

未知木马检测免疫算法的 r 值分析

王倩云¹, 罗玉漂¹, 陈云芳²

(1. 南京邮电大学 理学院, 江苏 南京 210003;

2. 南京邮电大学 计算机学院, 江苏 南京 210003)

摘要:免疫算法具备的轻量级、分布式和自我进化等特性能够有效解决信息安全领域中的未知木马的检测问题, 弥补传统木马技术上的缺陷。文中详细阐述了未知木马检测免疫算法及详细的检测流程, 从检测器的生成及检测过程两个角度总结了检测中存在的问题, 并给出了抗原检测方向的一个分析图。根据人工免疫原理, 采用阴性选择算法和 r 连续位匹配算法, 对人工免疫未知木马检测系统中的一个重要参数 r 值进行了仔细的分析与测试。实验测试结果表明检测器集越大, 系统检测性能越好, 系统的检测性能与检测参数 r 值是紧密相关的。

关键词:人工免疫; 木马检测; 检测器

中图分类号: TP309

文献标识码: A

文章编号: 1673-629X(2012)12-0175-04

Numerical Analysis of r in Immune Algorithm of Unknown Trojan Detection

WANG Qian-yun¹, LUO Yu-piao¹, CHEN Yun-fang²

(1. College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

2. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Immune algorithm with the lightweight, distributed and self evolution characteristics can solve the problem efficiently of unknown Trojan detection which is in the domain of information security. It makes up for the defect of traditional Trojan detection technology. It discusses concretely the immune algorithm of unknown Trojan detection and the process of detection, and summarizes the existent problem from both detector generating and detection's process perspective. A chart of trend of antigen detection is also proposed. According to the artificial immune theorem, negative selection algorithm and r bit sequence match algorithm is applied, the significant parameter r in the artificial immune system about the unknown Trojan detection is carefully analysed. Experimental results show that the greater the detection set is, the better detecting performance the system has, and the system's performance and detecting parameter r are closely related.

Key words: artificial immune; Trojan detection; detector

0 引言

21世纪计算机网络技术的快速发展使得网络安全的重要性日益凸显, 在众多网络安全问题中, 木马反检测、反清除、易植入性, 与控制端通信的特性使得许多反病毒软件对它束手无策。

生物免疫系统作为人体重要的防御系统, 能够有效识别已知和未知抗原, 并能够将非我分类清除, 保护生物机体自身不受外界病毒, 细菌等病菌的伤害^[1]。1971 第一届国际免疫学大会的召开, 标志着免疫学成

为一门独立、成熟的学科^[2]。学者们仿真生物免疫系统构造一个人工免疫系统, 并将它运用于其他领域, 解决工程与信息处理等问题。近年来, 人工免疫与计算机安全领域相结合的安全技术工作成为了相关学者的研究热点。目前人工免疫运用于许多领域, 如信息安全、机器学习、数据挖掘、模式识别等^[3]。

传统的木马检测技术根据分析方法, 可分为: 误用检测技术和基于异常检测技术。误用检测技术又称为特征码技术, 通过预设入侵模式, 检测当前用户的行为特征。它的准确率高, 但只能检测到已知木马。而基于异常的检测技术是通过建立正常行为模式, 检查系统的运行情况, 虽能检测到未知木马, 但漏报、误报率高。后出现了一种新型技术——行为分析技术^[4]。然而依然不能满足计算机网络系统安全性和性能的要求。入侵者可以逐渐改变自己的行为模式来逃避检

收稿日期: 2012-03-28; 修回日期: 2012-07-01

基金项目: 中国博士后科学基金项目(20090451241)

作者简介: 王倩云(1991-), 女, 从事信息与计算科学研究; 陈云芳, 副教授, 硕士生导师, 主要研究方向为信息安全、入侵检测、移动代理、人工免疫等。

测。对于未知木马和已知木马的变种,传统的木马技术存在检测率低、误报率高等缺点,这要求着计算机网络通信领域必须开发出一个新的检测系统。人工免疫的自适应、自学习、记忆等优点给了计算机工作者一个启发,由此产生了将人工免疫运用于木马检测技术的运用。

