposal of Double Agent Architecture using Actor-critic Algorithm for Penetration Testing[C]// ICISSP. 2021: 440-449.

[17] SARRAUTE C, BUFFET O, HOFFMANN J. POMDPs Make Better Hackers: Accounting for Uncertainty in Penetration Testing[C]// Proc. of the 26th AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada, 2012: 1816-1824.

[18] ZHANG Y, LIU J, ZHOU S, et al. Improved Deep Recurrent Q-Network of POMDPs for Automated Penetration Testing[J]. Applied Sciences, 2022, 12(20): 10339.

[19] GHANEM M C, CHEN T M. Reinforcement Learning for Efficient Network Penetration Testing[J]. Information, 2019, 11(1): 6.

[20] GHANEM M C, CHEN T M, NEPOMUCENO E G. Hierarchical Reinforcement Learning for Efficient and Effective Automated Penetration Testing of Large Networks[J]. Journal of Intelligent Information Systems, 2023, 60(2): 281-303.

[21] KORONIOTIS N, MOUSTAFA N, TURNBULL B, et al. A deep learning-based penetration testing framework for vulnerability identification in internet of things environments[C]// 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2021: 887-894.

[22] SCHWARTZ J, KURNIAWATI H, EL-MAHASSNI E. Pomdp+ information-decay: Incorporating defender's behaviour in autonomous penetration testing[C]// Proceedings of the International Conference on Automated Planning and Scheduling. 2020, 30: 235-243.

[23] ZHOU S, LIU J, HOU D, et al. Autonomous Penetration Testing Based on Improved Deep Q-network[J]. Applied Sciences, 2021, 11(19): 8823.

WANG Ziyang, born in 1996, postgraduate. His main research interests include reinforce learning and cyberspace security.

WANG Jia, born in 1987, Ph. D, associate professor, is a member of CCF (No. K8521M). Her main research interests include resource allocation in clouds, tasks scheduling in big data and cyberspace security.