计算机科学
COMPUTER SCIENCE

基于最大熵强化学习的最优渗透路径生成方法

王焱，王天荆，沈航，白光伟

引用本文

王焱，王天荆，沈航，白光伟. 基于最大熵强化学习的最优渗透路径生成方法[J]. 计算机科学，2024，51(3)：360-367.

WANG Yan, WANG Tianjing, SHEN Hang, BAI Guangwei. Optimal Penetration Path Generation Based on Maximum Entropy Reinforcement Learning [J]. Computer Science, 2024, 51(3): 360-367.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200104

基于最大熵强化学习的最优渗透路径生成方法

王 焱 王天荆 沈 航 白光伟
南京工业大学计算机科学与技术学院 南京 211816
(wy8573200@163.com)

摘 要 从攻击者角度分析入侵意图和渗透行为对于指导网络安全防御具有重要意义。然而,现有的渗透路径大多依据瞬时的网络环境构建,导致路径参考价值降低。针对该问题,文中提出了一种基于最大熵强化学习的最优渗透路径生成方法,该方法可以在网络环境动态变化的情况下,以探索的形式捕获多种模式的近似最优行为。首先,依据攻击图和漏洞评分对渗透过程进行建模,通过量化攻击获益来刻画渗透行为的威胁程度;然后,考虑到入侵行为的复杂性,开发基于最大熵模型的 Soft Q-学习方法,通过控制熵值和奖励的重要程度来保证求解渗透路径的过程具有稳定性;最后将该方法应用于动态变化的测试环境中,生成高可用的渗透路径。仿真实验结果表明,相比于现有基于强化学习的基准方法,所提方法具有更强的环境适应性,能够以更低的代价生成更高收益的渗透路径。

关键词:最大熵强化学习;攻击图;Soft Q-学习;渗透路径
中图分类号 TP393

Optimal Penetration Path Generation Based on Maximum Entropy Reinforcement Learning

WANG Yan,WANG Tianjing,SHEN Hang and BAI Guangwei
College of Computer Science and Technology,Nanjing Tech University,Nanjing 211816,China

Abstract Analyzing intrusion intentions and penetration behaviors from the attackers' perspective is of great significance for guiding network security defense. However,most existing penetration paths are constructed based on the instantaneous network environment,resulting in reduced reference value. Aiming at this problem,this paper proposes an optimal penetration path generation method based on maximum entropy reinforcement learning,which can capture the approximate optimal behavior of multiple modes in the form of exploration under dynamic network environments. Firstly,the penetration process is modeled according to the attack graph and the vulnerability score,and the threat degree of the penetration behavior is described by quantifying the attack benefits. Then,considering the complexity of the intrusion behavior,a soft Q-learning method based on the maximum entropy model is developed. The stability of the penetration path is ensured by controlling the entropy value and the importance of the reward. Finally,the method is applied to a dynamic environment to generate a highly available penetration path. Simulation experimental results show that,compared with the existing baseline methods based on reinforcement learning,the proposed method has more robust environmental adaptability and can generate higher-yielding penetration paths at a lower cost.

Keywords Maximum entropy reinforcement learning,Attack graph,Soft Q-learning,Penetration path

1 引言

随着互联网应用的共享开放以及新技术的爆发式发展,网络体系架构日益复杂,各种网络安全问题也层出不穷,给用户和企业带来了严重的安全威胁和经济损失。结合威胁情报分析、安全态势评估、入侵检测等技术构建网络安全体系可以应对诸多安全风险[1]。近年来,攻击者利用网络安全漏洞实施恶意攻击的手段和形式日益复杂。为了分析攻击者的行为,网络渗透一直是广受关注的网络安全问题。渗透过程可以分为直接渗透和间接渗透,其中直接渗透是从一个主机直接发起攻击并成功渗透到另一个主机的过程;间接渗透是通过渗透成功两个主机之间的跳板主机,才能从一个主机渗透到另一个主机的过程[2]。

在了解渗透原理的基础上,网络管理者可以模拟真实的网络攻防场景,分析可能被攻击者利用的渗透路径,从而设计有效的网络安全防御手段。然而,由于攻击者选择渗透行为的不确定性,加之网络态势的不断改变以及主机漏洞利用难度的客观评价等,渗透路径的构建难度大大增加。与此同时,网络安全防御系统的建立和完善也直接影响评判渗透行为是否客观高效。因此,合理科学地设计

到稿日期:2022-12-15 返修日期:2023-05-22
基金项目:国家自然科学基金(61502230,61501224);江苏省自然科学基金(BK20201357);江苏省"六大人才高峰"高层次人才项目(RJFW-020)
This work was supported by the National Natural Science Foundation of China(61502230,61501224),Natural Science Foundation of Jiangsu province,China(BK20201357) and Six Talent Peaks Project in Jiangsu Province(RJFW-020).
通信作者:沈航(hshen@njtech.edu.cn)

渗透路径生成系统面临严峻的挑战。

渗透路径生成主要依赖攻击图[3-5]与 Markov 链结合的模型来分析可能被攻击者渗透的路径。Kaynar[6]等详细说明了攻击图的生成是一个收集网络拓扑和应用信息、确定网络主机之间可达条件和设计核心图构建算法的过程。Sun 等[7]提出了一种基于攻击图和 Markov 链的网络安全风险评估模型，通过分析计算原子节点的攻击转移概率来获取最大可能被攻击的渗透路径。考虑到内部攻击和未知攻击，Wang 等[8]提出了一种基于知识图谱的双层威胁渗透图模型，基于知识图谱构建主机资源知识图谱，在此基础上生成主机威胁渗透图和网络渗透威胁图。实验结果表明，该方法能够描述未知攻击和内部攻击。上述常见的渗透路径生成方法未考虑渗透成功所导致的后续攻击路径的变化，为此 Tang 等[9]通过将告警集映射到因果知识网络来检测当前的攻击行为，根据能力等级动态调整知识分布，利用改进的 Dijkstra 算法计算出真实网络对抗环境下的最优攻击路径。为了解决攻击图的伸缩性限制，Zaid 等[10]通过使用 A* Prune 算法去除无用边来降低攻击图的复杂度，并采用随机森林算法预测网络拓扑中的攻击位置，从而生成概率最大的渗透路径。

传统的渗透路径生成仍然存在如下问题：1）人工建模渗透测试环境的成本高，无法应对大规模复杂体系网络场景[11-12]；2）人工渗透测试结果往往取决于测试人员的个人经验与能力，无法真实地反映出攻击者对目标网络和攻击路径选择的可能性。机器学习的方法成为了解决路径规划问题的重要手段[13]。

