一种改进的变阈值阴性选择免疫算法

翟宏群,冯茂岩

(江苏海事职业技术学院 信息工程系,江苏 南京 211170)

[摘要] 成功确定一个最有效检测元集是提高免疫阴性选择算法性能的关键步骤,它直接影响到系统的效率和准确度. 利用模糊思想,提出了一种生成最有效检测元集的变阈值阴性选择免疫算法. 采用最优搜索原理,有效提高了待检测的检测元成为成熟检测元的概率;匹配阈值可变,可大幅降低黑洞数量. 仿真结果表明,该算法与原算法相比,具有较高的检测率和较少的黑洞数量,算法具有较强的鲁棒性.

[关键词] 阴性选择,优先搜索,检测元集,黑洞

[中图分类号] TP309 [文献标志码] A [文章编号] 1672-1292(2011)03-0078-05

An Improved Adjustable Threshold Intrusion Detection Negative Selection Immune Algorithm

Zhai Hongqun, Feng Maoyan

(Information Department, Jiangsu Maritime Institute, Nanjing 211170, China)

Abstract: Success in confirming the most effective detector set is a key step to improve negative selection algorithm capability, which has a direct affect on efficiency and veracity of system. Fuzzy idea was used to put forward an adjustable threshold negative selection immune algorithm of creating the most effective detector set. The rate of mature detector activated can be improved effectively based on optimal search theory and the number of black holes can be reduced clearly through adjusting matching threshold in this algorithm. The simulation results indicate that this new algorithm in comparison with the original algorithm, is of higher detection efficiency and lower detection holes number, and thus the algorithm has better robustness.

Key words: negative selection, optimal search, detector set, black holes

生物免疫系统(Biological Immune System, BIS)是一个高度并行、分布、自适应和自组织的系统. 人工免疫系统(Artificial Immune System, AIS)是一类基于生物免疫系统功能、原理、基本特征以及相关理论免疫学说而建立的用于解决各种复杂问题的计算系统[1,2]. 随着人们对免疫系统认识的不断深入,更多的免疫机制将得到应用. 2008 年召开的第七届人工免疫系统国际会议收录的一些文章已经体现了这一趋势[3,4].

阴性选择算法(Negative Selection Algorithm)便是参照生物免疫系统阴性选择原理提出的仿生学算法,是人工免疫系统开发的核心算法之一[5]. 阴性选择算法的一个主要优点就是不需要有先验知识,可以对未知入侵模式进行有效的防御,但也存在不足:存在某些非我字符串,找不到有效的检测元发现它,导致所生成的检测元集合并不能完全覆盖所有可能的“非己”空间,形成检测黑洞.

由于黑洞不可能被任何检测元所检测到,所以应尽可能地减少黑洞的数量及其出现的概率. 本文对阴性选择算法进行了深入研究,利用模糊思想,提出一种改进的变阈值阴性选择免疫算法. 算法中,采用最优搜索原理,有效地提高了待检测的检测元成为成熟检测元的概率;匹配阈值的不断改变,成功地解决了传统阴性选择算法不可避免的“检测黑洞”数量问题.

收稿日期:2011-03-20.
基金项目:江苏省“网络与信息安全”重点实验室课题(BM2003201).
通讯联系人:翟宏群,讲师,研究方向:计算机应用技术、网络安全. E-mail: hqzhai@126.com

1 阴性选择算法及其分析

阴性选择算法采用基于免疫系统中的自体和异体识别的原理来进行变化监测,先随机产生检测器,然后删除与自身对抗的细胞并最终保留能检测异体的细胞. 如图 1 所示,算法过程如下所述:

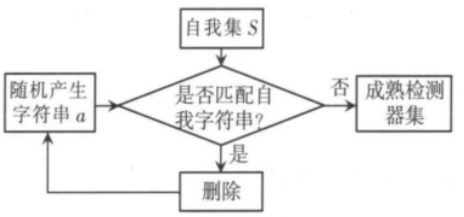

图 1 阴性选择算法流程示意图

Fig.1 Negative selection algorithm flow diagram

步骤 1 首先生成长为 L 的预检测器,然后与自我集 S 按照匹配规则进行匹配,若匹配成功,则删除; 否则,放入成熟检测器集合;

步骤 2 重复以上过程,直至生成预定数量的成熟检测器;

步骤 3 成熟检测器用于匹配待检测串,若匹配,则表明发生了异常变化.

