doi: 10.7690/bgzdh.2021.01.003

网络空间作战中利用强化学习方法防御放大式拒绝服务攻击

陈　泱[1,2]，李卓禹[1,2]，闫海港[1,2]，张元天[3]
（1. 海军研究院，北京 100161；2. 复杂舰船系统仿真重点实验室，北京 100161；
3. 中国科学院信息工程所，北京 100091）

摘要：为解决放大式拒绝服务攻击给赛博安全带来的风险，提出一种基于强化学习的方法。以 DNS 的放大攻击为对象，构建一个简化的放大攻击模型，利用 model-free 方法获取不同状态间的转移概率，采用强化学习方法建立防御放大攻击模型，通过对放大攻击模式的学习制定流量抑制策略，并对其进行仿真实验验证。结果表明：该方法能够有效挖掘出放大攻击的流量模式，智能化抵御来自放大攻击的威胁。

关键词：网络空间作战；赛博安全；强化学习；放大式拒绝服务攻击
中图分类号：TP393.08　**文献标志码：**A

Using Reinforcement Learning Method to Defense Amplification DDoS Attack in Cyberspace Operations

Chen Yang[1,2], Li Zhuoyu[1,2], Yan Haigang[1,2], Zhang Yuantian[3]
(1. *Naval Research Academy*, *Beijing* 100161, *China*;
2. *Science & Technology on Complex Ship Systems Simulation Laboratory*, *Beijing* 100161, *China*;
3. *Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100091, *China*)

Abstract: To resist the risk of amplification DDoS attack, which is likely to cause significant damage to cyber security, a reinforcement learning method is proposed. Taking DNS as the target of attack, a simplified amplification attack model is constructed. The transition probability between different states is obtained by using the model-free method. Then, the reinforcement learning method is used to build up to defense the attack, and the traffic suppression strategy is formulated by learning the amplification attack mode. Finally, the simulation results show that the proposed reinforcement learning method can effectively dig out the traffic pattern of amplification DDoS attack and intelligently resist the threat.

Keywords: cyberspace operations; cyber security; reinforcement learning; amplification DDoS attack

0　引言

网络空间作战是对网络空间能力的作战运用，是国家战略竞争的重要手段，也是联合作战的重要组成部分。在十九大报告中，习近平强调坚持走中国特色强军之路，全面推进国防和军队现代化，要加快军事智能化发展，提高基于网络信息体系的联合作战能力[1]。海湾战争以来，战争形态逐渐转变为基于网络信息的体系作战。随着网络空间作战的发展，人们逐渐发现网络空间先天存在漏洞，并意识到网络体系的安全是其发挥效能的前提和基础。2003 年美国首次颁布《网络空间安全国家战略》，其后于 2011 年出台了《网络空间国际战略》、2015 年出台了《网络空间安全战略》等维护网络体系安全的政策法规[2]。

赛博安全(Cyberspace security)即网络空间安全，由网络防御者和攻击者(黑客)开展策略博弈，两者的行动相互独立，且无任何合作或沟通[3]。在网络空间攻击者探测和利用技术漏洞的能力和操作称为赛博威胁。拒绝服务(denial of service，DoS)攻击，是当前最严重的赛博威胁之一。这种攻击的目的是耗尽服务器的资源或阻塞网络，使被攻击者无法向其合法用户或客户提供服务，攻击的对象可以是任何联网计算机、路由器或整个网络。

DoS 攻击的方式在不断演变，其中，放大式拒绝服务攻击已成为当今 DoS 攻击中最流行和最危险的攻击方式之一。在放大式拒绝服务攻击中，攻击者伪装自己的地址，发送请求给开放的网络服务器，利用协议的固有特性放大返回流量造成带宽阻塞[4-6]。

放大攻击具有 IP 欺骗的性质，放大攻击的攻击者永远隐藏在幕后。对于被攻击者来说能看到的只有充当了放大器角色的开放服务器。与其他拒绝服

收稿日期：2020-09-15；**修回日期：**2020-10-29
作者简介：陈　泱(1989—)，女，江苏人，硕士，助理研究员，从事多深度学习、智能化技术军事应用、作战体系研究。E-mail: 657060823@qq.com。