Wang 等[14]提出了一种全局引导强化学习方法，该方法在移动机器人遇到障碍时，利用环境的时空信息引导机器人做出局部路径调整，而无需重新调用规划算法寻找替代路径，提高了模型的泛化性。将 AI 路径规划引入渗透路径生成过程可以显著提升生成效率。Koo 等[15]在漏洞数据库创建的攻击图上，利用机器学习和深度学习生成渗透路径。Cody 等[16]提出了发现渗透路径的 RL 方法，智能体通过与环境交互学习而在攻击图中发现最优的多条攻击路径。这为子网间的路由器防火墙、认证日志跟踪和基于主机的防病毒等防御措施提供了参考依据。Zhang 等[17]引入了智能体多域动作选择模块的方法来发现更多隐藏的多域渗透路径，提出了改进的 DDPG 算法，使得智能体能在不同的状态下选择不同的动作，提高了网络多域安全防御能力。目前渗透测试的智能化方法的侧重点为如何高效地提高渗透路径的生成效率，没有考虑面对网络态势环境动态改变的情况下，智能体如何选择最优的动作行为设计。

针对上述问题，本文提出了一种基于最大熵强化学习模型的最优渗透路径生成方法，根据网络环境变化高效快速地生成最优渗透路径。本文的主要贡献为如下 3 个方面：

1）通过将漏洞评分和渗透路径长度相结合，来量化攻击者发起渗透攻击的获益，在此基础上来评判每一条渗透路径的优劣程度。

2）开发基于最大熵模型的 Soft Q-learning 方法，鼓励智能体在受到干扰时探索和开发大部分的状态空间，并设置密集的奖励机制和经验学习机制促使智能体快速学习到最优策略，从而获得最优渗透路径。

3）构建了一种具备网络安全防御功能的模拟环境，测试所提方法在网络环境动态变化下的性能。仿真结果表明，本文方法在适应性和路径质量上优于基准方法。

2　最优渗透路径问题

攻击者从一个主机发动攻击渗透到另一个主机最大可能的渗透路径称为最优渗透路径。可以利用攻击图（Attack Graph）和漏洞评分系统（Common Vulnerability Scoring System，CVSS）进行漏洞分析，从而对渗透路径进行问题建模。

2.1　攻击图构建

攻击图（Attack Graph，AG）是一种基于模型的网络安全评估技术。该技术可以建模真实网络场景，以评估目标网络中攻击者的攻击行为以及攻击状态的转移过程。

定义 1（攻击图）　攻击图可以用四元组 $AG=(S,E,A,\Delta)$ 表示，其中：

1）$S=\{s_i \mid i\in N\}$ 表示 N 个不同状态节点的集合，$s_i=\{0,1\}$，$s_i=1$ 表示该节点 i 已被渗透成功，反之，$s_i=0$ 表示节点 i 未被渗透成功。

2）$E=\{e_{i,j} \mid i,j\in N\}\subseteq S\times S$ 表示状态节点之间有向边的集合，$e_{i,j}$ 表示节点 i 和 j 之间存在有向边，即能够实现从 s_i 到 s_j 的一次状态转移。

3）A 代表原子攻击集合，原子攻击 a 表示对状态节点漏洞的一次成功利用。每个原子攻击动作触发攻击者转移到一个攻击状态 S，且原子攻击 a 的成功概率为 $P(a)$。

4）$\forall\Delta(e_{i,j})\in\Delta$，$\Delta(e_{i,j})$ 表示从状态 s_i 转移到状态 s_j 的概率 $P(e_{i,j})$，且 $P(e_{i,j})=P(a_i)$。

根据上述定义，给出图 1 所示的攻击图案例，其中 s_1 和 s_4 分别为起始状态和目的状态。不妨标注每条边上基于漏洞评分系统（Common Vulnerability Scoring System，CVSS）的漏洞评分和状态转移概率，以便通过目标值计算寻找最优渗透路径。为此，本文给出以下 3 个定义。

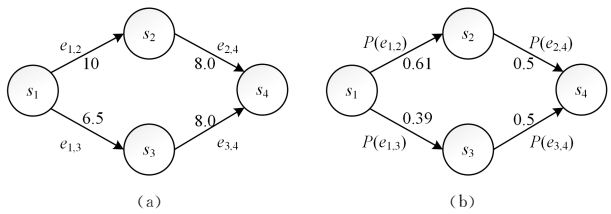


图 1　攻击图示例

Fig. 1　Example of attack graph

定义 2（攻击路径）　攻击路径（Attack Path，AP）指攻击者从初始状态节点到达目的状态节点的状态转移序列。例如，图 1(a)中 $AP_{1,4}:s_1\rightarrow s_2\rightarrow s_4$ 为一条从状态 s_1 到状态 s_4 的攻击路径。

定义 3（攻击路径长度）　攻击路径长度（Attack Path Length，APL）是一条攻击路径包含的有向边数量。例如，攻击路径 $AP_{1,4}$ 的长度为 2。

定义 4（攻击成功概率）　攻击成功概率（Probability of successful Attack，PA）是攻击路径中所有状态转移都成功的概率。例如，图 1 中 $P(AP_{1,4})=P(e_{1,2})\times P(e_{2,4})=0.305$。

利用脆弱性扫描工具 Nessus 等对网络进行脆弱性扫描，再依据网络拓扑结构和所得的脆弱性结果，结合攻击图自动生成工具 MulVAL，即可构建出攻击图。

2.2 漏洞评分

漏洞被利用的难易程度可以从可利用度、影响度和影响范围等方面进行量化。一般采用美国国家漏洞数据库(National Vulnerability Database，NVD)发布的通用 CVSS[18-19]进行评估。如表 1 所列，CVSS 的量化细则包括：

1)可利用度评价(Exploitability Metric)，包括攻击途径(Attack Vector，AV)、攻击复杂度(Attack Complexity，AC)、权限要求(Privilege Required，PR)和用户验证(User Interaction，UI)，用于反映漏洞不跟随时间和环境变化而变化的固有特征。

2)影响度评价(Impact Metrics)，包括机密性(Confidentiality Impact，C)、完整性(Integrity Impact，I)、可用性(Availability Impact，A)，用于反应漏洞被成功利用后造成的威胁程度。

