

一种 r 可变阴性选择算法及其仿真分析

张 衡¹⁾ 吴礼发²⁾ 张毓森²⁾ 曾庆凯^{3),4)}

¹⁾(解放军理工大学通信工程学院 南京 210007)

²⁾(解放军理工大学指挥自动化学院 南京 210007)

³⁾(南京大学计算机科学与技术系 南京 210093)

⁴⁾(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘要 论文首先简要介绍了人工免疫系统的基本概念,然后着重分析了人工免疫系统中的主要算法“阴性选择算法”,并提出一种 r 可变阴性选择算法。同传统的阴性选择算法相比,该算法大大减少了不可避免的“黑洞”数量。仿真结果表明: r 可变阴性选择算法产生成熟检测器的迭代次数、黑洞数量均大幅下降,同时检测率有显著提高。

关键词 人工免疫系统; 阴性选择算法; 匹配阈值; 黑洞

中图法分类号 TP301

An Algorithm of r -Adjustable Negative Selection Algorithm and Its Simulation Analysis

ZHANG Heng¹⁾ WU Li-Fa²⁾ ZHANG Yu-Sen²⁾ ZENG Qing-Kai^{3),4)}

¹⁾(Institute of Communication Engineering, PLA University of Science and Technology, Nanjing 210007)

²⁾(Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007)

³⁾(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

⁴⁾(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

Abstract The Artificial Immune System (AIS) is introduced briefly, then as one of the main algorithms in AIS, negative selection algorithm is discussed. In order to reduce the number of holes which is inevitable in negative selection algorithm, an algorithm of r -adjustable negative selection algorithm is put forward. The new algorithm can reduce the number of holes through adjusting threshold of matching. The simulation results show that both the number of iteration to generate mature detector and the number of holes decline quickly, while the rate of detecting abnormality raises.

Keywords artificial immune system; negative selection algorithm; threshold of matching; hole

1 引言

作为生物系统中的信息处理系统,生物免疫系统模式识别、记忆、学习、多样性产生、分布式检测、

噪声耐受等功能对于解决目前许多实际问题很有启发,形成了人工免疫系统(Artificial Immune System, AIS)这个新的研究领域^[1,2]。

在生物免疫系统中,免疫识别是一项主要功能,其本质是区分“自我”与“非我”。免疫识别是通过淋

收稿日期:2004-01-05;修改稿收到日期:2005-05-21. 本课题得到国家“八六三”高技术研究发展计划项目基金(2002AA141090, 2004AA147070)、国家自然科学基金(60473053)资助。张衡,男,1977年生,博士研究生,研究方向为安全操作系统、人工免疫系统。E-mail: e_zheng@sohu.com. 吴礼发,男,1968年生,教授,研究领域为网络管理、网络安全。张毓森,男,1949年生,教授,博士生导师,研究领域为指挥自动化系统、系统仿真、信息安全。曾庆凯,男,1963年生,教授,研究领域为网络安全、分布计算。

巴细胞上的抗原识别受体(receptor)与抗原的结合实现的,结合强度称为亲合度(affinity). 淋巴细胞的产生是免疫系统一切功能的基础,美国New Mexico大学的Forrest教授模拟淋巴细胞的产生过程,提出阴性选择算法^[3~5]. 在阴性选择算法中,初始检测器的生成是一个重要的步骤,目前主要有穷举法和D'haesleer给出的基于连续位匹配规则的生成算法^[6],但由于该算法的时间和空间复杂度与匹配阈值成指数关系,为 $O((l-r) \cdot 2^r) + O((l-r) \cdot N_S)$ 和 $O((l-r)^2 \cdot 2^r)$,它不适用于 l 和 r 较大的情况.

作为人工免疫系统中的核心算法之一的阴性选择算法,其性能对整个系统具有重要意义. 本文对阴性选择算法进行了深入研究,为了减少传统的阴性选择算法不可避免的“黑洞”数量,提高系统检测率和性能,提出一种 r 可变阴性选择算法. 同时对在这种算法下,产生成熟检测器的迭代次数、黑洞数量、检测率进行了仿真分析.

