(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110166428 A
(43)申请公布日 2019.08.23

(21)申请号 201910292304.2

(22)申请日 2019.04.12

(71)申请人 中国人民解放军战略支援部队信息工程大学
地址 450000 河南省郑州市高新区科学大道62号

(72)发明人 胡浩 张玉臣 杨峻楠 谢鹏程 刘玉岭 马博文 冷强 张畅 陈周文 林野

(74)专利代理机构 郑州大通专利商标代理有限公司 41111
代理人 周艳巧

(51)Int.Cl.
*H04L 29/06*(2006.01)
*H04L 12/24*(2006.01)

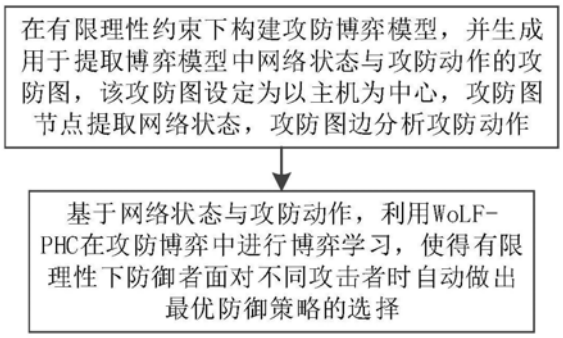
权利要求书2页 说明书10页 附图5页

(54)发明名称

基于强化学习和攻防博弈的智能防御决策方法及装置

(57)摘要

本发明属于网络安全技术领域，特别涉及一种基于强化学习和攻防博弈的智能防御决策方法及装置，该方法包含：在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；防御者在网络状态转移概率未知时，通过在线学习得到防御收益使得防御者面对不同攻击者时自动做出最优防御策略的选择。本发明有效压缩博弈状态空间，降低了存储和运行开销；防御者在与攻击者对抗中依据环境反馈进行强化学习，在面对不同攻击时能自适应做出最优选择；提升防御者学习速度，提高了防御收益，减少对历史数据依赖，有效提升防御者决策时的实时性和智能性。




CN 110166428 A

1.一种基于强化学习和攻防博弈的智能防御决策方法,其特征在于,包含如下内容:

A)在有限理性约束下构建攻防博弈模型,并生成用于提取博弈模型中网络状态与攻防动作的攻防图,该攻防图设定为以主机为中心,攻防图节点提取网络状态,攻防图边分析攻防动作;

B)基于网络状态与攻防动作,结合攻防博弈模型,对攻防博弈过程进行强化学习,攻防双方对抗中依据环境反馈,使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择。

2.根据权利要求1所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,A)中,攻防博弈模型用六元组表示,即AD-SGM＝(N,S,D,R,Q,π),其中,N表示参与博弈的局中人,S表示随机博弈状态集合,D表示防御者动作集合,R表示防御者立即回报,Q表示防御者状态—动作收益函数,π表示防御者防御策略。

3.根据权利要求1所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,攻防图用二元组表示,即G＝(S,E),其中,S表示网络节点安全状态集合,E表示攻击动作或防御动作发生引起节点状态的转移。

4.根据权利要求3所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,生成攻击图时,首先对目标网络扫描获取网络安全要素,然后与攻击模板结合进行攻击实例化,与防御模板结合进行防御实例化,最后生成攻防图,其中,攻防博弈模型的状态集合由攻防图节点提取,防御动作集合由攻防图边提取。

5.根据权利要求1所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,B)中,强化学习中,采用狼爬山策略WoLF-PHC免模型强化学习机制,通过与环境交互获取回报和环境状态转移知识,知识利用收益表示,设定防御者高低策略学习率以适应不同攻击者策略,收益更新过程利用强化学习机制,确定防御者最优防御策略。

6.根据权利要求5所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,收益表示为$Q_d(s,d)=Q_d(s,d)+\alpha[R_d(s,d,s')+\gamma\max_{d'}Q_d(s',d')-Q_d(s,d)]$,强化学习的策略为:$\pi_d(s)=\arg\max_d Q_d(s,d)$,其中,α为收益学习率;γ为折扣因子,$R_d(s,d,s')$表示防御者在状态s执行防御动作d网络转移到状态s'后的立即回报。

7.根据权利要求6所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,采用平均策略作为胜利和失败的评判标准,公式表示为:

$$\bar{\pi}_d(s,d)=\bar{\pi}_d(s,d)+\frac{1}{C(s)}(\pi_d(s,d)-\bar{\pi}_d(s,d)),\quad C(s)=C(s)+1。$$

8.根据权利要求6所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,免模型强化学习机制中,引入用于跟踪最近访问的状态-动作轨迹的资格迹,将当前回报分配给最近访问的状态-动作,利用资格迹对收益进行更新。

9.根据权利要求8所述的基于强化学习和攻防博弈的智能防御决策方法,其特征在于,强化学习中,定义每个状态-动作的资格迹为e(s,a),设当前网络状态为s*,资格迹以$e(s,d)=\begin{cases}\gamma\lambda e(s,d) & s\neq s^*\\ \gamma\lambda e(s,d)+1 & s=s^*\end{cases}$方式进行更新,将当前回报分配给最近访问的状态-动作,其

中，γ为折扣因子，λ为轨迹衰减因子。

10.一种基于强化学习和攻防博弈的智能防御决策装置，其特征在于，包含：

攻防图生成模块，用于在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；

防御策略选取模块，基于网络状态与攻防动作，结合攻防博弈模型，对攻防博弈过程进行强化学习，攻防双方对抗中依据环境反馈，使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择。