3)影响范围(Scope)，包括范围固定(Unchanged)和范围变化(Changed)，用于反应漏洞是否会影响组件外的资源或获取其他权限。为了量化漏洞被利用的难易程度，下面给出原子攻击 a 依赖的漏洞得分 $Score(e_{i,j})$ 的计算式：

if $0<Impact+Exp<10$

$$Score(e_{i,j})=\begin{cases}Roundup(1.08(Impact+Exp)), & Scope=C\\ Roundup(Impact+Exp), & Scope=U\end{cases}$$

if $Impact+Exp\geqslant 10$

$$Score(e_{i,j})=10$$

其中：

$$Impact=\begin{cases}6.42ISCbase, & Scope=C\\ 7.52(ISCbase-0.029)-3.25(ISCbase-0.02)^{15}, & Scope=U\end{cases}$$

$$ISCbase=1-((1-C)(1-I)(1-A))$$

$$Exp=8.22AV\cdot AC\cdot PR\cdot UI \tag{1}$$

在实际应用中，$Roundup()$ 保留小数点后一位且小数点后的第二位如果大于零则进一位。例如：$Roundup(5.02)=5.1$，$Roundup(5.00)=5.0$；$Impact$ 表示影响度分值，$ISCbase$ 为临时变量，Exp 为漏洞可利用性得分。

表 1 CVSS 组件权重

Table 1 CVSS component weight

基础指标	影响因素	度量值	评分
可利用度评价	AV	Network	0.85
		Adjacent	0.62
		Local	0.55
		Physical	0.20
	AC	Low	0.77
		High	0.44
	PR	Low	0.62
		High	0.27
		None	0.85
	UI	Required	0.62
		None	0.85
影响度评价	C, I, A	None	0.00
		Low	0.22
		High	0.56
影响范围	S	Changed	
		Unchanged	

2.3 问题建模

最优渗透路径旨在在攻击图模型中的起始节点和目标节点之间寻找一条攻击获益最大的路径。攻击获益的衡量指标主要体现在：1)漏洞得分 $Score(e_{i,j})$，分值越高，证明漏洞越容易被利用，攻击难度越低；2)发起攻击的持续时间，即攻击者要在一定时间内完成渗透，否则防御者可能预警攻击修复网络。

$P_{S,D}$ 被定义为攻击图(AG)中初始节点 N_S 和目的节点 N_D 之间的路径。$P_{S,D}$ 由状态节点序列 $s_1,\cdots,s_n$ 组成，满足两个条件：1) $\forall i, 1\leqslant i\leqslant N$ 都有 $(s_i,s_{i+1})\in E$；2)每个节点仅在路径中作用一次。记 $P_{S,D}^1$ 和 $P_{S,D}^k(k>1)$ 分别为首次和第 k 次获得的路径，$L_{S,D}^1$ 和 $L_{S,D}^k$ 分别为首次和第 k 次获得的路径长度。

定义 0-1 变量 $z_{i,j}$ 表示攻击者是否选择从状态节点 s_i 转移到状态节点 s_j，满足：

$$z_{i,j}=\begin{cases}1, & \text{attacker moves from state } s_i \text{ to state } s_j\\ 0, & \text{otherwise}\end{cases}$$

则给定一个路径 $P_{S,D}$，路径的长度 $L_{S,D}$ 表示为：

$$L_{S,D}=\sum_{(i,j)\in E}z_{i,j} \tag{2}$$

攻击图中边上的漏洞分值计算式为 $d_{i,j}=Score(e_{i,j})$。$C(z_{i,j})$ 表示攻击者从状态 s_i 转移到状态 s_j 的攻击代价。$C(z_{i,j})$ 必须满足 3 个条件：1)每发生一次状态转移，就有唯一的 $C(z_{i,j})$ 对应；2) $L_{S,D}^1\geqslant L_{S,D}^k$ 时，不考虑攻击代价；3)当 $L_{S,D}^1<L_{S,D}^k$ 时，记 $L_{S,X}^k$ 为 $P_{S,D}^K$ 中的部分路径长度，且 $L_{S,X}^k=L_{S,D}^1$，$N_X=s_m$。$P_{S,D}^K$ 中从状态节点 s_m 开始计算后续路径中每一次状态转移后的攻击代价 $C(z_{m,j})$。v 表示攻击代价指数，E^* 表示当前攻击图中有向边的总数，H 表示漏洞评分 $Score(e_{i,j})$ 的最大值。攻击代价表示为：

$$C(z_{m,j})\triangleq d_{i,j}^v \tag{3}$$

其中，攻击代价指数 v 由 E^*，H，$L_{S,D}^1$ 共同确定：

$$v=e^{\frac{\ln H}{E^*-L_{S,D}^1}} \tag{4}$$

给定首次获得的渗透路径 $P_{S,D}^1$，当 $L_{S,D}^1\geqslant L_{S,D}^k$ 时，$C(z_{i,j})=0$，攻击者每次成功渗透一个节点 s_j 并发生一次状态转移，即 $z_{i,j}=1$，将获得有向边 $e_{i,j}$ 标注的漏洞得分 $d_{i,j}$ 作为一次攻击获益；当 $L_{S,D}^1<L_{S,D}^k$ 时，攻击者继续发起渗透攻击时，除了获得有向边 $e_{i,j}$ 标注的漏洞得分 $Score(e_{i,j})$ 外，还需考虑每一次成功渗透后的攻击代价 $C(z_{i,j})$。因此，攻击获益被量化为：

$$f=\begin{cases}\sum_{(i,j)\in E}d_{i,j}\cdot z_{i,j}, & L_{S,D}^1\geqslant L_{S,D}^K\\ \sum_{(i,j)\in E}d_{i,j}\cdot z_{i,j}-\sum_{(i,j)\in E}C(z_{i,j})\cdot z_{i,j}, & L_{S,D}^1<L_{S,D}^K\end{cases} \tag{5}$$

综上所述，最优渗透路径的实质是给定 $P_{S,D}^1$，在攻击图中找到一条能够最大化攻击获益的路径。该问题被建模为：

$$\max_{\{z_{i,j}\}}: x\sum_{(i,j)\in E}d_{i,j}\cdot z_{i,j}+(1-x)\left(\sum_{(i,j)\in E}d_{i,j}\cdot z_{i,j}-\sum_{(i,j)\in E}C(z_{i,j})\cdot z_{i,j}\right)$$

$$\text{s.t. } x=\begin{cases}1, & L_{S,D}^1\geqslant L_{S,D}^K\\ 0, & L_{S,D}^1<L_{S,D}^K\end{cases} \tag{6}$$

$$\begin{cases}\sum_{j=1}^{N} z_{i,j}=1, \forall s_i \in P_{S,D}^k \cap i \neq n \\ \sum_{i=1}^{N} z_{i,j}=1, s_j \in P_{S,D}^k \cap j \neq 1\end{cases} \tag{7}$$