务攻击不同，放大攻击的流量包格式与日常流量完全没有区别，流量来源的 IP 地址也是日常使用的服务器，很难用传统的数据包过滤方法来防御[7-8]；因此，笔者建立一种能防御分布式拒绝服务攻击的强化学习模型，能够将放大攻击的影响力降到最低。

1　背景

1.1　强化学习

强化学习是机器学习的一个分支，核心思想是利用智能体(Agent)与环境的交互获得奖励或惩罚形式的数字反馈，并用收到的反馈改进决策操作。解决强化学习任务的问题是找到一个策略来最大化累积的奖励。强化学习的基本解决框架是马尔可夫决策过程。假设系统具有马尔可夫性质，即状态转移概率和奖励取决于环境的状态和 Agent 在环境中的作用。

图 1 为强化学习的马尔可夫模型。Agent 在完成某项任务时，观察现有的状态 S_t，并选择一个动作 A_t与周围环境进行交互，在动作 A_t和环境的作用下，智能体会产生新的状态 S_{t+1}，同时环境会给出一个立即奖励 R_t。如此循环下去，Agent 与环境进行不断地交互，从而产生很多数据。强化学习算法利用产生的数据修改自身的动作策略，再与环境交互，产生新的数据，并利用新的数据进一步改善自身的动作，经过数次迭代学习后，Agent 能最终地学到完成相应任务的最优动作(最优策略)[9]。

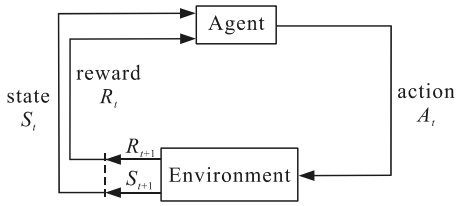

图 1　强化学习模型

强化学习的学习过程是个动态、不断交互的过程，转移概率 P 和奖励 R 构成了模型。γ 是折扣因子，用于计算累积的回报。在模型 γ 已知的情况下，马尔可夫决策过程可以用动态规划方法进行求解，核心方法是 Bellman 方程：

$$v(s)=R_s+\gamma\sum_{s'\in S}P_{ss'}v(s')。\tag{1}$$

式中 v 是价值函数，表示从状态到价值的映射，可以用$\langle\text{state},\text{action}\rangle$二元组来替代 v 值，用 action-value 函数替代价值函数：

$$Q(s,a)=R_s+\gamma\sum_{s'\in S}P_{ss'}Q(s',a')。\tag{2}$$

在大多数的现实领域中，环境模型是未知的，model-free(无模型)方法构成了强化学习的主干。蒙特卡罗方法和时序差分算法是 model-free 算法的代表。以时序差分算法为例，该算法提出了时序差分值 $R_S+\gamma Q(s,a)$，并利用时序差分更新 state-action 状态对 $Q(s,a)$[10]。时序差分方法实质上是从一个不完整的状态序列学习，经过采样不断更新 state-action 值的过程。可按照下式更新 state-action 值[11]：

$$Q(s,a)\leftarrow Q(s,a)+\alpha[R_s+\gamma Q(s',a')-Q(s,a)]。\tag{3}$$

其中 α 是学习速率。在状态 s 时采取了动作 a 获得回报 R_S，转移到新的状态 s'以及新的动作 a'，并更新原值。

强化学习的最终目标是获取策略并以此选择动作。定义策略 π: $\pi(s,a)\rightarrow[0,1]$，$\pi(s,a)$是动作 a 在状态 s 中发生的概率。在文中模型中，由于状态空间数量有限，可以使用传统的表格式方法存储 state-action 信息。