2 问题定义

定义1. 模式. 由 l 个符号组成的字符串, $X = X_1 X_2 \cdots X_l$. 其中 $X_i (i=1, 2, \dots, l)$ 在本文中符号限取0或1,这样模式就是长度为 l 的二进制串.

本文中用 U 表示所有模式的集合, N 表示所有非我模式(nonself)集合,简称非我集, S 表示所有自我模式(self)集合,简称自我集. 显然公式 $U = N \cup S$ 成立. 通过阴性选择算法生成的、可以检测出非我模式的模式集合称为检测器集,

定义2. 匹配. 在一定的匹配规则下,两个模式串 a 和 b 的相似程度超过匹配阈值,则称 a 和 b 匹配,记为 $\text{Match}(a, b)$.

设有两个模式串 a 和 b , $a = X_{a1} X_{a2} \cdots X_{al}$, $b = Y_{b1} Y_{b2} \cdots Y_{bl}$. 在人工免疫系统中,一般采用的匹配规则是 r 连续位距离,其定义如下.

定义3. r 连续位距离:对于两个模式串 a 和 b ,当且仅当它们在 r 或大于 r 个连续的位置上有相同的字符时,则它们在 r 连续位规则下匹配.

同样例如, $a = 1011010101$, $b = 1001011101$,当 $r \leq 3$ 时, $\text{Match}(a, b)$.

两个随机的模式串 a 和 b 在连续位规则下匹配的概率为^[7,8]

$$P(\text{Match}(a, b)) = 2^{-r} \left(\frac{l-r}{2} + 1 \right).$$

3 阴性选择算法^[4]

首先介绍阴性选择算法:

1. 定义自体为一长度为 L 的字符串的集合 S ;
2. 随机产生一长度为 L 的字符串 a ;
3. 将字符串 a 依次与集合 S 中的字符串匹配;
4. 根据匹配规则,如果 a 遇到与之匹配的字符串,则结束匹配,转到步2;
5. 如果 a 不与 S 中任何字符串匹配,则 a 成熟,将 a 加入到检测器集中.

图1表示了这个过程.

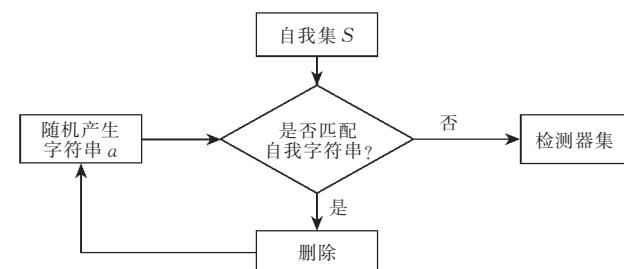


图1 阴性选择算法流程示意图

由阴性选择算法产生的检测器是成熟的,可以参加实际的检测活动. 从这里我们可以得到:每个检测器只覆盖非我集的一部分,实际上可以独立地行使功能,相互之间无需协调,一组检测器可以随意分布于多个位置;由这些检测器组成的系统对非我的检测是概率检测;由于字符串空间是有限的,如果所定义的自体是完整的,则不会有误报产生.

在阴性选择算法下,无论采用何种匹配规则,都有“黑洞”存在.“黑洞”的定义如下.

定义4. 黑洞. 一非我模式串 $a \in N$ 是一个黑洞,如果存在检测器 s ,使得 $\text{Match}(a, s)$,则 $\exists t \in S$,使得 $\text{Match}(t, s)$. 也就是说,黑洞中的非我模式串是无法产生相应的检测器来检测的.