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,不足的是所生成的检测元集合并不能完全覆盖所有可能的"非己"空间,存在着"检测黑洞"问题.

为清晰地阐述"黑洞"问题,引入以下记法, U_U 表示所有模式串组成的集合, N_S 表示自我模式串集合(Self 集) , N_N 表示非我模式串集合(Non-self 集) , N_U 表示所有字符串组成的集合.

定义1 黑洞: 一非我模式串 $a \in N_N$ 是一个黑洞,当且仅当 $\forall s \in U_U$, Match(s, a) 成立,其中, $s \in N_S$. 即黑洞就是不能产生相应的检测元来检测到的非我模式串.

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. 周建国[6] 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[7] 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.

2 变阈值阴性选择免疫算法

为了减少黑洞的数量,本文提出一种变阈值阴性选择免疫算法. 为了更好地描述该算法,首先给出以下几个相关定义:

定义 2 字符串的 r-模板是指长度为 l 的字符串,在 r 个连续的位置上具有确定的字符,而在其余的 $l - r$ 个位置上为通配符 * 的字符串.

定义3 第 i 模板是指从第 i 个位置上开始有确定字符的模板. 例如, $**0111****$ 就是一个在长度 $l = 10$ 的字符串的 4-模板,同时也是一个第 3 模板.

定义4 一个模板 t_r 与一个模式串 s 匹配,当且仅当在模板的 r 个已定义字符的连续位置上与 s 匹配,记为 Match(t_r, s) ,否则不匹配,记为 ¬ Match(t_r, s) . 例如, $**0111****$ 与 1101110100 匹配,则 Match($**0111***$, 1101110100) .

2.1 算法描述

(1) 构造一个与模式串长度相同的 r-模板 t_r, t_r 匹配非我模式串 a,但不匹配任何一个自我模式串,即 Match(t_r, a) , ¬ Match(t_r, s) , $s \in N_S$,如果不能构造出这样一个模板,则 a 为黑洞.

(2) 利用 t_r 作为匹配模板,在由所有匹配 t_r 的字符串组成的空间进行深度优先搜索,为非我模式串 a 构造一个有效的检测元. 预先确定搜索树的长度 C,即确定搜索节点数. 每次搜索,将一个确定的字符 a_i 增加到模板上,并首先将其初始化为 $a_i = 1$,这时的模板为 $t_r a_i$,如果此时的模板不与任一自我模式串 $s \in N_S$ 匹配,则表明模板是有效的,继续搜索; 反之,初始化 $a_i = 0$,继续搜索.

(3) 重复步骤(1) 和(2) ,直到构造成功一个有效的检测元,如果在预定搜索长度 C 内不能构造出一个有效的检测元,则可认为 a 为一个黑洞.

2.2 算法的伪代码

```
#define 搜索树的长度( 即搜索节点数) 为某一常数 C;
int c = 0; // 计数器,为全局变量
```

```
char a_i, t_r [C + 1], a ∈ N_R, s ∈ N_S;
while( c < k) {
    scan( "%s", t_r);   // 输入与模式串等长的 r-模板 t_r
    if( !( Match( t_r, a) ∧ ¬ Match( t_r, s) ) )
    print( "%s 为黑洞 \n", a);
    else{ /* 利用 t_r 作为匹配模板,深度优先搜索出所有匹配 t_r 的字符串组成空间,为 a 构造一个有效的检测元. 每次搜索,将一个确定的字符 a_i 增加到模板上,这时的模板为 t_r a_i. * /
    c ++;
    a_i ← 1;
    t_r ← t_r a_i;
    if( ¬ Match( t_r, s) ) //t_r 有效,向左继续搜索;
    else{
    c ++;
    a_i ← 0;
    t_r ← t_r a_i;
    if( ¬ Match( t_r, s) ) //t_r 有效,向右继续搜索;
    }
}
} //while
if( Match( t_r, a) ∧ ¬ Match( t_r, s) )
    print( "%s 为一有效的检测元 \n", t_r);
else
    print( "%s 为一黑洞 \n", a)
```

2.3 算法的示例说明

假设模式长度 $l = 10$,Self 模式集 $N_S = \{0100111100, 1000111111, 1100110100, 0010010011\}$. 设 Non-self 模式串 $a = 0010110100$,匹配阈值 $r = 3$,利用上述算法判断 Non-self 模式串 a 是否为一黑洞.