1 基于人工免疫的木马检测系统

人工神经网络,进化算法后,人工免疫成为通信安全系统领域的一个新颖研究点,引起学者的关注。Forrest 提出二进制下 r 连续位匹配规则 and 否定选择^[5]; D'haeselleer 提出了线性时间算法和贪婪算法^[6];而国内研究有:电子科技大学刘鹏飞提出了新的带变异否定选择算法^[7];燕山大学杨华玲完整介绍了穷举、线性检测器生成算法及其生成过程,公式的推导^[8];哈尔滨理工大学李鑫鑫等分别提出了 r -chunk 检测器^[9],混合检测器的思想以及多级否定选择算法^[10],混合匹配规则的检测器生成算法^[11];中国人民解放军理工大学张衡提出了 r 可变阴性选择算法,并对黑洞问题进行了分析^[12]等。人工免疫运用于木马检测系统中,可以动态监控网络,检测到各种变形的木马,保护计算机网络通信的安全。

1.1 相关定义

文中选用了二进制编码方式, r 连续位匹配规则^[4-6]。

- r 连续位匹配规则: 计算两个字符串中对应位置连续 r 位相同的个数;
- 匹配概率 P_M : 随机抽取的一字符串与任一检测器匹配的概率;
- 检测失败率 P_f : 表示一“非我”字符串不与 N_R 中的检测器所匹配的概率;
- 自我集大小 N_s ; 初始检测器集的大小 N_{R0} ; 成熟检测器集的大小 N_R ;
- 匹配长度 r : 表示检测器与字符串匹配相同符的长度;
- f : 表示一随机字符串不与自我集中元素匹配的概率;
- 字符串长度 l ; 字符串字母的种类 m 。

1.2 检测流程

检测流程主要包括了检测器的生成过程和检测过程,如图 1 所示。

1.3 算法

系统不对自我正常的网络活动产生免疫行为,这就是美国学者 Forrest 在 1994 年提出了阴性选择算法,它实现检测器的自我耐受过程。

最初 Forrest 的这个算法是在二进制环境下选用

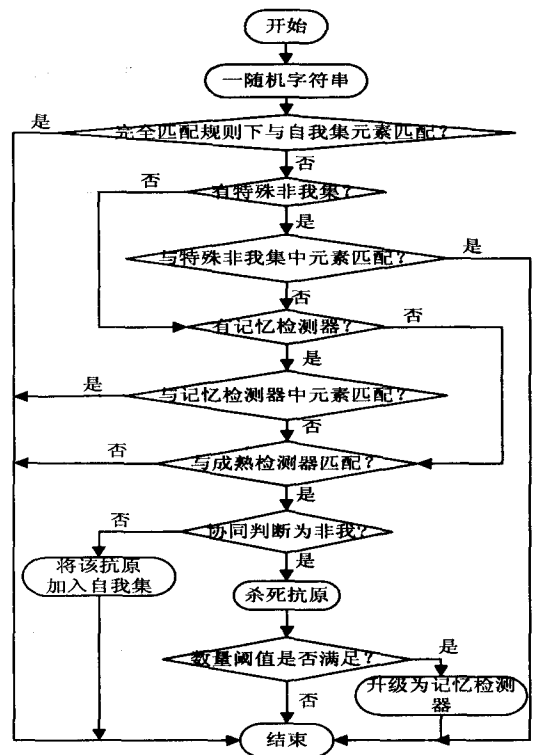


图 1 检测器检测过程

随机方式生成检测器,该算法完全将初始检测器一一列举出来,覆盖非我的范围广,但由于计算机的内存有限,生成检测器速度慢,且大量初始检测器重复,穷举法生成检测器并不是最优选择。而后为了解决这一问题,D'haeselleer 提出了针对 r 连续位匹配规则的检测器生成算法:线性时间算法和贪婪算法^[10]。线性时间检测器生成算法在缩短时间的目标下,进行一个有限递归运算,得到一定数量的不与自我匹配的字符串。贪婪算法则通过消除冗余的检测器来改进算法的效率,保证生成的检测器能够尽可能多的覆盖非己空间。