图2给出了黑洞的直观形象表达. 在模式空间

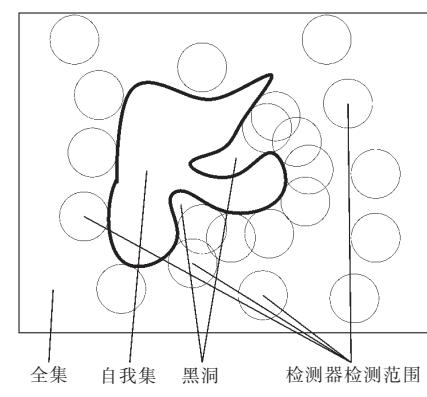


图2 黑洞的直观表示

中,自我模式集与非我模式集的界面往往是不规则的,而匹配阈值是固定的,因此有一些非我模式不能被任何检测器检测.

例如,设自我集 $S=\{001101,001111,011111\}$,匹配阈值 r 为 3,则 011101 是一个黑洞. 黑洞存在于任何匹配规则中. 事实上,在生物免疫系统中,也存在黑洞问题,而且病原体总在向黑洞中进化,使其更难被检测到. 生物免疫系统采用 MHC 机制^[9] 来解决黑洞问题. 每一种 MHC 类型都可认为是一种表示蛋白质的方式.

黑洞的存在取决于模式集的结构和模式匹配所采用的匹配规则. 自我模式越相似,黑洞数量越少. 对于同一种匹配规则,匹配阈值越大,黑洞数量越少. D'haesleer 给出了基于连续位匹配规则的计算黑洞数量的算法^[6]. 周建国^[10] 提出一个用于判断某一非我模式串是否属于黑洞的算法,其空间复杂度为 $O(l-r)$. 为了解决黑洞问题, Hofmeyr^[11] 提出采用多重表示模拟生物免疫系统的 MHC 机制,减少黑洞的数量. 多重表示法使用同一种匹配规则,但运用不同的模式表示法,其最大缺点是对系统性能有较大影响.

4 r 可变阴性选择算法

为了减少黑洞的数量,本文提出一种 r 可变的检测器产生算法,如下所示:

1. 定义自体为一长度为 L 的字符串的集合 S ;
2. 随机产生一长度为 L 的字符串 a , 初始匹配阈值为 r_1 ;
3. 将字符串 a 依次与集合 S 中字符串匹配;
4. 根据匹配规则,如果 a 不与 S 中任何字符串匹配,则 a 成熟,将 a 与匹配阈值加入到检测器集中. 转到步 2;
5. 当 a 遇到与之匹配的字符串,则将匹配阈值调整为 r' ,如果 $r' > r_c$,转到步 2;否则转到步 3.

在算法中,设匹配阈值 r 的变化为 r_1, r_2, \dots, r_c ,共 C 个,且 $r_1 < r_2 < r_3 < \dots < r_c$, r_c 为最大匹配阈值. 在后面的分析中,匹配阈值的调整策略是 $r_i = r_{i-1} + 1$,共 C 个,即最大匹配阈值为 $r_1 + C - 1$.

r 可变的阴性选择算法的核心思想是通过调整匹配阈值这一比较简单的方法大幅度降低黑洞数量,与普通阴性选择算法的不同之处主要在于以下几点:

(1) 产生过程不同, r 可变的阴性选择算法有多个可调的匹配阈值,普通阴性选择算法只有固定的匹配阈值;

(2) 检测集内容不同, r 可变的阴性选择算法产生的检测器集不仅有检测器,还包括其对应的匹配阈值,普通阴性选择算法产生的则只有检测器;

(3) 后续的检测过程不同,在 r 可变的阴性选择算法产生的不同检测器检测范围不同,普通阴性选择算法产生的不同检测器检测范围固定.

因此,普通阴性选择算法是 r 可变的阴性选择算法的一个特例.

5 r 可变阴性选择算法的仿真分析

本文对这一算法进行了仿真,分析采用 r 可变的检测器产生算法对产生一个成熟检测器需要的迭代次数、检测率、检测器数目及黑洞数目的变化情况.

仿真所使用的数据为:

(1) 采集数据集(CDS). 获取正常运行 FTP 时产生的 LSM 截获点数据^[12],对所采集的数据,经学习和处理所得的长为 80 的模式集合. 集合有 1566 个模式. 在 r 连续位距离下这个数据集的平均距离为 14.881.