1.2　放大式拒绝服务攻击

与传统 DoS 攻击一样，放大攻击的主要目标是耗尽被攻击者的带宽资源或计算资源造成服务停止。一般来说，放大攻击有以下特征[12-14]：

1) 大量借用基于 UDP 协议的公共服务器发起攻击，这些服务器通常不会对攻击者的身份(如 IP 地址)做验证；

2) 伪装性，攻击者会将自己的 IP 地址伪造成被攻击者的地址，被攻击者会接受到来自放大服务器的大量流量。放大服务器指可以被攻击者滥用发起攻击行为的服务器；

3) 反射性，攻击者从来不直接发送流量给被攻击者，而是通过放大服务器的反射间接进行攻击；

4) 分布性，即攻击者可以从单一攻击源连接多个放大服务器，进行高分布式的攻击；

5) 放大性，从放大服务器反射到被攻击者的流量要远远大于攻击者发送给放大服务器的流量。

依据以上特征建立如图 2 所示的放大式拒绝服务攻击模型。攻击者向 3 个异构的放大器发送较小的请求包，把攻击者伪装成目标主机的 IP 地址作为发送给放大器请求的源地址，从而导致放大器向目标主机发送响应流量包。由于协议存在放大漏洞，放大器通常会发送比请求包大得多的响应包，从而在目标主机处造成带宽拥塞。

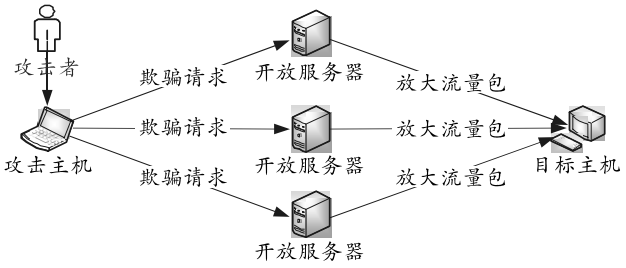

图 2 放大攻击模型

放大攻击的核心是攻击者滥用的具有放大漏洞的协议，决定了放大倍数以及放大器的数量。目前发现的放大攻击大部分利用 DNS 解析服务器来放大流量。攻击者将精心编制的请求包发往大量 DNS 解析服务器。这些服务器完成域名解析后将回复数百倍于请求包的流量，造成链路拥塞。笔者从网络空间作战的实际情形出发，以基于 DNS 的放大攻击为范例。

2 模型设计

2.1 威胁模型

如图 3 所示，最左侧节点是用户区，区域内受信任的主机用 H 表示。主机只接收流量，不对外发包。事实上正常的 DNS 请求包所消耗的带宽同 DNS 放大攻击所消耗的带宽相比微乎其微，可以忽略不计。右侧的节点集 S 表示 DNS 解析服务器集合。依据放大攻击的特性，集合 S 内的任意一台服务器都有一定的概率被攻击者利用成为放大服务器。在本例中，S1 和 S2 是合法的 DNS 解析服务器，S3～S5 是发起放大攻击的服务器，中间节点表示路由器。该路由器是距离用户区最近的，直接负责转发或丢弃所有通往用户区的流量。路由器与 DNS 服务器之间的网络结构省略，看作直连。该路由器是绝对可信的，因为与其他网络设施相比，路由器支持的服务数量有限，几乎没有可以被放大攻击利用的服务，并且想要控制路由器非常困难。路由器每连接一个 DNS 主机，路由器系统将划分一部分计算资源生成一个新的虚拟强化学习 Agent。每一个 Agent 对应一条链路的流量进行学习，学习的结果将汇总进行进一步的全局性学习。

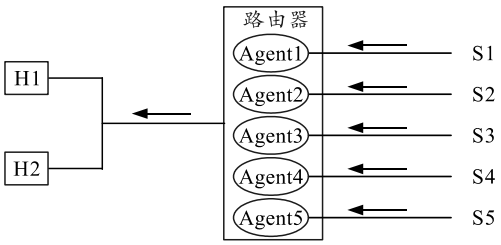

图 3 威胁模型

DNS 服务器集合 S 中的合法服务器以 v_n 的速率向路由器发送 DNS 数据包，攻击者以 v_a 的速率发送数据包。默认 $v_a \geq v_n$，如果攻击者发送流量的速率与合法流量相似，攻击者需要利用的攻击服务器将数百倍于合法服务器，大大增加了攻击成本。路由器到服务区之间的链路负载上限为 U。如果路由器转发的 DNS 流量小于 U，则用户可以正常工作；反之，放大攻击达成。