约束条件(6)中，0-1 变量 x 依据不同条件下的取值决定攻击获益表达式。约束条件(7)保证第 k 次获得的路径 $P_{S,D}^k$ 满足：1)除目的状态节点外，每一个状态节点的出度为 1；2)除初始状态节点外，其余状态节点入度为 1。

3　方法设计

针对最优渗透路径问题，本节首先描述最优渗透路径问题可抽象为马尔可夫决策过程(Markov Decision Process，MDP)[20]模型，在此基础上提出 Soft Q-learning 方法，最后给出真实网络环境下的动态攻击图分析。

3.1　MDP 模型

强化学习可定义为马尔可夫决策过程的策略搜索，将之应用于最优渗透路径生成时，智能体(Agent)根据当前网络部署的状态信息和节点漏洞给出的环境反馈进行策略学习，完成最优渗透路径规划。下面详细描述 MDP 的四元组(S，A，R，P_S)。

1)状态空间 S：$S=\{s_t\}$ 表示当前网络状态的集合，其中 s_t 是智能体所感知的主机状态信息，例如主机权限、操作系统种类、访问接口等。

2)动作空间 A：$A=\{a_t\}$ 表示智能体动作的集合，其中 a_t 表示智能体是否可利用目标主机的漏洞进行渗透攻击。若 $a_t=0$，则不能渗透；若 $a_t=1$，则能够渗透。

3)状态转移概率 P_S：假设状态空间 S 和动作空间 A 都是连续的，状态转移概率 P_S 表示从状态 s_t 转移到下一步状态 s_{t+1} 的概率。

4)回报值 R：智能体学习期间需要对每一步选择执行的动作做出优劣评价，即获得一个短暂回报值。$R=\{r_t\}$ 表示回报值的集合，其中 r_t 表示给定状态 s_t 和动作 a_t 时，从状态 s_t 转移到下一步状态 s_{t+1} 反馈所得的奖励价值。

强化学习是智能体以试错的方式进行学习，通过与环境进行交互获得奖赏来不断指导自身行为，且环境产生的强化信号可以对动作好坏进行评价。如图 2 所示，智能体在某一时刻 t 观测到环境反馈的状态 $s_t \in S$，依据策略 $\pi(s_t, a_t)$ 选取动作 $a_t \in A$，完成动作后获得实时奖励值 r_t。动作 a_t 同时会改变环境，使得状态从 s_t 转移到 s_{t+1}。智能体根据 $t+1$ 时刻的状态 s_{t+1} 选择下一个动作，进入下一时间节点的迭代。

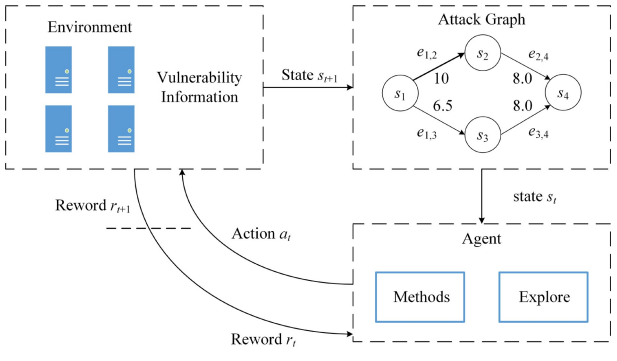


图 2　强化学习交互过程

Fig. 2　Reinforcement learning interaction process

3.2　基于最大熵的 Soft Q-learning 方法

传统强化学习中智能体的目标是最大化累计折扣回报值，智能体在单个场景督促下只能学到单一的策略，对于多任务最优或次优行为的场景(例如渗透路径场景)而言，智能体倾向于学习到一个较为随机的策略，并且可以探索到更多评价良好的行为。因此，在连续动作空间中用近似推理进行最大熵策略学习[21-22]是解决此类问题的可行方案。

标准的强化学习目标是学习到一个最大化期望收益的最优策略：

$$\pi_{\text{std}}^*=\arg\max_{\pi}\sum_t E_{(s_t,a_t)\sim\rho_\pi}[\gamma^t r(s_t,a_t)] \tag{8}$$

其中，ρ_π 表示状态和动作在轨迹上的分布[23-24]。在式(8)的基础上，最大熵强化学习在已有奖励项上增加一个熵值选项，使得最优策略同时最大化累积奖励和熵值的期望：

$$\pi_{\text{MaxEnt}}^*=\arg\max_{\pi}\sum_t E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)+\alpha H(\pi(\cdot|s_t))] \tag{9}$$

其中，$H(\pi(\cdot|s_t))$ 表示状态 s_t 下策略分布的熵[25]，且策略 $\pi(\cdot|s_t)$ 越随机，熵值越大。温度因子 α 用于平衡奖励和熵，若 α 较大，表明策略随机性强，若 $\alpha=0$，则策略趋近于确定性。

从式(9)出发，定义 Soft Q-learning 的学习算法为

$$Q_{\text{soft}}(s_t,a_t)\leftarrow r_t+\gamma E_{s_{t+1}}[V_{\text{soft}}(s_{t+1})], \forall s_t, a_t \tag{10}$$

$$V_{\text{soft}}(s_t)\leftarrow\alpha\log\int_A\exp\left(\frac{1}{\alpha}Q_{\text{soft}}(s_t,a')\right)\mathrm{d}a', \forall s_t \tag{11}$$

在这些规则下，Q_{soft} 和 V_{soft} 最终会收敛到 Q_{soft}^* 和 V_{soft}^*。

基于式(10)和式(11)，可以得到更新后的最大熵策略：

$$\pi_{\text{MaxEnt}}^*(a_t,s_t)=\exp\left(\frac{1}{\alpha}(Q_{\text{soft}}^*(s_t,a_t)-V_{\text{soft}}^*(s_t))\right) \tag{12}$$

该策略可以广泛分布在连续的动作空间中，在最大熵的框架下更具有随机性，从而提升智能体的探索率。

所提 Soft Q-learning 方法提供了一种隐式的探索策略，它鼓励智能体在受到干扰时探索和开发大部分的状态空间，以充分的探索来捕获多种模式的近似最优行为，降低算法对模型与估计误差的敏感性。因此，面对复杂多变的网络环境，该方法控制熵值和奖励的重要程度可以保证求解最优渗透路径的过程具有稳定性。

