基于SMDP强化学习的电力信息网络入侵检测研究

李 帅[1],王先培[2],王泉德[2],牛胜巍[3]
(1. 北京超高压公司,北京 100045;2. 武汉大学 电子信息学院,湖北 武汉 430072;
3. 濮阳电力公司 信息中心,河南 濮阳 457001)

摘要: 介绍了电力信息网络总体防护体系结构及安全现状,阐述了在电力信息网中常用的防火墙、入侵检测系统(IDS)等防护手段,分析了当前入侵检测方法及难以确定正常与异常的阀值、误报率和漏报率高的不足。提出了基于半马尔可夫决策过程(SMDP)强化学习的 IDS 模型。论述了强化学习的理论、算法及衡量标准,马尔可夫决策过程,SMDP 在电力信息网络中的应用。改进后的 SMDP 学习算法,使系统的误报率降低、检测率提高。

关键词: 电力系统; 强化学习; 半马尔可夫过程; 入侵检测
中图分类号: TM 73;TP 393.08 **文献标识码:** B **文章编号:** 1006-6047(2006)12-0075-04

1 入侵检测系统

随着信息技术的快速发展,计算机及网络技术已引入到电力生产、建设、经营、管理等各个环节,尤其是在电网调度自动化、厂站自动控制、管理信息系统、电力营销系统、办公自动化系统、电力负荷管理系统等方面,计算机网络系统的应用已有相当的基础,并且取得了较好的效果。

然而,在计算机网络得到广泛应用的同时,网络安全问题也日益突出。例如,2003 年美加电网发生的特大面积停电事故,造成事故的主要原因是蠕虫病毒阻碍了加拿大安大略省从停电事故中恢复正常供电的进程[1-4]。

为保障信息网络的安全,电力企业都配置了如防火墙、入侵检测系统[5-7] IDS(Intrusion Detection System)等网络安全产品,构成了如图 1 所示的电力系统信息网络的总体防护体系。

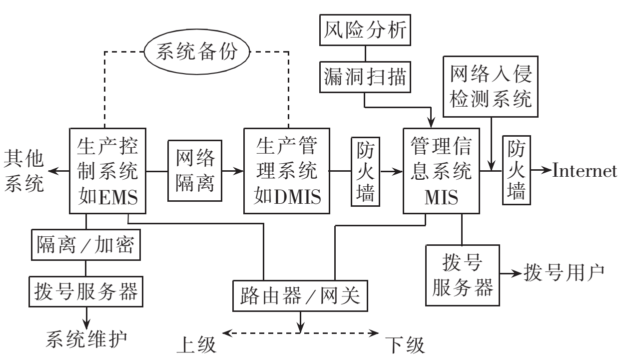

图 1 电力信息网络总体防护结构
Fig.1 The defense architecture of electric power information network

网络隔离技术将不同功能要求和不同保密层次的网络加以分隔,提高了网络整体的安全性;认证技术实现了访问控制的目的;加密技术的使用能较好地实现敏感文件的机密性;防火墙[7]起到过滤非法数据包的作用;入侵检测技术[8]则可以发现攻击行为。IDS 则是整个防护体系结构中的关键。

IDS 是一种主动防御攻击的网络安全系统,按检测原则分为滥用检测和异常检测。滥用检测是建立在已知攻击基础上的,对系统的未知攻击缺乏自适应性,漏报率很高。异常检测则是建立在网络及系统的正常行为模型上的,目的是希望降低入侵检测的误报和漏报率。基于异常检测的方法很多,有统计分析、贝叶斯网络、神经网络、数据挖掘、遗传算法等,困难之处在于难以确定正常与异常的阀值,误报率和漏报率也较高。鉴于这些原因,本文引入了一种基于强化学习的异常入侵检测模型,能较好地改善系统误报率和漏报率高的问题。