# 基于强化学习和攻防博弈的智能防御决策方法及装置

## 技术领域

[0001] 本发明属于网络安全技术领域，特别涉及一种基于强化学习和攻防博弈的智能防御决策方法及装置。

## 背景技术

[0002] 近年来，信息安全事件日趋频繁，给网络安全带来了巨大的损失，据统计，阿里云在2017年仅每天就要遭受16亿次左右的攻击，对于不同攻击者，可能每个攻防场景只会出现一次，但对于以阿里云为代表的防御者来说，其每天都要面对大量相同的攻防场景。考虑到网络设备硬件资源有限，如何综合考虑防御成本和收益，以防御收益最大化为目标，使防御者在风险与投入之间达成一种均衡，如何使防御者在大量相同的攻防场景中对收益进行在线学习和更新，安全管理员面临适度安全条件下“最优策略难以选取”的困境。博弈论与网络攻防所具有的目标对立性、关系非合作性和策略依存性高度契合。目前基于博弈论的防御决策方法可以分为基于完全理性假设和有限理性假设两类：一是基于攻防参与者完全理性的防御决策方法。完全理性假设的前提是每个参与者既能理智选择最优策略使自己利益最大化，同时能预测其他参与者的策略选择。应用到无线传感器安全领域，通过建立攻击者与传感器信任节点间的非合作博弈模型，依据纳什均衡给出最优攻击策略，可以对蠕虫病毒攻击和防御策略的效能进行分析。通过建立入侵检测系统和无线传感器节点间的重复博弈模型，可以分析节点包的转发策略。二是基于攻防参与者有限理性的防御决策方法。有限理性意味着攻防双方不会在一开始就找到最优策略，会在攻防博弈中学习攻防博弈，合适的学习机制是在博弈中取胜的关键。该类方法主要围绕演化博弈展开，演化博弈以群体为研究对象，采用生物进化机制，通过模仿其它成员的优势策略来完成学习。演化博弈中参与人之间信息交换过多且主要是对攻防群体策略的调整过程、趋势和稳定性进行研究，不利于指导个体成员的实时策略选择。如何采取更好的学习机制模拟攻防过程，提高防御决策的准确性和时效性成为亟待解决的技术问题。

## 发明内容

[0003] 为此，本发明提供一种基于强化学习和攻防博弈的智能防御决策方法及装置，适用于现实攻防网络环境，实现在线学习能力的智能化防御决策，具有较强的实用性和可操作性。

[0004] 按照本发明所提供的设计方案，一种基于强化学习和攻防博弈的智能防御决策方法，包含如下内容：

[0005] A)在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；

[0006] B)基于网络状态与攻防动作，依托攻防博弈模型，对攻防博弈过程进行强化学习，攻防双方对抗中依据系统反馈，使得有限理性下防御者面对不同攻击者时自动做出最优防

御策略的选择。

[0007] 上述的，A)中，攻防博弈模型用六元组表示，即AD-SGM＝(N,S,D,R,Q,π)，其中，N表示参与博弈的局中人，S表示随机博弈状态集合，D表示防御者动作集合，R表示防御者立即回报，Q表示防御者状态—动作收益函数，π表示防御者防御策略。

[0008] 上述的，攻防图用二元组表示，即G＝(S,E)，其中，S表示节点安全状态集合，E表示攻击动作或防御动作发生引起节点状态的转移。

[0009] 优选的，生成攻击图时，首先对目标网络扫描获取网络安全要素，然后与攻击模板结合进行攻击实例化，与防御模板结合进行防御实例化，最后生成攻防图，其中，攻防博弈模型的状态集合由攻防图节点提取，防御动作集合由攻防图边提取。

[0010] 上述的，B)中，强化学习中，采用狼爬山策略WoLF-PHC免模型强化学习机制，通过与环境交互获取回报和环境状态转移知识，知识利用收益表示，设定防御者高低策略学习率以适应攻击者策略，通过更新收益进行强化学习，确定防御者最优防御策略。

[0011] 优选的，收益表示为 $Q_d(s,d)=Q_d(s,d)+\alpha[R_d(s,d,s')+\gamma\max_{d'}Q_d(s',d')-Q_d(s,d)]$，强化学习的策略为：$\pi_d(s)=\arg\max_d Q_d(s,d)$，其中，α为收益学习率；γ为折扣因子，Rd(s,d,s')表示防御者在状态s执行防御动作d网络转移到状态s'后的立即回报。

[0012] 更进一步，采用平均策略作为胜利和失败的评判标准，公式表示为：

[0013] $$\bar{\pi}_d(s,d)=\bar{\pi}_d(s,d)+\frac{1}{C(s)}(\pi_d(s,d)-\bar{\pi}_d(s,d)),\quad C(s)=C(s)+1。$$

[0014] 更进一步，免模型强化学习机制中，引入用于跟踪最近访问的状态-动作轨迹的资格迹，将当前回报分配给最近访问的状态-动作，利用资格迹对收益进行更新。

[0015] 更进一步，强化学习中，定义每个状态—动作的资格迹为e(s,a)，设当前网络状态为s*，资格迹以 $e(s,d)=\begin{cases}\gamma\lambda e(s,d) & s\neq s^*\\ \gamma\lambda e(s,d)+1 & s=s^*\end{cases}$ 方式进行更新，将当前回报分配给最近访问的状态-动作，其中，γ为折扣因子，λ为轨迹衰减因子。

[0016] 更进一步，一种基于强化学习和攻防博弈的智能防御决策装置，包含：

[0017] 攻防图生成模块，用于在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；

[0018] 防御策略选取模块，基于网络状态与攻防动作，结合攻防博弈模型，对攻防博弈过程进行强化学习，攻防双方对抗中依据环境反馈，使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择。

[0019] 本发明的有益效果：

[0020] 本发明中以主机为中心的攻防图模型用于网络状态及攻防动作，有效压缩博弈状态空间；防御者采用强化学习机制，在与攻击者对抗中依据环境的反馈进行学习，使得有限理性下的防御者在面对不同攻击者时都能自动做出最优选择；在决策装置中加入资格迹，提升了防御者的学习速度，减少了对历史数据的依赖，有效提升防御者决策时的实时性和智能性。