下面将分析穷举检测器生成算法的相关参数^[11]:

$$P_M = m^{-r} + m^{-r}[(m-1)/m](l-r) \quad (1)$$

$$f = (1 - P_M)^{N_s} \quad (2)$$

$$P_f = (1 - P_M)^{N_s} \Rightarrow \ln P_f = N_R \times \ln(1 - P_M) \Rightarrow \ln P_f = -N_R \times P_M \quad (3)$$

$$N_R = N_{R0} \times f \Rightarrow N_{R0} = \frac{N_R}{f} = -\frac{\ln P_f}{P_M} \times \frac{1}{f} = -\frac{\ln P_f}{P_M (1 - P_M)^{N_s}} \quad (4)$$

时间复杂度为: $O\left(-\frac{\ln P_f \times N_s}{P_M (1 - P_M)^{N_s}}\right)$, 空间复杂度为: $O(l \cdot N_s)$ 。

2 分析 r 值的必要性

2.1 存在的检测问题

在否定选择算法下,无论采用何种匹配规则,都存

在某些非自体字符串却找不到有效的检测器来发现,称其为检测漏洞^[6]。基于人工免疫的未知木马检测系统也存在检测不到某些非我抗原的缺陷。称这一类检测不到的抗原为黑洞。

首先,从检测器的生成角度上来看,黑洞存在的根源是检测这一类非我抗原的检测器不存在。而检测器不存在主要有两种可能:一、这部分检测器从未存在过;二、这部分检测器存在过,但在耐受过程中被删除了。

r 值的设定不当会使这两类可能性都增大。在相同环境下, r 值设定过小,则大量的初始检测器会被删除,造成检测器生成困难或是相应的非我抗原进入系统时不能被检测到。且在检测过程中容易造成误报。而 r 值设定过大,会有大部分初始检测器被保留下来,检测器的覆盖范围也较广,但在检测过程中,非我抗原可能会由于亲和力的不足而被判定为不匹配,从而使系统将该非我抗原误认为自我抗原,发生了漏报,对网络安全造成一定的威胁。

对于检测器不存在的第二类可能,还有一个因素会造成它的产生,即这类检测器由于与自我集中的自我相似而被删除了。

其次,从检测器检测抗原的过程上来看, r 值设定过小,模式字符串匹配度增大,每个检测器将匹配更多的串,这意味着需要的检测器数量可以减少且在检测器固定的情况下不会带来黑洞问题,但它会产生误报问题。 r 值设定过大,则对于每一个保留下来的成熟检测器,它的覆盖抗原的范围就变小,这由匹配规则为完全匹配时,一个检测器只能检测到某个特定的非我抗原事件中可得出此结论。所以,黑洞便因检测器覆盖不到而产生了。

2.2 检测性能指标

目前的检测器生成算法均不同程度地存在检测器生成效率低,存在漏洞或检测器冗余问题。那么测试一个检测系统的性能,到底应该选用何种指标呢?

由匹配概率的定义知,它并不能真实地反映一个检测系统性能的优劣,匹配概率只是表示随机抽取的一字符串与任一检测器匹配的概率。从实际操作来看,计算匹配概率难度较大,实现起来不太可能。所以应该选择失败率作为检测这一系统性能的标准。

一个抗原有两种命运:被检测器检测到和测不到,具体抗原检测分析如图2所示。

1、被检测器检测到又分为被记忆检测器检测到和被成熟检测器检测到。

(a)当被记忆检测器测到时,由于记忆检测器已是经过时间和实际训练得到的一个相当成熟稳定的检

测器,这时候可直接默认该抗原为非我抗原,直接删除该抗原。

(b)若是被成熟检测器检测到,则需要协同判断,若是协同判断为非我抗原,那也直接删除该抗原;否则需要将这个自我抗原加入自我集中,因为这个自我抗原一般来说是与自我集中的自我元素不相似。