本文首先介绍了强化学习理论与方法,提出了一种基于半马尔可夫决策过程 SMDP(Semi-Markov Decision Process)强化学习的入侵检测系统模型。它不仅可以检测已知的入侵,而且对未知入侵的检测也比较有效,更重要的是通过 SMDP 学习算法的改进,缩短了建模时间,提高了检测率,降低了误报率和漏报率。

2 理论基础

2.1 强化学习

2.1.1 强化学习理论

强化学习[9-12] RL(Reinforcement Learning)又称再励学习,是一种试探评价的学习过程,广泛应用于智能控制、机器人及其他领域[8]。其理论基础是马尔可夫决策过程 MDP(Markov Decision Process)。入

收稿日期:2006-01-23;修回日期:2006-05-24

侵检测系统的环境状态可以看成是马尔可夫过程或者是近似马尔可夫过程，本文将其引入到入侵检测领域，图 2 为强化学习的基本模型。图中，智能体是进行状态感知、学习训练、动作选择的模块，环境是马尔可夫型的，state 是由一系列能描述环境的参数组成，action 的作用是将环境状态变迁，reward 则是环境状态奖赏值。它的基本思想是通过学习选择一个动作 a_t 作用于环境，环境接收该动作后状态会发生变化，于是产生一个强化信号 r_t（奖赏值）反馈给智能体，智能体根据强化信号和环境当前状态再选择下一个动作 a_{t+1}。选择的原则是受奖赏概率值增大，即如果智能体的某个决策行为导致来自外部的评价信号的改善，此后产生这个决策行为的趋势会更强，反之系统产生这个动作的趋势便减弱。

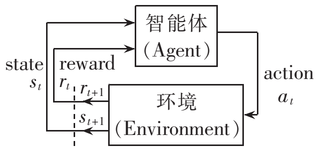

图 2 强化学习模型

Fig.2 The reinforcement learning model

2.1.2 强化学习算法

主要有动态规划 DP(Dynamic Programming)、蒙特卡罗方法 MC(Monte Carlo)、瞬时差分学习算法 TD(Temporal Differences)和 Q 学习算法[9]。

在得知状态转移概率和奖赏函数的情况下，可以用 DP 方法求解这个优化问题。而在绝大多数环境模型中，状态转移概率和奖赏函数是未知的，本文研究的入侵检测系统环境就属于该情况，笔者结合 SMDP 强化学习的特点，引入了 Q 学习算法，并在此基础上作了一些规则上的改进，系统性能得到很大的提高。

2.1.3 强化学习性能的衡量标准

a. 收敛性。所谓收敛性就是 Agent 能否最终收敛于最优行为。

b. 收敛速度。最优性通常是一个渐进的结果，因此收敛速度很难定义，一般用接近最优的程度作为衡量收敛程度的一种测量，通常选用给定时间长度观察 Agent 的性能。

2.2 马尔可夫决策过程

马尔可夫决策过程[11-12]由 4 元组$<S,A,P,R>$描述。其中，S 为环境有限状态集；A 为智能体选择动作集；状态转移函数 P：$S\times A\to PD(S)$；奖赏函数 R：$S\times A\to\Re$，记 $R(s'/s,a)$为系统在状态 s 用 a 动作使环境状态转移到 s'获得的瞬时奖赏值；记 $p(s'/s,a)$为系统在状态 s 采用 a 动作使环境状态转移到 s'的概率。智能体面临的任务是决定一个最优策略，使得总的折扣奖励信号期望值最大。定义了在 s 状态下的值函数 $V^{\pi}(s)$和在动作 a 评估函数 $Q(s,a)$。

$$V^{\pi}(s)=E_{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t r_t\,/\,s_0=s\right\} \tag{1}$$

这一值函数是无限时域累积折扣奖励期望值，其中，π 表示一种策略，r_t 和 s_t 分别是时刻 t 的奖赏值和状态，折扣系数 $\gamma(0\leqslant\gamma\leqslant 1)$使得临近奖赏比未来奖赏更重要。