2.2 强化学习方法

强化学习方法的目标是最大限度丢弃攻击流量包并转发合法流量包。采用 model-free 方法获取不同状态间的转移概率。在一个回合中，威胁模型中的 5 个 Agent 通过观察流量状态作出决策。方法的主要要素包括状态、动作、奖励、折扣率、学习算法。

状态集：系统的 state 也即 Agent 获取的环境数据。对于每一个 Agent，设定其状态集为二元组 (L_{server}, L_{total})，L_{server} 是时间窗口 T 内从对应的 DNS 解析服务器接收到的流量负载，而 L_{total} 是链路总负载，是 T 时间内 5 个 Agent 的 L_{server} 值之和。学习算法的时间步长应同样设置为 T。

动作集：在每一个时间窗口内 Agent 在获取 state 二元组(L_{server}, L_{total})后，依据学习到的策略做出决策。Agent 的动作有 2 种：转发或者丢弃在时间窗口 T 内接收到的所有流量。5 个 Agent 按顺序执行动作，算法如下：

Algorithm 1. Take-action function:
Get L_{server}, L_{total}
Take action
If action is 转发:
$L_{total} = L_{total} + L_{server}$//update 主负载
End if

奖励：消除由放大攻击造成的网络拥塞，即避免 $L_{total} \geq U$，同时让尽可能多的合法流量到达用户区；因此，负载到达上限时给予 Agent 一个负奖励。负载未达上限时，合法流量被转发的比例即为奖励值(例如 50%的合法流量被转发给用户，奖励值即为 0.5)。此外，加入一定的先验知识，如果单一 Agent 接受到的流量 L_{server} 达到总负载的 30%，奖励值为 −1。算法如下：

Algorithm 2. Reward function:
If Ltotal＞U or $L_{server} > U*0.3$
Reward = −1
Else

Reward = (legitimate traffic reached)/(legitimate traffic)

End if

折扣率：折扣率决定未来的奖励对现在的 Q 值的影响程度，一般用 γ 表示，$0 \leqslant \gamma \leqslant 1$。$\gamma$=0 意味着 Agent 只关注于当前获取的奖励。文中设置 γ 值为 0.8。

学习算法：使用 Q-learning 方法来计算 $q(s,a)$ 的值。Q-learning 方法是一种异步策略的 model-free 方法，即一个回合生成的策略和想要优化的策略并不相同。这种方法有助于 Agent 探索整个 state-action 空间。Q-learning 方法采用以下的公式更新 Q 值：

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R_s + \gamma \max_a Q(s',a) - Q(S,A)]。(4)$$

图 4 描述了整个强化学习框架。使用 ε-greedy 策略生成一个回合并初始化 state-action 空间。每一个时间窗口为一步。每步依据算法 1、2 来获取动作和奖励，按照式(4)更新 Q 值。经过大量的回合后，Q 值将收敛，从种获取最优策略。

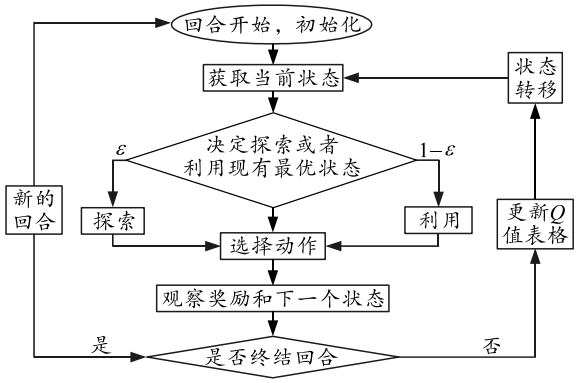

图 4　强化学习模型

完整算法的伪代码如下：

Algorithm 3: Q-learning for traffic throttling

Initialize $Q((s1,s2),a)$, s1 is Agent payload, $s2$ is total payload, Q(terminal, .)=0

Loop for each episode:

Initialize $S(s1,s2)$

Loop for each step of episode:

Loop for each Agent:

Choose A from S using policy ε-greedy，Take action A, observe R, S′($s1'$,$s2'$)

If $s1' > U$ or $s2' > U$*0.3, $R \leftarrow -1$

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R_s + \gamma \max_a Q(s',a) - Q(s,a)]$$