(2) 随机生成数据集(RDS). 随机生成的长为 80 的二进制串的模式集合. 数据集的模式数目同样是 1566. 在 r 连续位距离下这个数据集的平均距离为 5.676,远远小于采集数据集的 14.881.

定义各种符号如下^[4]:

N_{R0} 为耐受前检测器数目;

N_R 为耐受后检测器数目;

N_S 为自我串数目,即自我集大小;

P_M 为两个随机串能够匹配的概率;

f 为随机串不与任何自我串匹配的概率,易有 $f=(1-P_M)^{N_S}$;

P_f 为漏报率;

P_s 为检测率, $P_s=1-P_f$.

5.1 产生一个成熟检测器需要的迭代次数

设 P_M^i 为两随机串在 r_i 连续位匹配规则下的匹配概率. 则有

$$P_M^i = 2^{-r_i} \left(\frac{l-r_i}{2} + 1 \right).$$

对于随机的串 a ,能够通过 r 可变阴性选择算法成为检测器的概率 P_D 满足

$$\begin{aligned} P_D &= (1-P_M^1) + P_M^1(1-P_M^2) + \\ &\quad P_M^1 P_M^2 (1-P_M^3) + \dots + P_M^1 P_M^2 \dots P_M^{C-1} (1-P_M^C) \\ &= 1 - P_M^1 P_M^2 \dots P_M^{C-1} P_M^C. \end{aligned}$$

产生一个成熟检测器需要的迭代次数满足

$$D(E) = \frac{1}{(P_D)^{N_s}}.$$

仿真实验和理论计算中, 首先要确定 r_1, r_1 较小时, 迭代次数过大; 过大时迭代次数少, 但要达到一定的检测率需要的检测器数目多, 通过仿真分析和综合判断, 确定 $r_1 = 13$. 在不同最大匹配阈值下, 产生 1 个成熟检测器需要的迭代次数如图 3 所示.

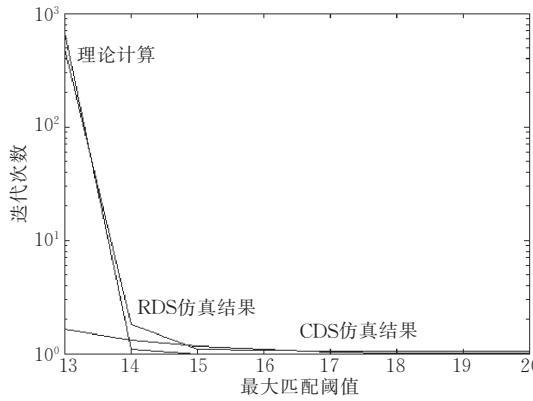


图 3 理论计算和仿真实验的迭代次数

最大匹配阈值为 13 时即在 $r = 13$ 的普通阴性选择算法(r 不变). 当最大匹配阈值增加时, 迭代次数减小. 采用采集数据集(CDS)的迭代次数与理论计算有较大差距, 而采用随机生成数据集(RDS)的迭代次数与理论计算非常接近, 这说明迭代次数与自我集内部模式之间的相似度相关.

5.2 不同匹配阈值的检测器的分布

对于随机串, 由前面的分析可知:

随机串在匹配阈值为 r_i 时, 不与任何自我串匹配的概率 $f_i = (1 - P_M^i)^{N_s}$.

成为匹配阈值为 r_i 的检测器概率 $P^i = (1 - f_1) \cdot (1 - f_2) \cdots (1 - f_{i-1}) f_i$.

设耐受前检测器数目为 N_{R0} , 则经过 r 可变阴性选择算法后检测器数目为 $N_R = N_{R0} (1 - P_D)$, 根据理论分析, 不同匹配阈值的检测器的分布满足

$$N_{Ri} = N_{R0} P^i.$$

在实验中, 当产生 20000 个检测器, 即 $N_R = 20000$ 时, 在不同 C 下, 不同匹配阈值的检测器的分布情况如表 1 所示.