在策略 π 的作用下，状态 s_t 的值函数应该满足 Bellman 方程：

$$V^{\pi}(s_t)=\max_{a_t\in A}Q^{\pi}(s_t,a_t) \tag{2}$$

$$Q^{\pi}(s_t,a_t)=R_{s_t}(a_t)+\gamma\sum_{s_{t+1}\in S}p(s_t,a_t,s_{t+1})V^{\pi}(s_{t+1}) \tag{3}$$

MDP 求解的目标是寻找一个最优策略 π^*，以使每个状态 $s(s\in S)$的值同时达到最大。即

$$V^{\pi^*}(s_t)=\max_{a_t\in A}\left\{R_{s_t}(a_t)+\gamma\sum_{s_{t+1}\in S}p(s_t,a_t,s_{t+1})V^{\pi^*}(s_{t+1})\right\} \tag{4}$$

2.3 SMDP 模型

Bradtke 和 Duff 提出半马氏决策过程[13-17]模型，Sutton 对该理论作了进一步的发展。它与马尔可夫模型的主要不同在于其状态之间的转移概率不是常数，而与状态保持时间分布有关。SMDP 可以看成是 MDP 的扩展，应用范围更广泛，它提供了智能体如何在抽象动作集中学会选取最优动作集合的理论基础，抽象时序动作在标准的 SMDP 概念中被认为成黑匣子技巧集合，也就是在任意一个时间段内和固定的状态空间下执行一系列的动作。SMDP 可以用一个 5 元组$<S,A,P,R,F>$定义。F 为 $P(t/s,a)$即在时刻 t 时，智能体在状态 s 情况下执行动作 a 到达下一个 SMDP 状态的概率值。如图 3 所示 SMDP 的过程。

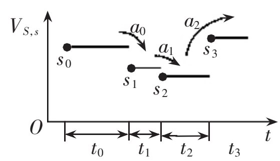

图 3 SMDP 过程

Fig.3 SMDP process

在 SMDP 的学习过程中，采用了 Q 学习算法，在规则上作了一些改进，可使 Q 值迅速地收敛。Q 学习迭代公式如下：

$$Q_{k+1}(s_t,a_t)\leftarrow(1-\alpha)Q_k(s_t,a_t)+\alpha[r+\gamma'\max_{a\in A}Q_k(s_{t+1},a_{t+1})] \tag{5}$$

式中 α 为学习率；γ 为折扣系数；t 为执行时间；r 为在执行时间 t 内的积累回报值。

3 SMDP 在电力信息网络应用

3.1 入侵检测系统总体结构

如图 4 所示，采用具有 SMDP 强化学习能力系统的总体结构。该系统由网络数据包捕获、数据包解析和预处理、SMDP 强化学习、响应处理 4 部分模块组成。其中，强化学习模块分为 SMDP 模型的学习模块和检测模块。

首先，捕获模块从数据网络系统的不同主机和网段上的若干关键节点收集和过滤一系列数据包，其次将这些数据包送到解析和预处理模块将数据包包头的特征提取和编码处理，形成特定格式的分类编码序列（如基于 TCP 端口号范围和标志位，UDP 端口号范围、ICMP 报文类型等），通过 SMDP 模型学习模块对这些不同协议类型的序列进行分类训练和学习，使之不断地健壮正常行为轮廓，再经过检测模块的实时分析检测，如发现异常，由响应处理模块报警和相关处理。

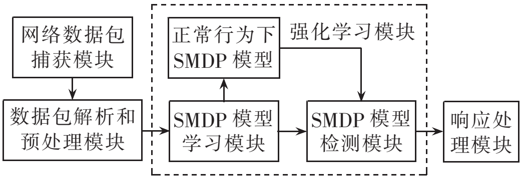

图 4　入侵检测系统的总体结构
Fig.4 The architecture of IDS

3.2　系统各模块及其功能设计

3.2.1　网络数据包捕获模块

网络数据包捕获模块为系统提供数据，数据的可靠性和正确性则是检测网络攻击的关键。模块具有捕获和过滤功能，将网络接口设置为混杂模式，在该模式下从计算机网络系统的不同网段和主机若干关键点截获数据包，通过过滤功能将目标 IP 地址不是本地主机的数据包和一些明显无用信息过滤掉，将得到的基于 IP、TCP、UDP 等协议数据包送到下一模块单元。