**附图说明：**

[0021] 图1为实施例中智能防御决策流程示意图；

[0022] 图2为实施例中攻防状态转移示意图；

[0023] 图3为实施例中强化学习机制原理图；

[0024] 图4为实施例中实验网络结构；

[0025] 图5为实施例中网络脆弱性信息示意图；

[0026] 图6为实施例中攻击动作图；

[0027] 图7为实施例中防御动作图；

[0028] 图8为实施例中防御动作描述；

[0029] 图9为实施例中实验设置参数图；

[0030] 图10为实施例中防御决策态势图；

[0031] 图11为实施例中防御收益态势图。

**具体实施方式：**

[0032] 为使本发明的目的、技术方案和优点更加清楚、明白，下面结合附图和技术方案对本发明作进一步详细的说明。实施例中涉及到的技术术语如下：

[0033] 强化学习是一种经典的在线学习方法，其参与人通过环境的反馈进行独立学习，相比生物进化型学习方式，学习速度快，符合攻防转换快，时效性强的特点。博弈的非合作性、目标对立性和策略依存性等特点均符合网络攻防的基本特征。本发明实施例，参见图1所示，提供一种基于强化学习和攻防博弈的智能防御决策方法，包含：

[0034] 在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；

[0035] 基于网络状态与攻防动作对攻防博弈模型进行强化学习，攻防双方对抗中依据系统反馈，使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择。

[0036] 基于属性攻击图的动态威胁跟踪分析，在攻击路径推断、威胁转移概率、前后件推断、消解环路、实时分析、综合多路径、权限提升和存取访问关系等方面具有明显优势。

[0037] 将强化学习机制引入到攻防博弈中，在有限理性约束下构建攻防博弈模型，并生成以主机为中心的攻防图，用于提取博弈模型中的网络状态与攻防动作；通过强化学习实现在线实时自动化的防御决策。

[0038] 网络攻防博弈模型采用概率值描述网络状态转移的随机性，由于当前网络状态主要与前一个网络状态有关，采用一阶马尔可夫来表示状态转移关系，如图2所示，转移概率为$P(s_t,a_t,d_t,s_{t+1})$，其中，s为网络状态，(a,d)为攻防动作。由于网络攻防双方具有目标对立性和非合作性，攻防双方会刻意隐藏自己的关键信息，转移概率设定为攻防双方的未知信息。在此基础上，构建博弈模型。本发明另一个实施例中，攻防随机博弈模型(attack defense stochastic game model,AD-SGM)用一个六元组AD-SGM＝(N,S,D,R,Q,π)表示，其中，N＝(attacker,defender)为参与博弈的两个局中人，分别代表网络攻击者和防御者；$S=(s_1,s_2,\cdots,s_n)$为随机博弈状态集合，由网络状态组成；$D=(D_1,D_2,\cdots,D_n)$为防御者动作集合，其中$D_k=\{d_1,d_2,\cdots d_m\}$为防御者在博弈状态$S_k$的动作集合；$R_d(s_i,d,s_j)$为防御者在状态

$s_i$执行防御动作d网络转移到状态$s_j$后的立即回报;$Q_d(s_i,d)$为表示在状态$s_i$下防御者采取动作d后的期望收益;$\pi_d(s_k)$为防御者在状态$s_k$的防御策略。

[0039] 防御策略与防御动作是两个不同的概念,防御策略是防御动作的集合。防御策略以概率向量的形式规定了防御者在每个网络状态选择什么动作,如$\pi_d(s_k)=(\pi_d(s_k,d_1),\cdots,\pi_d(s_k,d_m))$为防御者在网络状态$s_k$的策略,$\pi_d(s_k,d_m))$为其选择动作$d_m$的概率,其中$\sum_{d\in D_k}\pi_d(s_k,d)=1$。

[0040] 通过创建网络攻防图G,从攻防图G的节点提取网络状态,攻防图G的边分析攻防动作,用于提取攻防策略。本发明另一个实施例中,攻防图表示为一个二元组G=(S,E),其中$S=\{s_1,s_2,\cdots,s_n\}$是节点安全状态集合,$s_i=<host,privilege>$,其中host是节点的唯一标识,privilege={none,user,root}分别表示不具有任何权限、具有普通用户权限、具有管理员权限。$E=(E_a,E_d)$为有向边,表示攻击动作或防御动作的发生引起节点状态的转移,$e_k=(s_r,v/d,s_d)$,k=a,d,其中$s_r$为源结点,$s_d$为目的结点。

[0041] 进一步地,攻防图的生成时,首先对目标网络扫描获取网络安全要素,然后与攻击模板结合进行攻击实例化,再与防御模板结合进行防御实例化,最后生成攻防图。攻防随机博弈模型的状态集合由攻防图节点提取,防御动作集合由攻防图的边提取。具体步骤可设计为如算法1所示:

[0042] 算法1.攻防图生成算法

[0043]

**输入:**网络安全要素 $NSE$，攻击模板 $AM$，防御模板 $DM$

**输出:**攻防图 $G=(S,E)$

**1)** $S \leftarrow NSE, E \leftarrow \varnothing$ /*生成所有节点*/

2) for each $S$ do:/*攻击实例化，生成攻击边*/

3) update $NSE$ in $s$ /*更新网络安全要素*/

4) if $C.shost = s.host$ and $C.dhost.V \geq AM.prec.V$ and $C.dhost.F \geq AM.prec.F$ and $C.dhost.P.privilege \geq AM.prec.P.privilege$:

5) $s_r.host \leftarrow C.shost$

6) $s_d.host \leftarrow C.dhost$

7) $s_d.privilege \leftarrow AM.postc.P.privilege$

8) $E_a \leftarrow E_a \cup \{e_a(s_r, AM.tid, s_d)\}$

9) end if

10) end for

11) for each $S$ do:/*防御实例化，生成防御边*/

12) if $E_a.s_d = s$ and $DM.tid = E_a.tid$:

13) $E_d \leftarrow E_d \cup \{e_d(E_a.s_d, DM.dset.d, E_a.s_r)\}$

14) end if

15) end for

16) for each $S$ do:/*去除 $S$ 中的孤立节点*/

17) if $e_a(s,tid,s_d) = \varnothing$ and $e_d(s_r,d,s) = \varnothing$:

18) $S \leftarrow S - s$

19) end if

20) end for

21) Return $G$

[0044] 其中，第1)步是利用网络安全要素生成所有可能状态节点并初始化边；第2)—11)步是攻击实例化，生成所有攻击边；第12)—18)步是防御实例化，生成所有防御边；第19)—23)步是去除所有孤立节点；第24)步是输出攻防图。

[0045] 本发明实施例中，将强化学习机制引入攻防博弈中，描述攻防策略的学习与改进过程。WoLF-PHC是一种典型的免模型强化学习算法，其学习机制如图3所示。本发明另一个实施例中，强化学习中Agent通过与环境的交互获得回报和环境状态转移的知识，知识用收

益$Q_d$来表示，通过更新$Q_d$来进行学习。其收益函数$Q_d$为：

[0046] $$Q_d(s,d) = Q_d(s,d) + \alpha[R_d(s,d,s') + \gamma \max_{d'} Q_d(s',d') - Q_d(s,d)] \quad (1)$$

[0047] 式(1)中，α为收益学习率；γ为折扣因子。强化学习的策略为：$\pi_d(s) = \arg\max_d Q_d(s,d)$

[0048] 进一步地，WoLF-PHC狼爬山策略通过引入WoLF机制，使防御者具有两种不同的策略学习率，当获胜时采用低策略学习率$\delta_w$，当失败时采用高策略学习率$\delta_l$，如式(5)所示。两个学习率使得防御者在比预期表现差时能快速适应攻击者的策略，比预期表现好时能谨慎学习，同时保证了算法的收敛性。WoLF-PHC算法采用平均策略作为胜利和失败的判断标准，如式(6)(7)所示。

[0049] $$\delta = \begin{cases} \delta_w & \sum_{d \in D_k} \pi_d(s_k,d) Q_d(s_k,d) > \sum_{d \in D_k} \bar{\pi}_d(s_k,d) Q_d(s_k,d) \\ \delta_l & \text{其它} \end{cases} \quad (5)$$

[0050] $$\bar{\pi}_d(s,d) = \bar{\pi}_d(s,d) + \frac{1}{C(s)}(\pi_d(s,d) - \bar{\pi}_d(s,d)) \quad (6)$$

[0051] C(s)＝C(s)+1 (7)

[0052] 为了提高WoLF-PHC算法的学习速度，减少算法对数据量的依赖程度，本发明另一个实施例中，在WoLF-PHC中引入资格迹。资格迹能跟踪最近访问的特定状态—动作轨迹，然后将当前回报分配给最近访问的状态—动作。进一步地，定义每个状态—动作的资格迹为e(s,a)，设当前网络状态为s*，资格迹以式(8)所示方式进行更新，其中λ为轨迹衰减因子。

[0053] $$e(s,d) = \begin{cases} \gamma\lambda e(s,d) & s \neq s^* \\ \gamma\lambda e(s,d) + 1 & s = s^* \end{cases} \quad (8)$$

[0054] 基于WoLF-PHC防御决策方法要想取得较好效果，对α、δ、λ和γ四个参数进行合理设置。1)收益学习率α取值范围为0＜α＜1，α越大代表越靠后的累积奖赏越重要，学习速度也更快；α越小算法的稳定性越好。2)策略学习率δ取值范围为0＜δ＜1，根据实验得出，采取$\frac{\delta_l}{\delta_w} = 4$时能够取得较好效果。3)资格迹衰减因子λ取值范围为0＜λ＜1，负责对状态—动作分配信誉，可以看作时间的标度，λ越大则分配给历史状态—动作的信誉越大。4)折扣因子γ取值范围为0＜γ＜1，代表防御者对立即回报与未来回报的偏好。当γ接近于0时，表示未来回报无关紧要，更看重立即回报；当γ接近于1时，代表立即回报无关紧要，更看重未来回报。

[0055] WoLF-PHC中的Agent，如图3所示，对应攻防随机博弈模型AD-SGM中的防御者，Agent的状态对应AD-SGM中的博弈状态，Agent的行为对应AD-SGM中的防御动作，Agent的立即回报对应AD-SGM中的立即回报，Agent的策略对应AD-SGM中的防御策略。在上述基础上，具体的防御决策算法可设计为如算法2所示：

[0056] 算法2.防御决策算法

**输入：** $AD-SGM$;参数$\alpha$、$\delta$、$\lambda$和$\gamma$

**输出：** 防御动作$d$

1) initialize $AD-SGM$ 、 $C(s)=0, e(s,d)=0$/*网络状态和攻防动作由算法 1 提取*/

2) $s^*=get(E)$/*从网络 E 中获取当前网络状态*/

3) repeat:

4) $d^*=\pi_d(s^*)$ /*选取防御动作*/

5) Output $d^*$;/*将防御动作反馈给防御者*/

[0057] 6) $s'=get(E)$/*获取执行动作$d^*$后的状态*/

7) $$\rho^*=R_d(s^*,d^*,s')+\gamma\max_{d'}Q_d(s',d')-Q_d(s^*,d^*)$$

8) $$\rho_g=R_d(s^*,d^*,s')+\gamma\max_{d'}Q_d(s',d')-\max_{d}Q_d(s^*,d)$$

9) for each state-action pair $(s,d)$ except $(s^*,d^*)$ do:

10) $e(s,d)=\gamma\lambda e(s,d)$

11) $Q_d(s,d)=Q_d(s,d)+\alpha\rho_g e(s,d)$

12) end for/*更新非当前$(s^*,d^*)$的资格迹和$Q_d$值*/

13) $Q_d(s^*,d^*)=Q_d(s^*,d^*)+\alpha\rho^*$/*更新$(s^*,d^*)$的$Q_d$值*/

14) $e(s^*,d^*)=\gamma\lambda e(s^*,d^*)+1$/*更新$(s^*,d^*)$的资格迹*/

15) $C(s^*)=C(s^*)+1$

16) 依据式(6)更新平均策略

17) 依据式(5)选择策略学习速率

[0058] 18) $$\delta_{s^*d}=\min\left(\pi_d(s^*,d),\frac{\delta}{|\mathrm{D}(s^*)-1|}\right),\forall d\in D(s^*)$$