3.3　动态环境下的渗透路径探索

传统基于攻击图的最优渗透路径生成方法主要存在两个问题。一方面，基于当前网络拓扑环境构建的生成的攻击路径无法适应网络环境参数的改变，需要重新构建攻击图来应对，灵活性差。另一方面，通过计算转移概率，分析不同攻击路径的概率分布和期望的方法响应速度较慢，无法适应大规模网络场景。

考虑到上述问题，智能体在探索学习的过程中，应该在环境因素变化的干扰下选取未来收益尽可能高的动作。以图 3 为例，s_1 和 s_{10} 分别作为初始状态节点和目标状态节点，有向边上标注的分值为箭头所指状态节点的漏洞评分。初始环境下智能体已经模拟攻击者的行为从状态节点 s_1 转移到状态节点 s_2，得到部分路径序列 $s_1\rightarrow s_2$。在继续探索的过程中，以下两种情况会导致攻击图中的参数发生动态变化。

1)漏洞评分改变。一些节点的前驱节点被渗透后,自身组件脆弱性增加导致 CVSS 分值升高;另外,一些节点部署了入侵检测系统,可以预警网络攻击,及时修复自身的脆弱性组件,因此 CVSS 分值降低。如图 4 所示,当 s_5 已被攻击者成功渗透且获得主机权限后,因为 s_{10} 与 s_5 之间约定了协作防御,s_{10} 能被及时预警并修复自身脆弱性组件,其关联的有向边上 CVSS 分值由 7.2 分降为 2.3 分,即渗透成功的难度增大;s_7 与 s_5 之间未约定协作防御,则 s_5 自身脆弱性增加而导致关联的多跳范围内有向边上的漏洞分值均有不同程度的升高,即渗透成功难度减小。

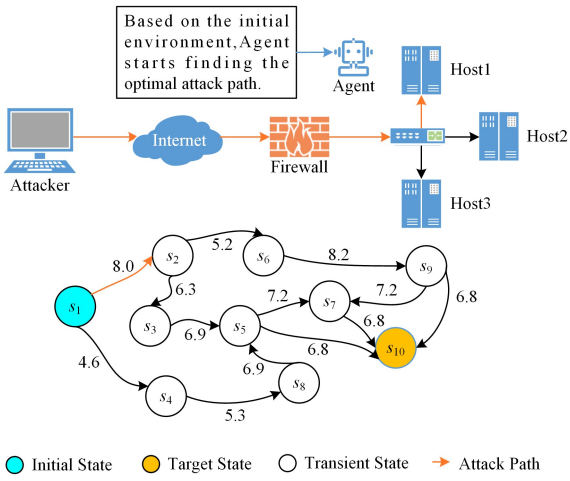


图 3 初始阶段攻击图中的部分渗透路径

Fig. 3 Partial penetration paths in initialized attack graph

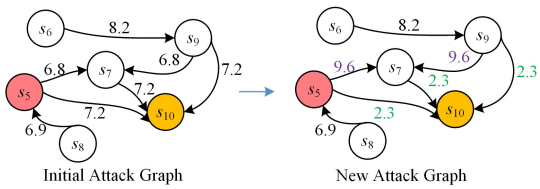


图 4 漏洞评分变化

Fig. 4 Vulnerability score changes

2)状态节点有向边改变。实际网络场景下主机之间设置了信息交互机制,即它们按照约定时间进行基于密钥协议的信息交互。当某个主机被渗透成功时,它可能无法与相邻主机进行信息交互,相邻主机会立刻预警且暂时关闭与此主机之间的通信服务。如图 5 所示,当 s_2 对应的主机 Host1 的权限被成功渗透时,因为主机 Host1 与 s_4 对应的主机 Host3 之间部署了基于密钥协议的信息交互机制,此时主机 Host3 无法与主机 Host1 进行信息交互。主机 Host3 立即预警并暂时关闭与主机 Host1 之间的链接服务($s_2 \rightarrow s_4$ 有向边),使得攻击者无法将主机 Host1 作为跳板主机对主机 Host3 发起攻击。

综上,图 4 和图 5 所示的两种网络环境的变化都会使得攻击图的参数动态调整,更接近于真实网络环境。同时,攻击图动态更新的过程作为连续的状态空间能够不断反馈给智能体,在最大熵策略驱动下,智能体能够及时依据环境的反馈选取动作,直到形成稳定策略。

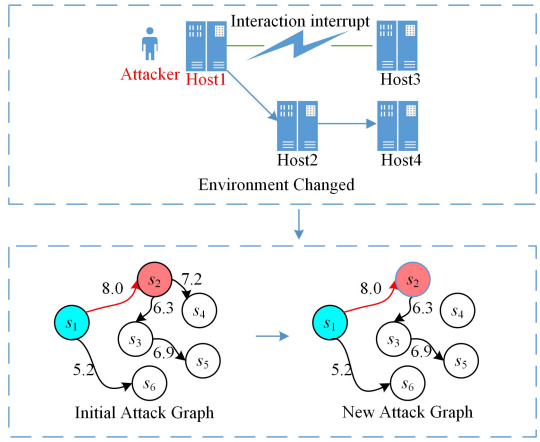


图 5 主机信息交互变化

Fig. 5 Host information interaction changes

4 仿真实验和结果分析

4.1 实验环境

为了验证本文方法的有效性,搭建一个实际网络环境来测试渗透路径生成。实验网络拓扑如图 6 所示,包括外部和内部攻击者、主机、防火墙、入侵检测系统(IDS)。防火墙 1 和防火墙 2、防火墙 3 之间形成 DMZZone,并分布着 5 台主机 H1－H5;另外的 20 台主机 H6－H20 分布在 TrustedZone。防火墙 1 禁止外网用户访问 TrustedZone,但完成对外地址到服务器实地地址的转换后可以访问 DMZZone。防火墙 2 和防火墙 3 允许 DMZZone 和 TrustedZone 的主机互相通信,但 TrustedZone 的主机仅能被动接受请求。