3.2.2　数据包解析和预处理模块

数据包解析主要是对过滤后的 IP、TCP、UDP、ICMP 等协议的数据包包头（header）进行解析，得到每个解析以太包头和 TCP/IP 包头的特征信息，如数据长度、源 IP、目的 IP、协议类型（TCP、UDP、ICMP）、源端口、目的端口等。

数据预处理主要功能是把解析得到的特征信息分类，同种类型的数据归纳在一起，将这些数据格式转换和编码处理，形成一系列的编码序列，使得这些序列可被 SMDP 模型学习模块和检测模块所识别和处理。

3.2.3　SMDP 模型的学习模块

在学习模块中，将解析和预处理模块传送的分类编码序列作为输入，将这些序列按照不同协议类型（如 TCP 端口号范围、TCP 标志位组合、UDP 端口号范围、ICMP 报文类型等），分类进行基于 SMDP 的训练学习，并使收敛的速度加快。如图 5 所示学习过程流程图。

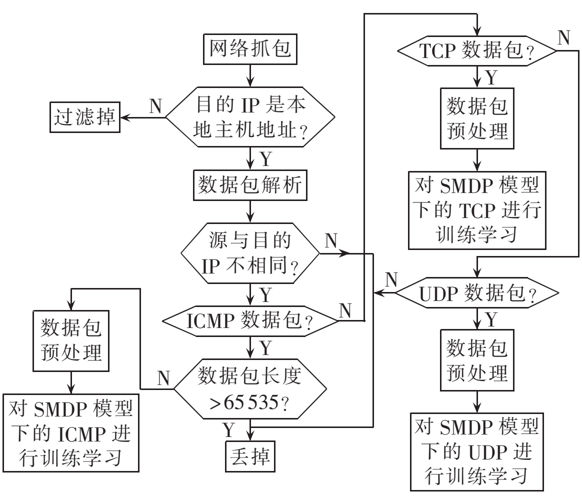

图 5 SMDP 模型学习过程流程图
Fig.5 Flowchart of learning process of SMDP model

目前，SMDP 强化学习有 3 种经典的方法[16-17]：Option、MAXQ、HAM。在该强化学习模块中本文引入 Option 方法。SMDP 可理解为在有穷序列状态 MDP 中加入一组确定的 Options，这时入侵检测系统的动作只在 Options 集合内选择，并且只有在该系统执行完整个 Options 后环境才返还给奖赏值，这样就可以运用 Q 学习算法解决如何在给定的 Options 集合上寻找最优策略 π^*。

Option 由一个 3 元组 $<\pi,\beta,\Gamma>$ 表示。其中，$\pi: S\times A\rightarrow[0,1]$，表示一个 Option 的内部策略，$\beta: S\rightarrow[0,1]$，表示一个终结状态判断条件，$\Gamma: \Gamma\subseteq S$，是一个激活 Option 时的初始状态。假设 Option 当前状态为 s，根据内部策略 π 选择某一动作，当确定选择下一动作后，动作作用于环境使环境状态转移为 s'，然后由 $\beta(s')$ 所给出的概率判定 Option 是终结还是继续执行。若当前 Option 执行结束，系统可以选择另一个 Option 开始执行上述的步骤。

利用 Q 学习算法先对值函数进行学习，值函数的每次更新都发生在 Option 执行结束之后。假设 Option o 在状态 s 开始执行，经过 τ 步在状态 s' 终结，且函数 $Q(s,o)$ 总是收敛于最优状态–动作对值函数。$Q(s,o)$ 的迭代算法如下：

$$Q_{k+1}(s,o)\leftarrow(1-\alpha)Q_k(s,o)+\alpha\left[r+\gamma^{\tau}\max_{o'\in O_{s'}}Q_k(s',o')\right] \tag{6}$$