表 1 不同匹配阈值的检测器的分布(占成熟检测器比例)

N_{Ri}						
$r_c=13$	$r_c=14$	$r_c=15$	$r_c=16$	$r_c=17$	$r_c=18$	$r_c=19$
$r=13$	20000	15985	14281	13393	12805	12503
	(100%)	(80%)	(71%)	(67%)	(64%)	(63%)
$r=14$	4015	3438	3177	3055	3093	3076
	(20%)	(17%)	(16%)	(15%)	(15%)	(15%)

(续 表)

N_{Ri}						
$r_c=13$	$r_c=14$	$r_c=15$	$r_c=16$	$r_c=17$	$r_c=18$	$r_c=19$
2281	2113	2139	2071	2012	2073	
(12%)	(11%)	(11%)	(10%)	(10%)	(10%)	
$r=16$	1317	1253	1223	1194	1206	
	(6%)	(6%)	(6%)	(6%)	(6%)	
$r=17$	748	720	721	754		
	(4%)	(4%)	(4%)	(4%)		
$r=18$	390	394	421			
	(2%)	(2%)	(2%)			
$r=19$	232	222				
	(1%)	(1%)				
$r=20$	112					
	(1%)					

当随机产生 20000 个模式串, 即 $N_{R0} = 20000$ 时, 在不同 C 下, 不同匹配阈值的检测器的分布情况如表 2 所示.

表 2 不同匹配阈值的检测器的分布(占未成熟检测器比例)

N_{Ri}						
$r_c=13$	$r_c=14$	$r_c=15$	$r_c=16$	$r_c=17$	$r_c=18$	$r_c=19$
12226	12160	12154	12302	12342	12222	12251
(61%)	(61%)	(61%)	(62%)	(62%)	(61%)	(61%)
$r=14$	3046	3047	2936	2948	2830	2980
	(15%)	(15%)	(15%)	(15%)	(14%)	(15%)
$r=15$	2006	1945	1939	2044	2040	1989
	(10%)	(10%)	(10%)	(10%)	(10%)	(10%)
$r=16$	1239	1212	1207	1196	1178	
	(6%)	(6%)	(6%)	(6%)	(6%)	
$r=17$	674	727	687	715		
	(3%)	(4%)	(3%)	(4%)		
$r=18$	409	378	391			
	(2%)	(2%)	(2%)			
$r=19$	214	232				
	(1%)	(1%)				
$r=20$	120					
	(1%)					

从表 2 可以看出: 不同匹配阈值的检测器占未成熟检测器的比例基本固定. 从表 1 中得到不同匹配阈值的检测器占全部成熟检测器的比例也趋于稳定, 这是由于随着最大匹配阈值 r_c 的增加, 未成熟检测器的成熟概率趋于 1.

5.3 检测率 P_s

漏报率 P_f 满足

$$P_f = (1 - P_M^1)^{N_{R1}} + (1 - P_M^2)^{N_{R2}} + \dots + (1 - P_M^C)^{N_{RC}},$$

检测率满足

$$P_s = 1 - P_f = 1 - (1 - P_M^1)^{N_{R1}} - (1 - P_M^2)^{N_{R2}} - \dots - (1 - P_M^C)^{N_{RC}}.$$

仿真实验中,采用了不同的 N_{R_0} 和不同的最大匹配阈值 r_c ,图 4 是实验结果.

图 4 中,检测率随着最大匹配阈值的增加而增加,而且当最大匹配阈值 r_c 超过 15 时,检测率趋于平缓,因此 r_c 取值 15. 最大匹配阈值为 13(即 $r=13$ 的普通阴性选择算法)时的检测率是最低的,说明在相同 N_{R_0} 下, r 可变阴性选择算法的检测率高于普通阴性选择算法.