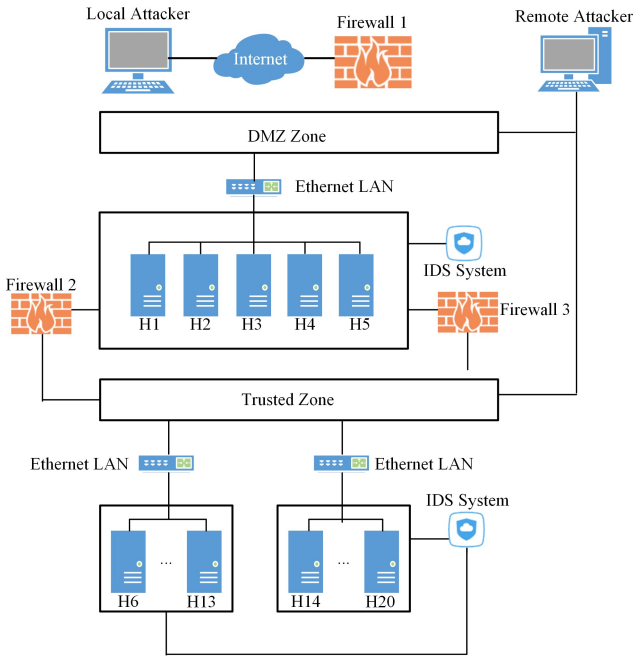


图 6 实验网络拓扑图

Fig. 6 Experimental network topology

本文依据公开漏洞数据库 NVD 为 25 台主机设置漏洞信息,表 2 列出了部分主机 H1,H3,H11,H13 的漏洞信息示例,其中包括漏洞标号、漏洞评分、漏洞信息描述以及发行日期。

表 2　主机漏洞信息

Table 2　Host vulnerability information

Host	CVEID	Score	CVE Description	Published Date
H1	CVE-2021-34631	8.8	Plugin is vulnerable to cross-site	2021-08-05
H3	CVE-2021-34635	6.1	Plugin is vulnerable to reflected cross-site scripting	2021-08-02
H11	CVE-2020-2569	3.9	Vulnerability in the oracle DBA component	2020-01-15
H13	CVE-2020-2687	4.3	Allows unauthenticated attacker with network	2020-01-15

根据实验网络拓扑图和采集的主机漏洞信息，利用自动生成工具 MulVAL 生成攻击图，如图 7 所示。攻击图中包含 23 种不同的状态节点，状态转移的有向边上标注漏洞评分。攻击图为有向图，可以用邻接矩阵存储攻击图中的信息。攻击图中的状态数作为邻接矩阵的维度，攻击图中边上的分值作为邻接矩阵中的对应元素值并记流向节点自身的状态转移漏洞评分 $Score(e_{i,i})=10$。

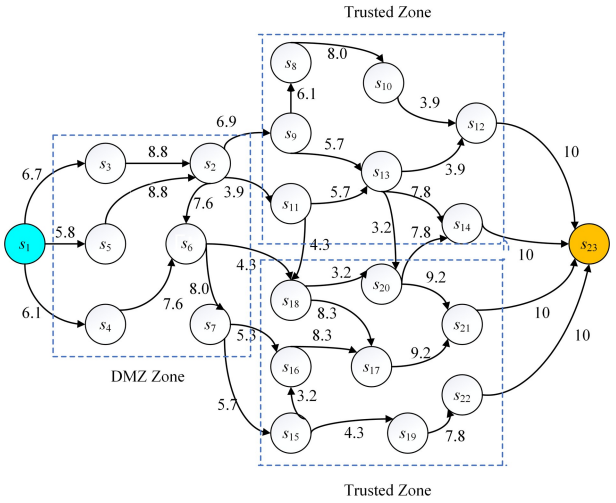


图 7　根据实验场景抽象出的攻击图

Fig. 7　Attack graph abstracted from experimental scenarios

4.2　训练和测试

根据攻击图映射的邻接矩阵，智能体通过不断探索来学习攻击者的行为策略，并设定每一轮探索的中止条件可为以下 3 种情况：1）成功渗透到目标主机并获取权限；2）训练步数达到设定值；3）状态不停转移到自身导致训练陷入死循环。

为了找到攻击者发起渗透攻击的最大可能路径，智能体需要通过不断试错学习来得到最优策略。初始状态时，因未经过预先学习，智能体倾向于使用随机策略来选择所需执行的动作。因此，智能体首先进行第一轮学习，并且找到到达目标状态 s_{23} 的全局最优渗透路径 $P^1_{S,D}$（$s_1\rightarrow s_3\rightarrow s_2\rightarrow s_9\rightarrow s_{13}\rightarrow s_{14}\rightarrow s_{23}$），并记初始环境下最优渗透路径长度 $L^1_{S,D}=6$。

第二轮学习开始后，在学习过程中，智能体的防御系统对异常或攻击行为的响应使得图 7 中一些有向边的赋值或连通性发生了变化，如表 3 所列。智能体需要在更新后的攻击图给出的反馈下继续更新最优渗透路径，模拟攻击者发动网络攻击获取攻击收益最大化的过程。

表 3　攻击图的动态变化列表

Table 3　Dynamically changing list of attack graphs

变化方式	变化结果
CVSS 评分	$s_8\rightarrow s_9$ 的有向边分值降为 2.3
CVSS 评分	$s_{17}\rightarrow s_{21}$ 的有向边分值降为 3.1
CVSS 评分	$s_4\rightarrow s_6$ 的有向边分值升为 8.2
CVSS 评分	$s_9\rightarrow s_{13}$ 的有向边分值降为 4.3
攻击图有向边	删除节点 $s_{17}\rightarrow s_{21}$ 的有向边
攻击图有向边	删除节点 $s_1\rightarrow s_5$ 的有向边

智能体在第 k 次（$k>1$）学习过程中已经探索得到的部分渗透路径长度为 $L^k_{S,X}$。智能体选择动作执行时获得的实时奖励值有以下两种计算方式：

1）当 $L^k_{S,X}\leqslant L^1_{S,D}$ 时，鼓励智能体倾向于使用随机策略进行路径探索学习，每渗透成功一个节点将获得当前节点流入边的漏洞评分 $Score(e_{i,j})$，即：

$$Reward=Score(e_{i,j}) \tag{13}$$

2）当 $L^k_{S,X}>L^1_{S,D}$ 时，不再鼓励智能体继续使用随机策略进行路径探索学习，而是在每个动作结束后对智能体施加惩罚（即计算攻击代价 $C(z_{i,j})$），并催促它尽快到达目标节点。此时，$Reward$ 被表示为：