在强化学习模型中，构造 Option 方法就是在入侵检测系统环境中随机的产生若干 Options，然后通过学习系统已经得到的正常行为策略 π，对该组 Options 测试进行筛选，进而构造相应的一组 Options，这样得到的 Options 提高了入侵检测系统的决策效率。Q 学习算法和 Option 方法的结合，已证明能更有效提高当前模块的学习速度，确保值函数迅速收敛。这样学习训练完成后得到新的 SMDP 模型被检测模块用于异常检测。

3.2.4　SMDP 模型的检测模块

采用 Q 学习算法和 Option 方法使各种协议类型的阀值按照规则得到更准确的更新，健壮了正常行为轮廓，当事件的状态编码序列通过检测模块时，根据 TCP/IP 协议组各自数据包包头类型阀值判断是不是正常行为。这样检测率得到了提高，同时也降低了误报率和漏报率。例如若在 1 s 内收到 200 个 TCP SYN 包，即判定是 SYN Flood 攻击，假设滑动窗口大小是 20，在 1 s 内接到 10 个，例如 cccc……ccc 异常这样大小的状态值序列，就可以判定是 SYN Flood 攻击。

3.2.5　响应处理模块

当检测出有异常行为时，响应模块会立即作出动作：向命令控制台发送报警信号；删除数据包，并且不向发端计算机发送错误信息；切断网络连接并记录事件。

以上介绍的基于 SMDP 强化学习入侵检测系统，采用 Q 学习算法和 Option 方法实时对电力信息网络上的数据包进行收集、解析、预处理和训练学习，

极大提高了学习的收敛速度,较好地解决了入侵检测系统误报率和漏报率高的问题,保证使电力信息网络有足够高的可靠性、可用性、机密性和完整性。

4 结语

电力系统是一个特殊的行业,与国民经济和人民群众的生活息息相关,而电能的安全传输又依赖于电力信息网络的正常工作。因此,保证电力信息系统网络安全已成为一项非常紧迫的任务。本文提出的基于SMDP强化学习的入侵检测系统不但检测率很高,而且大大降低了误报率和漏报率,有效地加强了电力信息网络的安全性。

参考文献:

[1] 韩祯祥,曹一家. 电力系统的安全性及防治措施[J]. 电网技术,2004,28(9):1-6.
HAN Zhen-xiang,CAO Yi-jia. Power system security and its prevention[J]. Power System Technology,2004,28(9):1-6.

[2] 印永华,郭剑波,赵建军,等. 美加"8.14"大停电事故初步分析以及应吸取的教训[J]. 电网技术,2003,27(10):8-11.
YIN Yong-hua,GUO Jian-bo,ZHAO Jian-jun,et al. Preliminary analysis of large scale blackout in interconnected North America power grid on August 14 and lessons to be drawn[J]. Power System Technology,2003,27(10):8-11.

[3] 戚宇林,刘文颖,杨以涵,等. 电力信息的网络化传输是电力系统安全的重要保证[J]. 电网技术,2004,28(9):58-61.
QI Yu-lin,LIU Wen-ying,YANG Yi-han,et al. Ensuring power security by networking transmission of electric power information[J]. Power System Technology,2004,28(9):58-61.

[4] 胡炎,谢小荣,韩英铎,等. 电力信息系统安全体系设计方法综述[J]. 电网技术,2005,29(1):35-39.
HU Yan,XIE Xiao-rong,HAN Ying-duo,et al. A survey to design method of security architecture for power information systems[J]. Power System Technology,2005,29(1):35-39.

[5] 王宁波,王先培. 容侵技术在电力系统数据网络安全中的应用[J]. 电力自动化设备,2004,24(10):35-38.
WANG Ning-bo,WANG Xian-pei. Application of intrusion tolerance technology in power system data network security[J]. Electric Power Automation Equipment,2004,24(10):35-38.