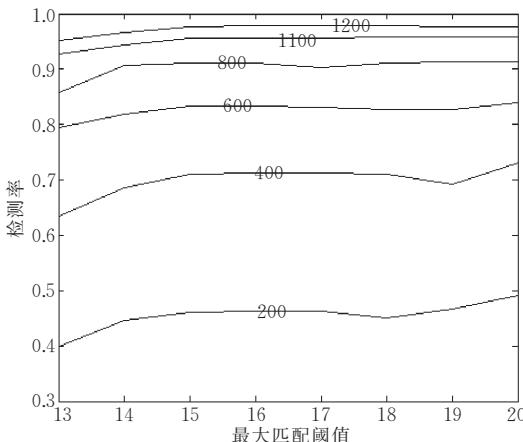


图 4 不同 N_{R_0} 下最大匹配阈值与检测率的关系

5.4 黑洞数目的变化

在 r 可变阴性选择算法中,黑洞的数量取决于最大匹配阈值 r_c 的值. 根据 D'haesleer 给出的基于连续位匹配规则的计算黑洞数量的算法,计算了在采集数据集 CDS 情况下的黑洞数量,如图 5 所示.

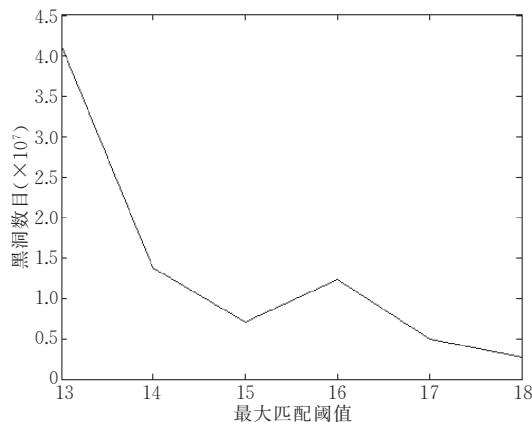


图 5 在 CDS 下不同最大匹配阈值的黑洞数量

黑洞数量随着最大匹配阈值的增加迅速下降,这是由于如图 6 所示,较大匹配阈值的检测器的加入(它们的检测范围缩小),一些原本是黑洞的模式也可被检测到. $r_c=15$ 时的黑洞数量是在 $r_c=13$ (即普通阴性选择算法)时的 17.2%. 在图 5 中当 $r_c=16$ 时的黑洞数量又有所上升,这是由于自我集模式的分布特点导致的.

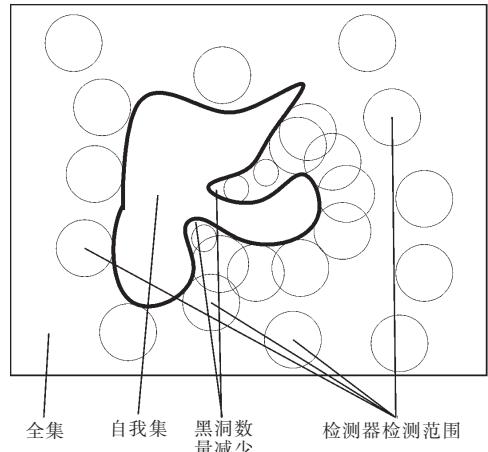


图 6 不同匹配阈值检测器减少黑洞数目直观图

6 总 结

作为人工免疫系统中的核心算法之一的阴性选择算法,其性能对整个系统具有重要意义. 为减少算法导致的“黑洞”数量,本文提出一种 r 可变阴性选择算法,算法的核心思想是通过调整匹配阈值的方法大幅度降低黑洞数量. 仿真结果表明,这种算法产生成熟检测器的迭代次数、黑洞数量均大幅下降,检测率则有提高;由于产生成熟检测器的迭代次数下降,产生检测器的时间耗费减小.