End loop

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R_s + \gamma \max_a Q(s',a) - Q(s,a)]$$

$S(s1,s2) \leftarrow S'(s1',s2')$

End loop

Until $S(s1,s2)$ is terminal

End loop

3　仿真实验及分析

3.1　参数设置

参数设置：链路负载以及流量单位为 Mbit/s。主链路负载设定为 10，正常 DNS 服务器发送的流量为区间[0,1]内的任意值，作为放大攻击的服务器发送的流量为区间[3,5]内的任意值。强化学习速率 α 设定为 0.1。

利用 2 个对比模型作有效性验证：模型 1 为不采取任何过滤措施的基线方法；模型 2 采用基于端口的过滤方法，即流量超过阈值时关闭相关端口。

3.2　仿真结果

为了验证模型的有效性，笔者进行了 3 个实验。每次实验训练 5 000 个回合，每回合结束后的奖励值能够方法性能。第 1 个实验是基于 ε-greedy 策略中 ε 值的性能对比，将 ε 值分别设置为 0、0.1、0.2。如图 5 所示，当 ε=0.1 时收敛最快，奖励值最高。在后续实验中 ε 值设置为 0.1。

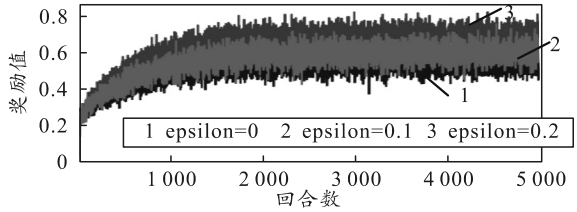

图 5　基于 ε 值的奖励对比

第 2 个实验是将强化学习方法与基于端口的过滤方法进行比较，实际上是防火墙最常用的策略。将每个 Agent 的流量阈值设置为 1.5。若一个 Agent 接收到的流量超过 1.5，则认为 Agent 将放弃相应 DNS 服务器发送的数据包。奖励值在每一个回合结束后分别计算。在此次实验中，默认固定攻击源，S3－S5 发送的流量为 3(如图 3 所示)。实验结果如图 6 所示。

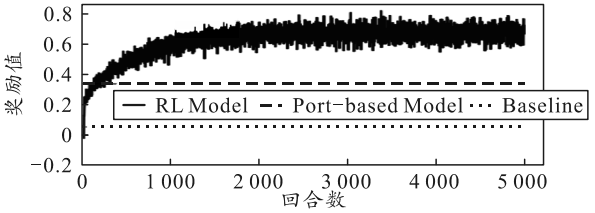

图 6　3 种模型的性能对比

在图 6 中，由于攻击源和攻击方式是固定的，每一回合结束后，基线模型和基于端口模型的性能相同，表现为直线。前者奖励接近于 0，后者奖励为 0.34。很明显，强化学习模型会随着时间学习并改进，直到最终收敛，显著优于其他 2 个模型。

为了更接近实际攻击模式，第 3 次实验不再固定攻击源和发送流量。实验结果如图 7 所示。基线模型的奖励仍然趋近于 0，而基于端口的模型奖励在 0 和 0.1 之间波动。基于端口的方法在前 1 000 个回合的性能优于强化学习模型。当训练趋于收敛时，强化学习模型具有明显的优势。