[6] 王先培,许靓,李峰,等. 一种3层结构带自保护的电力信息网络容入侵系统[J]. 电力系统自动化,2005,29(10):69-72.
WANG Xian-pei,XU Liang,LI Feng,et al. A tri-level self-protected intrusion tolerant system for electric power information network system[J]. Automation of Electric Power Systems,2005,29(10):69-72.

[7] 王先培,熊平,李文武. 防火墙和入侵检测系统在电力企业信息网络中的应用[J]. 电力系统自动化,2002,26(5):60-63.
WANG Xian-pei,XIONG Ping,LI Wen-wu. Application of firewall and IDS in the information network for power enterprise[J]. Automation of Electric Power Systems,2002,26(5):60-63.

[8] CHEBROLU S,ABRAHAM A,THOMAS J P. Feature deduction and ensemble design of intrusion detection systems[J]. Computer & Security,2005(24):295-307.

[9] SUTTON R S,BARTO A G. Reinforcement learning:an introduction[M]. London:Cambridge MIT Press,1998.

[10] MEHRAN A,MANFRED H. State space reduction for hierarchical reinforcement learning[C]//Proceeding of the 17th International FLAIRS Conference. Miami Beach,USA:AAAI Press,2004:509-514.

[11] DIETTERICH T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Artifical Intelligence Research,2000,13(2):227-303.

[12] MARBACH P,TSITSIKLIS J N. Simulation-based optimization of Markov reward processes[J]. IEEE Trans Automatic Control,2001,46(2):191-209.

[13] GHAVAMZADEH M,MAHADEVAN S. Hierarchical multi-agent reinforcement learning[R]. Canada:EI,2004.

[14] LIU Jian-yong,ZHAO Xiao-bo. On average reward semi-Markov decision processes with a general multichain structure[J]. Mathematics of Operations Research,2004,29(2):339-352.

[15] FEINBERG E A. Continuous time discounted jump Markov decision process:a discrete-event approach[J]. Mathematics of Operations Research,2004,29(3):492-524.

[16] SUTTON R,PRECUP D,SINGH S. Between MDPS and semi-MDPS:a framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence,1999,112(2):181-211.

[17] TANG Hao,LIANG Xiang-jun,GAO Jun,et al. Robust control policy for semi-Markov decision processes with dependent uncertain parameters [C] // Proceedings of the 5th World Congress on Intelligent Control and Automation. Hangzhou,China:IEEE,2004:515-518.

(责任编辑:汪仪珍)

作者简介:

李 帅(1980-),男,河北辛集人,硕士研究生,主要研究方向为电力实时系统安全与可靠性(**E-mail**:shuailimail@163.com);

王先培(1963-),男,湖北仙桃人,教授,博士研究生导师,主要研究方向为电力实时系统安全与可靠性。

Research on intrusion detection based on SMDP reinforcement learning in electric power information network

LI Shuai[1],WANG Xian-pei[2],WANG Quan-de[2],NIU Sheng-wei[3]

(1. Beijing Ultra-high Voltage Corporation,Beijing 100045,China;
2. Electron Information Institute,Wuhan University,Wuhan 430072,China;
3. Electric Power Information Network Center of Puyang,Puyang 457001,China)

Abstract: The defense architecture and current security conditions of the electric power information network system are introduced and general defensive technologies such as the firewall and the IDS (Intrusion Detection System) are described. Defects of current intrusion detection methods are analyzed:it is difficult to determine thresholds of normality and abnormality;the false-positive rate and the false-negative rate are high. A model of intrusion detection based on the SMDP(Semi-Markov Decision Process) reinforcement learning is proposed. Its principle,algorithm and standard, as well as the Markov decision process and applications of the SMDP in the electric power information network are discussed. The improved SMDP learning algorithm lowers the false-positive rate and the false-negative rate.

Key words: electric power systems; reinforcement learning; SMDP; intrusion detection system