参 考 文 献

- 1 Forrest S., Hofmeyr S. A. Immunology as information processing. In: Segel L. A., Cohen I. eds. Design Principles for the Immune System and Other Distributed Autonomous Systems. New York: Oxford University Press, 2000, 361~387
- 2 Hofmeyr S. A., Forrest S.. Immunity by design: An artificial immune system. In: Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, CA, 1999, 1289~1296
- 3 Forrest S., Hofmeyr S. A., Somayaji A., Longstaff T. A.. A sense of self for unix processes. In: Proceedings of the 1996 IEEE Symposium on Security and Privacy, Los Alamitos, CA, 1996, 120~128
- 4 Forrest S., Perelson A. S., Allen L., Cherukuri R.. Self-nonself discrimination in a computer. In: Proceedings of the 1994 IEEE Symposium on Security and Privacy, Los Alamitos, CA, 1994, 202~212
- 5 D'haesleer P., Forrest S.. An immunological approach to change detection algorithms analysis and implications. In: Proceedings of the 1996 IEEE Symposium on Security and Privacy, Los Alamitos, CA, 1996, 110~119

- 6 D'haeseler P.. Further efficient algorithms for generating antibody string. The University of New Mexico, Albuquerque, NM: Technical Report CS95-03, 1995
- 7 Percus J. K. , Percus O. , Perelson A. S.. Probability of self-nonsel discrimination. In: Perelson A. S. , Weisbuch G. eds.. Theoretical and Experimental Insights into Immunology. New York: Springer-Verlag Press, 1993, 63~70
- 8 Percus J. K. , Percus O. , Perelson A. S.. Predicting the size of the antibody-combining region from considering of efficient self/nonsel discrimination. In: Proceedings of the National Academy of Science, Washington, 1993, 1691~1695
- 9 Long Zhen-Zhou. Medical Immunology. Beijing: People's Medical Publishing House, 1995(in Chinese)
- (龙振洲. 医学免疫学. 北京:人民卫生出版社, 1995)
- 10 Zhou Jian-Guo. An immunological model of network intusion detection and its simulation research [Ph. D. dissertation]. Beihang University, Beijing, 2002(in Chinese)
(周建国. 网络入侵检测的免疫学建模及其仿真研究[博士学位论文]. 北京航空航天大学,北京, 2002)
- 11 Hofmeyr S. A.. An immunological model of distributed detection and its application to computer security [Ph. D. dissertation]. University of New Mexico, Albuquerque, NM, 1999
- 12 Wright C. , Cowan C. , Morris J. *et al.*. Linux security modules: General security support for the linux kernel. In: Proceedings of USENIX Security Symposium, San Francisco, CA, 2002, 17~31



ZHANG Heng, born in 1977, Ph. D. candidate. His research interests include security operation system, artificial immune system.

Background

The paper is supported by the National High Technology Research and Development Program of China (863 Program) under grant No. 2002AA141090 (Research and Development on Security Key Technology of Server), No. 2004AA147070 (“Research and Development on Security Evaluation of System Platforms”) and the National Natural Science Foundation of China under grant No. 60473053 (“Study on Semantic Constraint Approach to Control Program’s Behavior”).

The security of program roots from two ultimate problems: One is the behavior of program cannot abide by the intention of its designer, another is the access control of program cannot abide by the least privilege principle. In those research projects, static and dynamic methods, confining and monitoring technology are combined to turn the protection of program from passive to active. During the research process, immune mechanism is applied to the modeling and control of the program behavior, and also applied to the automated modeling and test of security function.

In recent years, the main work concerned with security

WU Li-Fa, born in 1968, professor. His research interests include network management, network security.

ZHANG Yu-Sen, born in 1949, professor, Ph. D. supervisor. His research interests include system simulation, information security.

ZENG Qing-Kai, born in 1963, professor. His research interests include network security, distributed computing.

that finished by authors' research team list as follow: developed an applied distributed network intrusion detection system “NetNumen” which based on rule and can combine anomaly detection with misuse detection, developed an applied military security operation system which can reach B security level. Now the research team focuses on security test and embed operation system.

The immune mechanism is used to solve the key problem of program security which is security state monitor and get nice results. But one of the important algorithms in immune mechanism named negative selection algorithm exists hole which can lead to False Negatives. In this paper, the authors applied a simple way through adjusting the matching threshold to reduce the number of holes. The simulation which adopts data got from experiment shows that both the number of iteration to generate mature detector and the number of holes decline quickly, while the rate of detecting abnormality raises.