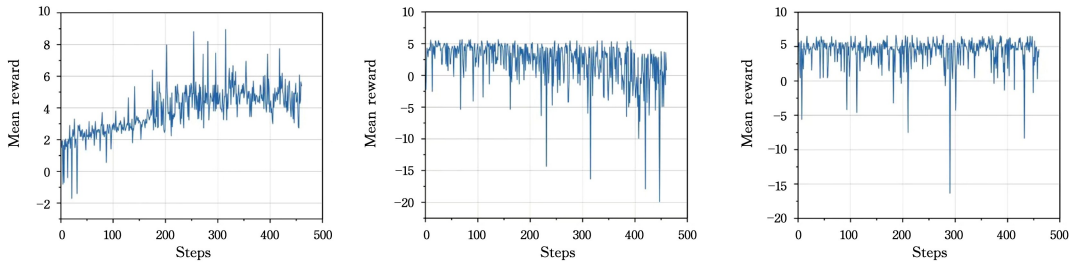
$$Reward=Score(e_{i,j})-Score\ (e_{i,j})^v \tag{14}$$

4.3　结果分析

以第二轮学习结果为例，智能体顺利到达目标节点 s_{23}，得到局部环境改变后的最优渗透路径 $P^2_{S,D}$：$s_1\rightarrow s_4\rightarrow s_6\rightarrow s_7\rightarrow s_{15}\rightarrow s_{19}\rightarrow s_{22}\rightarrow s_{23}$。对比第一轮的最优渗透路径 $P^1_{S,D}$ 发现：1）智能体能感知到攻击图的参数改变并做出相应的动作，探索得到的新路径适应了环境的改变；2）第二轮的路径长度 $L^2_{S,D}>L^1_{S,D}$，这是因为奖惩机制参考了首轮得到的 $L^1_{S,D}$，在环境参数改变的情况下，部分节点不易渗透，但考虑到攻击代价，攻击者必须尽快到达目的节点。智能体能够不断适应局部训练环境的改变而不断更新渗透路径，及时对攻击行为进行防护。

首先，考察整个训练过程中迭代次数对 $Reward$ 的影响。由图 8(a)可知，$Reward$ 呈现缓慢上升的趋势，直到最后趋于收敛。因为设置密集的奖惩机制、引入经验学习机制可以促使智能体尽快学习到最佳策略，故本文方法可以适应网络环境的动态改变。其次，传统的强化学习受最优性理论的影响，将最优策略在完全观测状态下看作确定性策略，但确定性的最优策略在具有最优控制和动作选择不确定的应用场景（本文的应用场景）下不适用。因此，针对本文应用场景，在相同的奖励值条件下，设计使用传统 Q-learning 与 Sarsa 的方法寻找最优渗透路径的对比实验，得到两种方法的平均奖励值，如图 8(b)和图 8(c)所示。两种方法平均奖励值抖动较大，最低值甚至接近 −20，模型的不确定性较大，得到的渗透路径也不一定就是最优渗透路径。

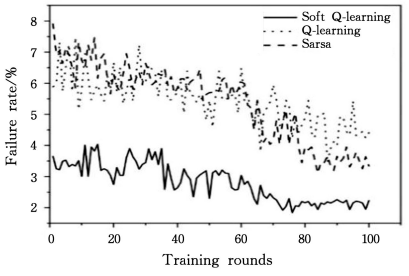
接着，通过改变训练步长和网络节点数量来计算任务失败率，比较 3 种方法在生成最优渗透路径上的效率。如图 9 所示，3 种方法的失败率均随着网络节点数量的增加呈现上升趋势，随着训练轮次的增加而呈现下降趋势，但 Soft Q-learning 方法的失败率整体低于其他两种方法。



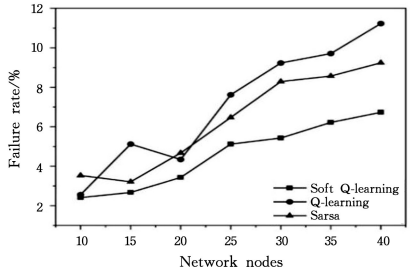
(a) Soft Q-learning (b) Q-learning (c) Sarsa

图 8 不同方法的奖励值

Fig. 8 Rewards of different methods



(a)



(b)

图 9 训练轮次和网络节点数量对失败率的影响

Fig. 9 Effect of network size and training epochs on failure rate

最大熵强化学习模型中的熵值反映系统的可靠性和稳定性，且熵值越小，策略越稳定。图 10 给出了两种不同规模网络节点个数下，模型熵值的变化情况。智能体的数量随着训练回合数的增加也增加，Soft Q-learning 方法的熵值逐渐下降，这是因为在训练开始时鼓励智能体进行探索，且节点个数较多的情况下，智能体探索性更强，熵值较大；当训练累积一段时间后，智能体学习到的策略趋于稳定，熵值逐渐减小。

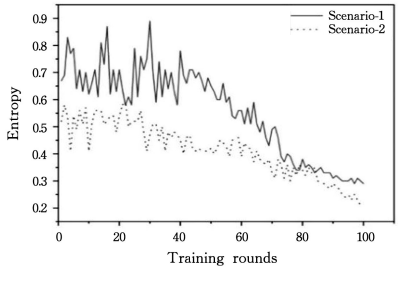


图 10 熵值变化图

Fig. 10 Entropy change graph

攻击获益可以衡量渗透路径被利用的可能性，攻击获益越大，证明该条渗透路径越有可能被攻击者利用。图 11 给出了在不同渗透路径长度下，3 种方法对应的攻击获益值，在路径长度分别为 5 和 9 时，可以看到通过 Q-learning 方法得出的攻击获益为负值，说明该方法得到的渗透路径质量不高，不是基于当前网络环境下的最优渗透路径。除此之外，Soft Q-learning 方法在不同路径长度下得到的攻击获益值明显高于两种对照方法。

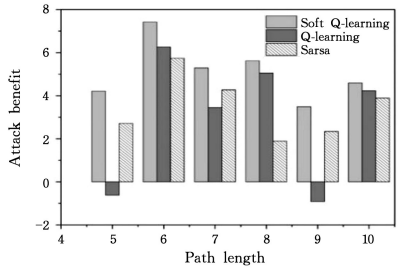


图 11 攻击获益

Fig. 11 Attack benefit

最后，在静态网络环境（即网络参数不发生改变）和动态网络环境（如 3.3 节所述）下，使用不同方法预测攻击者对目标节点所选择攻击路径的可能情况。图 12 给出了两种网络环境下对目标节点预测攻击路径的总体成功概率。通过对比发现，Soft Q-learning 方法预测的路径成功概率明显高于另外两种方法。这是因为 Soft Q-learning 方法中，智能体可以充分面对环境的干扰来选择未来收益尽可能高的动作，在最大熵策略驱动下，不断学习直到形成稳定的策略。