19) $$\Delta_{s^*d}=\begin{cases}-\delta_{s^*d} & d\neq\arg\max_{d_i}Q_d(s^*,d_i)\\ \sum_{d_j\neq d^*}\delta_{s^*d_j} & \text{其它}\end{cases}$$

20) $\pi_d(s^*,d)=\pi_d(s^*,d)+\Delta_{s^*d},\forall d\in D(s^*)$/*更新防御策略*/

21) $s^*=s'$

22) end repeat

[0059] 第1)步对攻防随机博弈模型AD-SGM和相关参数的初始化，其中网络状态和攻防动作由算法1提取，第2)步防御者检测当前网络状态，第3)—22)步进行防御决策和在线学习，其中4)—5)步依据当前策略选取防御动作，第6)—14)步利用资格迹对收益$Q_d$进行更新，第15)—21)步依据新的收益$Q_d$利用爬山算法更新防御策略$\pi_d$。算法的空间复杂度主要集中在

对$R_d(s,d,s')$、$e(s,d)$、$\pi_d(s,d)$、$\bar{\pi}_d(s,d)$和$Q_d(s,d)$的存储，设$|S|$为状态数，$|D|$为每个状态防御者的措施数，则空间复杂度为$O(4\cdot|S|\cdot|D|+|S|^2\cdot|D|)$。算法不需要对博弈均衡进行求解，与现有随机博弈模型相比大大减少了计算复杂度，增强了算法的实效性。

[0060] 基于上述的智能防御决策方法，本发明实施例还提供一种基于强化学习和攻防博弈的智能防御决策装置，包含：

[0061] 攻防图生成模块，用于在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作；

[0062] 防御策略选取模块，基于网络状态与攻防动作，结合攻防博弈模型，对攻防博弈过程进行强化学习，攻防双方对抗中依据环境反馈，使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择。

[0063] 采用上述的基于强化学习和攻防博弈的智能防御决策方法进行目标网络防御策略的智能选取。

[0064] 为进一步验证本发明实施例中技术方案的有效性，通过搭建如附图4所示的典型企业网络进行实验。攻防事件发生在内网，攻击者来自外网。网络管理员作为防御者，负责内网的安全。由于防火墙1和防火墙2的设置，外网正常用户只能访问Web服务器，而Web服务器可以访问数据库服务器、FTP服务器和电子邮件服务器。利用Nessus工具对实验网络进行扫描，实验网络脆弱性信息如附图5所示。

[0065] 参考MIT林肯实验室攻防行为数据库构建攻击、防御模板，采用A标识攻击者主机、W标识Web服务器、D标识数据库服务器、F标识FTP服务器、E标识电子邮件服务器，利用攻防图生成装置构建网络攻防图，为便于展示和描述，将攻防图分为攻击图和防御图，分别如附图6和附图7所示。防御图中防御动作描述如附图8所示。构建实验场景的攻防博弈模型：

[0066] ①$N=(attacker,defender)$为参与博弈的局中人，分别代表网络攻击者和防御者；

[0067] ②随机博弈状态集合$S=(s_0,s_1,s_2,s_3,s_4,s_5,s_6)$，随机博弈状态由网络状态组成，由图5和图6中的节点提取；

[0068] ③防御者动作集合为：$D=(D_0,D_1,D_2,D_3,D_4,D_5,D_6)$，其中$D_0=\{NULL\}$ $D_1=\{d_1,d_2\}$ $D_2=\{d_3,d_4\}$ $D_3=\{d_1,d_5,d_6\}$ $D_4=\{d_1,d_5,d_6\}$ $D_5=\{d_1,d_2,d_7\}$ $D_6=\{d_3,d_4\}$，由图6的边提取；

[0069] ④防御者立即回报$R_d(s_i,d,s_j)$的量化结果为：

[0070] $(R_d(s_0,NULL,s_0),R_d(s_0,NULL,s_1),R_d(s_0,NULL,s_2))=(0,-40,-59)$

[0071] $(R_d(s_1,d_1,s_0),R_d(s_1,d_1,s_1),R_d(s_1,d_1,s_2);R_d(s_1,d_2,s_0),R_d(s_1,d_2,s_1),R_d(s_1,d_2,s_2))=(40,0,-29;5,-15,-32)$

[0072] $(R_d(s_2,d_3,s_0),R_d(s_2,d_3,s_1),R_d(s_2,d_3,s_2),R_d(s_2,d_3,s_3),R_d(s_2,d_3,s_4),R_d(s_2,d_3,s_5);R_d(s_2,d_4,s_0),R_d(s_2,d_4,s_1),R_d(s_2,d_4,s_2),R_d(s_2,d_4,s_3),R_d(s_2,d_4,s_4),R_d(s_2,d_4,s_5))=(24,9,-15,-55,-49,-65;19,5,-21,-61,-72,-68)$

[0073] $(R_d(s_3,d_1,s_2),R_d(s_3,d_1,s_3),R_d(s_3,d_1,s_6);R_d(s_3,d_5,s_2),R_d(s_3,d_5,s_3),R_d(s_3,d_5,s_6);R_d(s_3,d_6,s_2),R_d(s_3,d_6,s_3),R_d(s_3,d_6,s_6))=(21,-16,-72;15,-23,-81;-21,-36,-81)$

[0074] $(R_d(s_4,d_1,s_2),R_d(s_4,d_1,s_4),R_d(s_4,d_1,s_6);R_d(s_4,d_5,s_2),R_d(s_4,d_5,s_4),R_d(s_4,d_5,s_6);R_d(s_4,d_6,s_2),R_d(s_4,d_6,s_4),R_d(s_4,d_6,s_6))=(26,0,-62;11,-23,-75;9,-25,-87)$