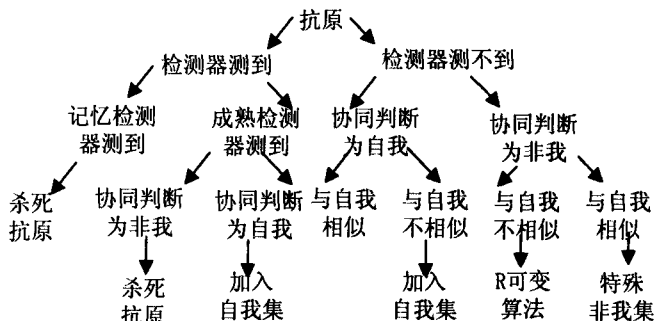


图2 抗原检测分析

2、检测器测不到时,这一类抗原中除了有自我抗原,还有非我抗原。

若是自我抗原,则从实际意义上来说它对系统无影响,先忽略之。

若是非我抗原时,面对这一检测漏洞问题时,需要对系统进行一些改进,以提高系统的匹配检测率,优化系统性能。而针对漏洞,要根据它们产生的原因提出改进的方法。经分析知,大致概括为两大点:a、该非我抗原与自我集中的自我元素相似;b、该非我抗原与自我集中的自我元素不相似。对于第一类概括所引起的检测漏洞,采取的改进方式是:建立一个特殊非我集,在检测过程中先将抗原与特殊非我集中的元素进行匹配检测,检测不成功再进行下面环节的检测。对于第二类概括所引起的检测漏洞,采取的改进方式是:选用 r 可变算法,通过 r 值的设定,调节检测器的覆盖半径。当 r 值越大时,检测半径越小; r 值越小时,检测半径越大。

所以 r 值的设定适当是至关重要的。

3 实验结果分析

本实验是在 matlab 环境下,模拟生物免疫系统,编写相对应的程序实现自我集,初始检测器,细胞耐受,成熟检测器,记忆检测器,特殊非我集的生成,并产生随机生成字符串模拟数据包,系统检测抗原的过程。

实验的思想是随机生成 n 个字符串,测出在检测过程中有多少个抗原被检测出来,表示字符为 count。在系统的一开始选用了完全匹配规则,将那些完全能确定的自我抗原筛选出来。设定好字符串长度 l ,阈值 line,通过调节自我集大小 sn,成熟检测器集大小 dn, r 值测试系统性能(记忆检测器元素 mdn,特殊非

我集元素 ssn 是在检测过程中,满足条件时自动生成)。

以下 3 个表格的实验结果,均在 $n = 2000, l = 32$, $line = 200$ 参数的设定下进行,其中表 2, 3 分别在表 1 的基础上改动检测器集合、自我集合的大小得到。

表 1 调节 r 值产生的实验结果

sn	dn	R	N	count	测不到	失败率	ssn	mdn
50	500	8	2000	2000	0	0	0	0
50	500	8	2000	2000	0	0	0	1
50	500	8	2000	2000	0	0	0	6
50	500	9	2000	2000	0	0	0	0
50	500	9	2000	1999	1	0.0005	1	0
50	500	9	2000	2000	0	0	1	0
50	500	10	2000	1996	4	0.002	3	0
50	500	10	2000	1994	6	0.003	7	0
50	500	10	2000	1994	6	0.003	13	0
50	500	11	2000	1891	109	0.0545	30	0
50	500	11	2000	1919	81	0.0405	54	0
50	500	11	2000	1926	74	0.037	83	0
50	500	12	2000	1533	467	0.2335	66	0
50	500	12	2000	1612	388	0.194	107	0
50	500	12	2000	1657	343	0.1715	146	0