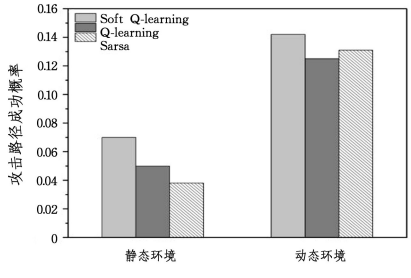


图 12 攻击路径成功概率比较

Fig. 12 Comparison of attack path success probabilities

结束语 渗透攻击给网络安全带来了巨大威胁，最优渗透路径生成能够反映攻击者入侵目标网络的最大可能攻击过程，为分析攻击行为及设计防御策略提供了重要依据。本文提出了一种基于最大熵强化学习的最优渗透路径生成方法，在动态攻击图上智能、高效地探索出攻击路径。实验结果表明，智能体使用最大熵模型训练出的学习策略可以提高渗透

路径的生成效率。下一步工作尝试建立网络态势的动态模型，对更加复杂的网络场景进行最优渗透路径分析，进而制定高效的安全防御策略。

参 考 文 献

[1] HOU J，JIA X. Research on enterprise network security system [C]// 2021 2nd International Conference on Computer Science and Management Technology(ICCSMT). IEEE，2021：216-219.

[2] HU H，LIU Y，ZHANG H. Route prediction method for network intrusion using absorbing Markov chain [J]. Journal of Computer Research and Development，2018，55(4)：831-845.

[3] MA Y，WU Y，YU D，et al. Vulnerability association evaluation of Internet of thing devices based on attack graph [J]. International Journal of Distributed Sensor Networks，2022，18(5)：1-10.

[4] HOU S，CHEN X，MA J，et al. An ontology-based dynamic attack graph generation approach for the internet of vehicles [J]. Frontiers in Energy Research，2022，10：1-12.

[5] SCHIELE N D，GADYATSKAYA O. A novel approach for attack tree to attack graph transformation [C] // International Conference on Risks and Security of Internet and Systems. Cham：Springer，2022：74-90.

[6] KAYNAR K. A taxonomy for attack graph generation and usage in network security[J]. Journal of Information Security and Applications，2016，29：27-56.

[7] SUN F，PI J，LV J，et al. Network security risk assessment system based on attack graph and Markov chain [C]// Journal of Physics：Conference Series，The 2017 International Conference on Cloud Technology and Communication Engineering (CTCE2017). Guilin，China，2017：1-10.

[8] WANG S，WANG J H，TANG G M，et al. Intelligent and efficient method for optimal penetration path generation[J]. Journal of Computer Research and Development，2019，56(5)：929-941.

[9] WANG S，TANG G，KOU G. Attack path prediction method based on causal knowledge net[J]. Journal on Communications，2016，37(10)：188-198.

[10] AL-ARAJI Z，SYED A S S，ABDULLAH R S. Attack prediction to enhance attack path discovery using improved attack graph [J]. Karbala International Journal of Modern Science，2022，8(3)：313-329.

[11] STERGIOPOULOS G，DEDOUSIS P，GRITZALIS D. Automatic analysis of attack graphs for risk mitigation and prioritization on large-scale and complex networks in Industry 4. 0[J]. International Journal of Information Security，2022，21(1)：37-59.

[12] BOUDERMINE A，KHATOUN R，CHOYER J H. Attack graph-based solution for vulnerabilities impact assessment in dynamic environment[C]// 2022 5th Conference on Cloud and Internet of Things(CIoT). IEEE，2022：24-31.

[13] ZHOU S C，LIU J J，ZHONG X F，et al. Intelligent penetration testing path discovery based on deep reinforcement learning[J]. Computer Science，2021，48(7)：40-46.

[14] WANG B，LIU Z，LI Q，et al. Mobile robot path planning in dynamic environments through globally guided reinforcement learning[J]. IEEE Roboticsand Automation Letters，2020，5(4)：6932-6939.

[15] KOO K，MOON D，HUH J H，et al. Attack graph generation with machine learning for network security [J]. Electronics，2022，11(9)：1-25.

[16] CODY T，RAHMAN A，REDINO C，et al. Discovering exfiltration paths using reinforcement learning with attack graphs[J]. arXiv：2201. 12416，2022.

[17] ZHANG L，BAI W，LI W，et al. Discover the hidden attack path in multi-domain cyberspace based on reinforcement learning[J]. arXiv：2104. 07195，2021.

[18] HOUMB S H，FRANQUEIRA V N，ENGUM E A. Quantifying security risk level from CVSS estimates of frequency and impact [J]. The Journal of Systems & Software，2009，83(9)：1622-1634.

[19] ATEFEH K，MOHAMMAD G，VALI D. An automatic method for CVSS score prediction using vulnerabilities description[J]. Journal of Intelligent & Fuzzy Systems，2015，30(1)：89-96.

[20] CHEN C L，CHEN J M. Use of markov chain for early detecting DDOS attacks[J]. International Journal of Network Security & Its Applications(IJNSA)，2021，13(4)：1-11.

[21] SHARMA K，SINGH B，HERMAN E，et al. Maximum information measure policies in reinforcement learning with deep energy-based model[C]// 2021 International Conference on Computational Intelligence and Knowledge Economy(ICCIKE). IEEE，2021：19-24.

[22] ZHANG T，LI Y，WANG C，et al. Fop：Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning [C]// International Conference on Machine Learning. PMLR，2021：12491-12500.

[23] SCHULMAN J，CHEN X，ABBEEL P. Equivalence between policy gradients and soft q-learning [J]. arXiv：1704. 06440，2017.

[24] GRAU-MOYA J，LEIBFRIED F，BOU-AMMAR H. Balancing two-player stochastic games with soft q-learning [J]. arXiv：1802. 03216，2018.

[25] GARG D，CHAKRABORTY S，CUNDY C，et al. IQ-Learn：Inverse soft-Q Learning for Imitation[J]. Advances in Neural Information Processing Systems，2021，34：4028-4039.

WANG Yan，born in 1999，postgraduate. His main research interest is network security based on AI.

SHEN Hang，born in 1984，Ph. D，associate professor，is a senior member of CCF(No. 19088S). His main research interests include network security and privacy computing.

（责任编辑：喻藜）