[0075] $(R_d(s_5,d_1,s_2),R_d(s_5,d_1,s_5),R_d(s_5,d_1,s_6);R_d(s_5,d_2,s_2),R_d(s_5,d_2,s_5),R_d(s_5,d_2,s_6);R_d(s_5,d_7,s_2),R_d(s_5,d_7,s_5),R_d(s_5,d_7,s_6))=(29,0,-63;11,-21,-76;2,-27,-88)$

[0076] $(R_d(s_6,d_3,s_3),R_d(s_6,d_3,s_4),R_d(s_6,d_3,s_5),R_d(s_6,d_3,s_6);R_d(s_6,d_4,s_3),R_d(s_6,d_4,s_4),R_d(s_6,d_4,s_5),R_d(s_6,d_4,s_6))=(-23,-21,-19,-42;-28,-31,-24,-49)$

[0077] ⑤为了更充分的检测算法的学习性能，防御者的状态动作收益$Q_d(s_i,d)$初始化时统一置0，不引入额外的先验知识。

[0078] ⑥防御者的防御策略$\pi_d$采取平均策略进行初始化，即$\pi_d(s_k,d_1)=\pi_d(s_k,d_2)=\cdots\pi_d(s_k,d_m)$)且$\sum_{d\in D(s_k)}\pi_d(s_k,d)=1,\forall s_k\in S$，不引入额外的先验知识。

[0079] 测试不同参数设置对算法的影响，以图6和图7中状态$s_2$为例，实验中攻击者初始策略为随机策略，分析不同的参数取值会影响学习的速度和效果，对不同的参数设置做进一步测试，对六种不同的参数设置进行测试，具体的参数设置如附图9所示。

[0080] 防御者在状态$s_2$对防御动作$d_3$和$d_4$的选择概率结果如图10所示。从图10中可以观测不同参数设置下算法的学习速度和收敛性。图10中显示设置1、3、6的学习速度较快，三种设置下算法经过1500次以内的学习即可得到最佳策略，但3和6的收敛性较差。虽然设置3和设置6能学习到最佳策略，但之后会出现震荡，没有设置1的稳定性好。

[0081] 防御收益可以代表算法对策略的优化程度，为了确保收益值不是只反应一次防御结果，取1000次防御收益的平均值，其每1000次的平均收益变化如图11所示。从图11中可以看到设置3的收益明显低于其它设置，但其它设置的优劣难以区分。因此，六组参数中设置1最适合于本场景。

[0082] 测试资格迹带来的运算开销，分别统计了20次有、无资格迹时算法进行10万次防御决策的时间，其20次的平均值为：有资格迹9.51s，无资格迹3.74s。虽然资格迹的引入会使得决策时间增加近2.5倍，但是引入资格迹后10万次的决策所需时间仍然只有9.51s，仍可以满足实时性的需求。

[0083] 通过以上实验，进一步验证了本发明在有限理性约束下构建攻防随机博弈模型并生成用于网络状态与攻防策略提取的网络攻防图，有效压缩了博弈状态空间；防御者通过学习可以获得针对当前攻击的最优防御策略，提升了对未知攻击的快速自动化防御能力，具有较强的实用性和可操作性。

[0084] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0085] 结合本文中所公开的实施例描述的各实例的单元及方法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已按照功能一般性地描述了各示例的组成及步骤。这些功能是以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不认为超出本发明的范围。

[0086] 本领域普通技术人员可以理解上述方法中的全部或部分步骤可通过程序来指令相关硬件完成，所述程序可以存储于计算机可读存储介质中，如：只读存储器、磁盘或光盘

等。可选地，上述实施例的全部或部分步骤也可以使用一个或多个集成电路来实现，相应地，上述实施例中的各模块/单元可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。本发明不限制于任何特定形式的硬件和软件的结合。

[0087] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

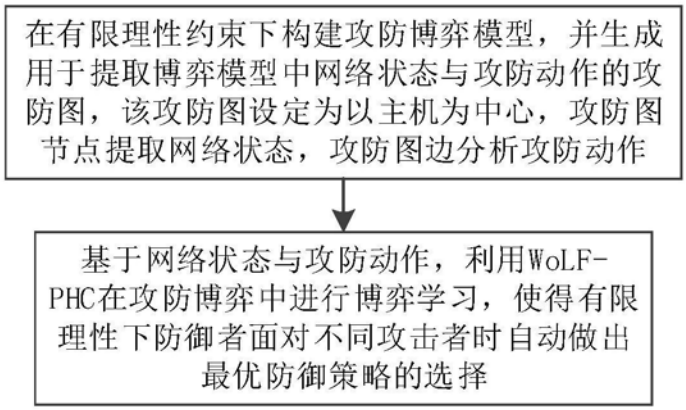

在有限理性约束下构建攻防博弈模型，并生成用于提取博弈模型中网络状态与攻防动作的攻防图，该攻防图设定为以主机为中心，攻防图节点提取网络状态，攻防图边分析攻防动作
基于网络状态与攻防动作，利用WoLF-PHC在攻防博弈中进行博弈学习，使得有限理性下防御者面对不同攻击者时自动做出最优防御策略的选择