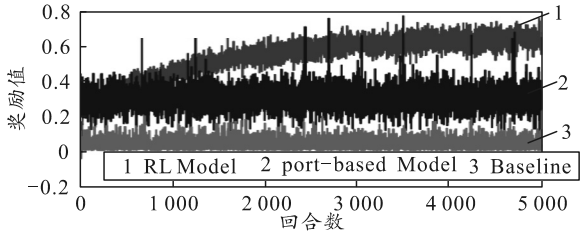

图 7 3 种模型的性能对比

图 8 为 3 种模型经过 10 次实验后的平均性能。性能指标是通过在一个回合中到达被攻击者的合法流量百分比来衡量的。很明显，强化学习模型性能最好。

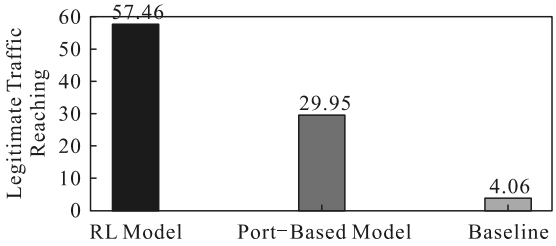

图 8 3 种模型到达被攻击者的合法流量百分比

3.3 仿真分析

实验的结果验证了强化学习方法具有一定的有效性与合理性。但此方法仍然具有很大的局限性：

1) 笔者采用离线学习。其优势主要是是模拟实验方便，实验环境可控，并且由于已知合法和非法流量，训练效率高；但劣势也很明显，模拟的攻击数据与实际数据有偏差，模拟数据中的标签在真实环境中不一定适用。下一步应当考虑在线学习算法，即系统直接在一个现实的网络中进行学习。

2) 文中算法只设置了转发或者丢弃所有流量 2 种动作。这样的设置便于仿真但粒度略粗。下一步将考虑一个连续的动作空间，可以任意地转发或丢弃一定百分比的流量。

4 结论

笔者提出一种基于强化学习的方法，将之动态地部署在核心路由器中，通过对放大攻击模式的学习制定流量抑制策略，目的是丢弃攻击流量并尽可能多的保留合法流量。通过建模仿真，验证了该法的有效性。结果表明：该强化学习方法能够在基本保留合法流量的前提下识别并丢弃攻击流量，并且效果优于传统的基于端口的流量抑制方法，在防御放大攻击方面有很好的应用前景。为应对赛博威胁，我军亟需充分利用各种高新技术手段，对军事网络实施重点防护，以形成网络空间作战赛博安全的联防、联管、联控。

参考文献：

[1] 习近平在中国共产党第十九次全国代表大会上的报告[N/OL]. 人民网：http://cpc.people.com.cn/BIG5/n1/2017/ 1028.

[2] 光明军事. 如何建设现代作战体系的“神经”和“血脉”[N/OL]. http://m.soho.com/a/258239446.gu.

[3] 杨林 译. 动态目标防御(Ⅱ)—博弈论与对抗模型的应用[M]. 北京: 国防工业出版社, 2014: 64.

[4] SAMAN T Z, JAMES J. A Survey of Defense Mechanisms Against Distributed Denial of Service Flooding Attacks[J]. IEEE Communications Surveys & Tutorials, 2013, 15: 2046−2069.

[5] MIRKOVIC J, REIHER P. A taxonomy of DDoS attack and DDoS defense mechanisms[J]. ACM SIGCOMM Computer Communication Review, 2004, 34(2): 39−54,.

[6] ROSSOW C, AMPLIFICATION H. Revisiting Network Protocols for DDoS Abuse[Z]. Network and Distributed System Security Symposium (NDSS '14), 2014: 23−26.

[7] CHEN Y, HWANG K. Collaborative change detection of DDoS attacks on community and ISP networks[Z]. IEEE Int' l Symp. on Collaborative Technologies and Systems, 2006: 401−410.

[8] ALOMARI E, MANICKAM S, GUPTA B B, et al. Alfaris. Botnet-based Distributed Denial of Service (DDoS) Attacks on Web Servers: Classification and Art[J]. International Journal of Computer Applications, 2012, 49(7): 24−32.

[9] RICHARD S S, ANDREW G B. Reinforcement learning: An Introduction[M]. The MIT Press, 2nd ed, 2017: 45−47.

[10] MARCO W, MARTIJN V O. Reinforcement Learning: State-of-the-Art[M]. Springer Science & Business Media, 2012: 35−48.

[11] TIMOTHY P L, JONATHAN J H, ALEXANDER P. Continuous control with deep reinforcement learning[Z]. ICLR, 2016: 12−18.

[12] 刘威. DNS 放大攻击的研究[J]. 信息网络安全, 2010(2): 46−48.

[13] 张小妹, 赵荣彩. 基于 DNS 的拒绝服务攻击研究与防范[J]. 计算机工程与设计, 2008(1): 27−30.

[14] 刘东鑫, 何明. 超大异常流量攻击的防御思路探讨[J]. 中兴通讯技术, 2015(6): 58−62.