表 2 检测器集合增大时调节 r 值产生的实验结果

sn	Dn	R	N	count	测不到	失败率	ssn	mdn
50	2500	8	2000	2000	0	0	0	0
50	2500	8	2000	2000	0	0	0	1
50	2500	8	2000	2000	0	0	0	4
50	2500	9	2000	2000	0	0	0	0
50	2500	9	2000	2000	0	0	0	0
50	2500	9	2000	2000	0	0	0	0
50	2500	10	2000	2000	0	0	0	0
50	2500	10	2000	2000	0	0	0	0
50	2500	10	2000	2000	0	0	0	0
50	2500	11	2000	2000	0	0	0	0
50	2500	11	2000	2000	0	0	0	0
50	2500	11	2000	2000	0	0	0	0
50	2500	12	2000	1997	3	0.0015	1	0
50	2500	12	2000	1995	5	0.0025	3	0
50	2500	12	2000	2000	0	0	3	0

表 3 自我集大小增大时调节 r 值产生的实验结果

Sn	dn	r	n	count	测不到	失败率	ssn	mdn
250	500	10	2000	1980	20	0.01	20	0
250	500	10	2000	1987	13	0.0065	33	0
250	500	10	2000	1994	6	0.003	39	0
250	500	11	2000	1904	96	0.048	88	0
250	500	11	2000	1949	51	0.0255	137	0
250	500	11	2000	1955	45	0.0225	174	0
250	500	12	2000	1779	221	0.1105	303	0
250	500	12	2000	1860	140	0.07	377	0
250	500	12	2000	1871	129	0.0645	435	0

由表 1 和表 2 可看出,检测器数量越多,检测的效果越好。但现实中检测器的数量是受限的,且检测器的生成与自我集的大小以及 r 的值有很大的关系。在其它参数的设定下, r 的取值为 10 时系统性能最优,10 之后失败率会逐渐增大。观察表 1 中 r 的取值为 8 或 9 的数据,可以发现这时抗原容易被测到,但不能排除这是由于 r 值设置过小,造成误报的可能。在生成表 3 的过程中,发现当 r 值为 8 时,在生成检测器时系统一直在运行而不出结果。由此分析知当自我集变大后, r 值要是设置过小,会造成检测器生成困难。

理论上的失败率为:

$$P_f = (1 - P_M)^{N_s}$$

$$(其中 P_M = m^{-r} [1 + \frac{m-1}{m}(l-r)])$$

$$P_M = 2^{-10} \times [1 + (32 - 10) \times (\frac{2-1}{2})] = 0.0017$$

$$P_f = (1 - 0.0017)^{500} = 0.0028$$

4 结束语

人工免疫运用于计算机网络安全领域是一个热点,但目前阶段它还只是处于研究阶段,并没有实际成品投入生活中使用。生物信息处理机制的复杂性还需要深入去挖掘,去完善人工免疫的理论。检测器的数量、 r 值的设定、字符串长度 l 与 r 值的比例、自我集大小都是影响系统生成检测器的效率以及系统检测性能的因素。检测器的生成必是研究的重点。

在未来,研究工作重点可以放在提高检测器的生成效率上。

参考文献:

- [1] Hosseinpour F, Bakar K A, Hardoroudi A H, et al. Survey on Artificial Immune System as a Bio-inspired Technique for Anomaly Based Intrusion Detection Systems [C]//International Conference on Intelligent Networking and Collaborative Systems. [s. l.]: [s. n.], 2010.
- [2] 孙勇智. 人工免疫算法系统模型、算法及其应用研究 [D]. 杭州: 浙江大学, 2004.
- [3] 张泽明. 人工免疫算法及其应用研究 [D]. 合肥: 中国科学技术大学, 2007.
- [4] 范明钰. 基于人工免疫的未知木马检测系统研究与实现 [D]. 成都: 电子科技大学, 2009.
- [5] Forrest S, Perelson S, Cherukuri R. Self-nonspecific Discrimination in a Computer [C]//Proceedings of IEEE Symposium on Research in Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1994: 202-212.
- [6] D'haeseleer P, Forrest S, Helman P. An Immunological Ap-