图1

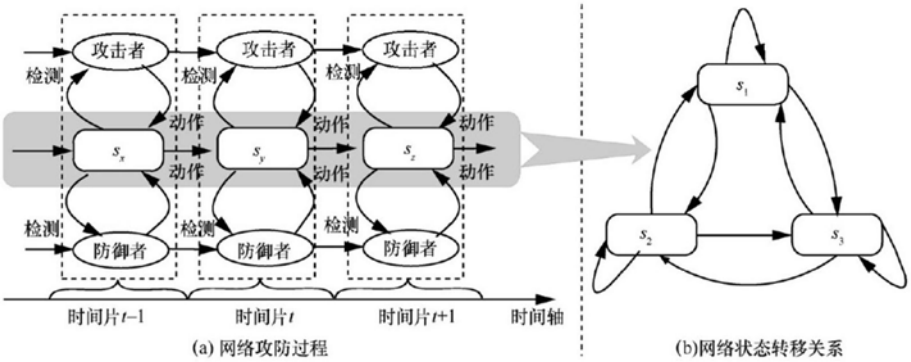

攻击者
检测
动作
防御者
时间片t-1
时间片t
时间片t+1
时间轴
(a) 网络攻防过程
(b)网络状态转移关系

图2

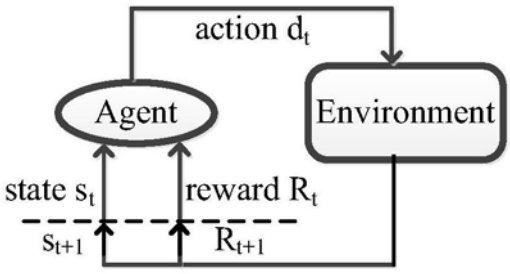

action $d_t$
Agent
Environment
state $s_t$
reward $R_t$
$s_{t+1}$
$R_{t+1}$

图3

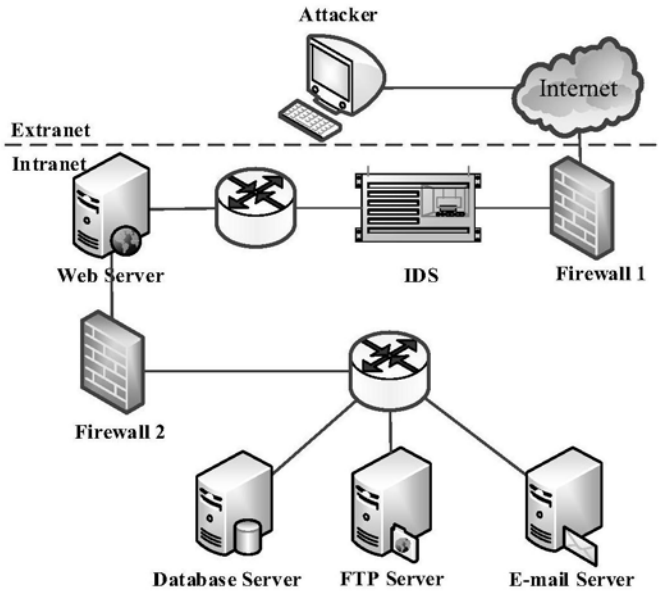



图4

| Attack Identifier | Host | CVE | Target Privilege |
|---|---|---|---|
| $tid_1$ | Web server | CVE-2015-1635 | user |
| $tid_2$ | Web server | CVE-2017-7269 | root |
| $tid_3$ | Web server | CVE-2014-8517 | root |
| $tid_4$ | FTP server | CVE-2014-3556 | root |
| $tid_5$ | E-mail server | CVE-2014-4877 | root |
| $tid_6$ | Database server | CVE-2013-4730 | user |
| $tid_7$ | Database server | CVE-2016-6662 | root |

图5

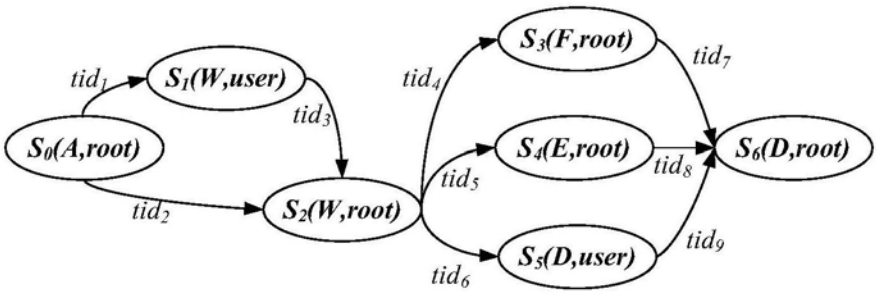



图6

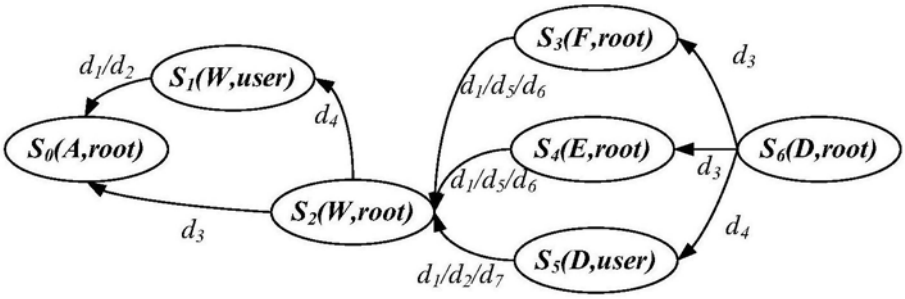


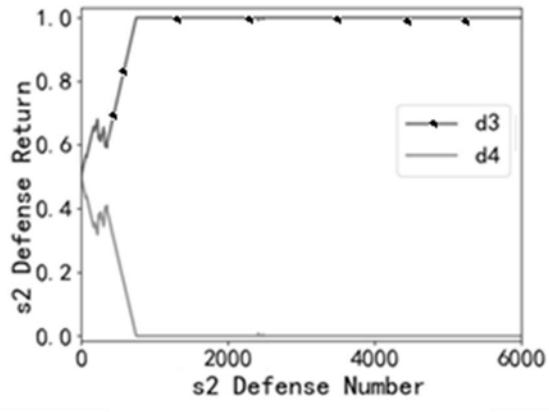

图7

| Atomic Defense Action | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|---|---|---|---|---|---|---|---|
| Renew root data | √ | | √ | | √ | √ | |
| Limit SYN/ICMP packets | | √ | | | | | |
| Install Oracle patches | √ | | | | | | √ |
| Reinstall Listener program | √ | | | | √ | | |
| Uninstall delete Trojan | | √ | | | | √ | |
| Limit access to MDSYS | | √ | | √ | | | |
| Restart Database server | | | √ | √ | √ | | |
| Delete suspicious account | | √ | | | | | √ |
| Add physical resource | √ | | | √ | √ | √ | |
| Repair database | | | √ | √ | | | √ |
| Limit packets from ports | √ | √ | √ | | | √ | |

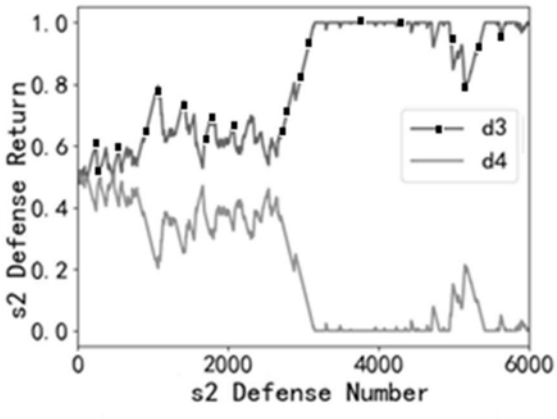
图8

| Set | $\alpha$ | $\delta_l$ | $\delta_w$ | $\lambda$ | $\gamma$ |
|---|---|---|---|---|---|
| 1 | 0.01 | 0.004 | 0.001 | 0.01 | 0.01 |
| 2 | 0.1 | 0.004 | 0.001 | 0.01 | 0.01 |
| 3 | 0.01 | 0.004 | 0.001 | 0.01 | 0.1 |
| 4 | 0.01 | 0.004 | 0.001 | 0.1 | 0.01 |
| 5 | 0.01 | 0.04 | 0.01 | 0.01 | 0.01 |
| 6 | 0.01 | 0.008 | 0.001 | 0.01 | 0.01 |

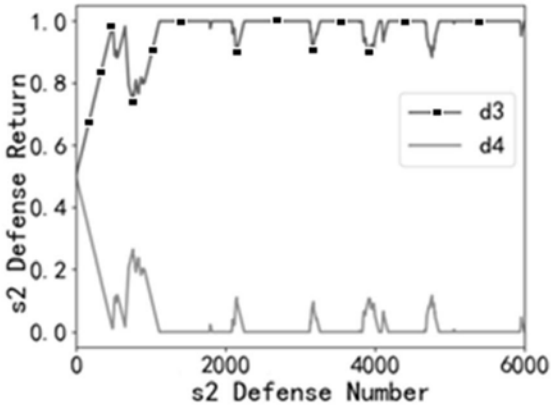
图9



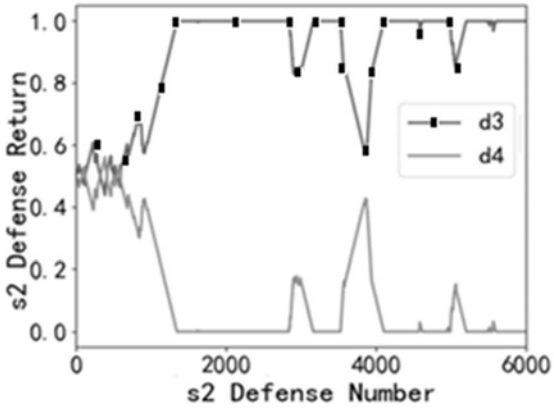

(a) Defense Decision under Set 1



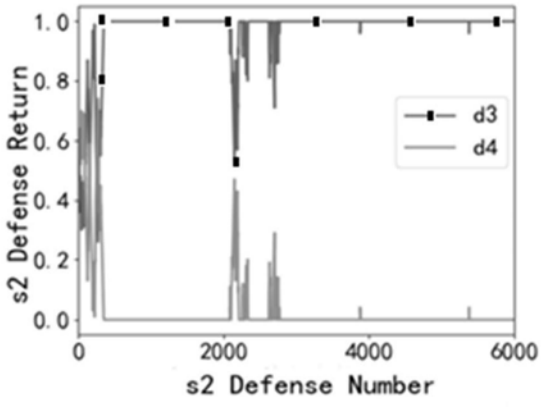

(b) Defense Decision under Set 2



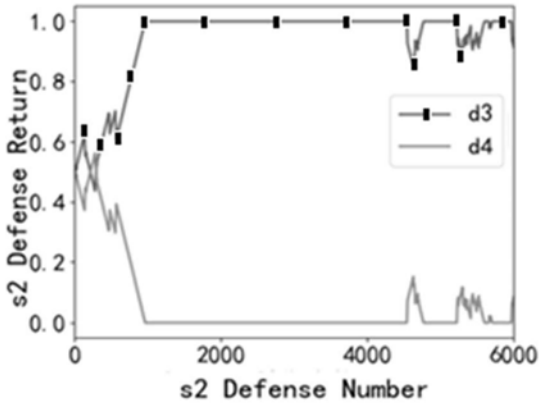

(c) Defense Decision under Set 3




(d) Defense Decision under Set 4




(e) Defense Decision under Set 5




(f) Defense Decision under Set 6

图10

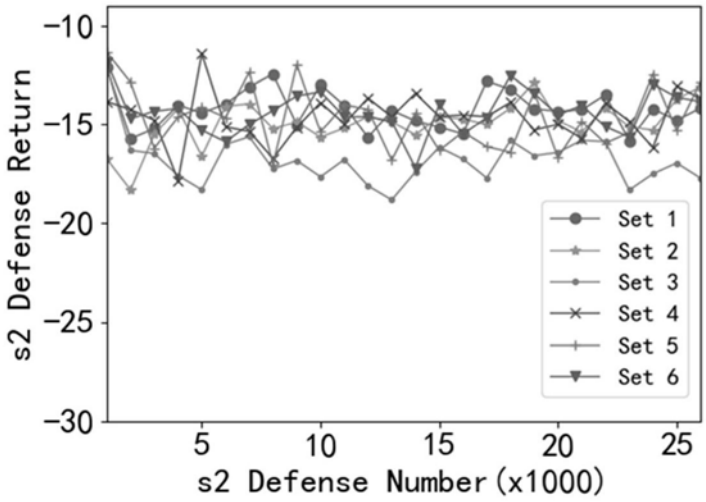

−10
−15
−20
−25
−30
5
10
15
20
25
s2 Defense Return
s2 Defense Number (x1000)
Set 1
Set 2
Set 3
Set 4
Set 5
Set 6

图11