首先构造一个匹配 Non-self 模式串 a,但不匹配任一 Self 模式串的 3-模板. $t_r = **101*****$ 就是这样一个模板,由于存在有这样的模板,故进行如图 2 的搜索,搜索结果产生 **10101010 即为一个有效的检测元,故 Non-self 模式串 a 不是黑洞.

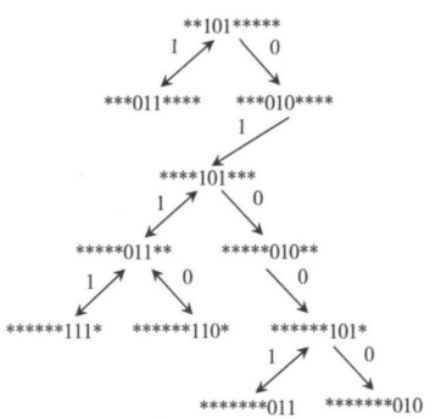

图 2 判断 Non-self 模式串是否为黑洞的算法示意图

Fig.2 The algorithm diagram whether Non-self model to be the black hole

2.4 变阈值策略

由于免疫系统的多样性机理,要使尽可能多样的抗体对抗千变万化的抗原,抗体必须具有泛化能力和联想记忆能力,也就是抗体与抗原之间并不是绝对的一一对应的关系,抗体与抗原之间的匹配具有不确定性和模糊性.

本算法利用此思想并结合模糊思想来确定检测元之间是否匹配,匹配的模糊程度由匹配阈值 r 来确定,通过调整匹配阈值 r 的方法来大幅度降低黑洞数量. 首先设定初始匹配阈值 r_1,当模糊相似度超过 r_1 时,调整匹配阈值. 算法中匹配阈值的调整策略为 $r_i = r_{i-1} + 1$,最大匹配阈值 $r_{\max} = C - 1$(C 为模式串长度). 与普通阴性选择算法相比,匹配阈值可变使得由本算法产生的不同检测元检测范围不同,产生的检测元集中不仅有检测元,还包括其对应的匹配阈值 r,且匹配阈值 r 可调,而普通阴性选择算法的匹配阈值 r 和检测元的检测范围均是固定的. 由此可见,普通阴性选择算法是变阈值阴性选择免疫算法的一个特例.

3 仿真分析

本文对所改进的算法进行了仿真,分析采用变阈值检测元产生算法对成熟检测元分布及黑洞数目的

变化情况. 仿真数据使用 randerr(1 000,32, [0:32]) 函数随机生成长度为32的二进制串的模式集合,首先随机生成 1 000 个长度为 32 的二进制串的模式集合,作为自我集.

3.1 匹配概率分析

两个随机串在 r-连续位匹配规则下的匹配概率为 P_M

$$P_M = (l - r + 1) \cdot 2^{-r} - (l - r) \cdot 2^{-r-1} = 2^{-r}\left(\frac{l - r}{2} + 1\right). \tag{1}$$

其中,l 为字符串的长度. 则一随机串不和自我集匹配的概率为:

$$f = (1 - P_M)^{N_S}. \tag{2}$$

其中,N_S 为自我集的大小. 由分析可知,漏报率 P_f 满足:

$$P_f = (1 - P_M)^{N_R}. \tag{3}$$

其中,N_R 为由算法生成的最小有效检测元数目. 则检测率为:

$$P_S = 1 - P_f. \tag{4}$$

对式(3) 两边同时取对数运算,可得最小有效检测元的数量 N_R 为:

$$N_R = \frac{\ln P_f}{\ln(1 - P_M)} = \frac{\ln P_f}{\ln(1 - (l - r + 1) \cdot 2^{-r} + (l - r) \cdot 2^{-r-1})}. \tag{5}$$

从式(3) 可以看出,对于固定的 P_M 和 P_f,N_R 与 N_S 无关,即有效检测元的数量不必随被保护内容的增大而增加. 这意味着一定大小的检测元集合可以有效地保护较大的自我集 N_S. 系统的漏检率 P_f 越低,即系统检测能力越强,所需的检测元集合就越大,会导致系统效率降低(计算代价增加).

全集 $N_U = 2^l$,检测元集在全集中可以覆盖的最大空间为:

$$B_S = N_R \times (2^l - T(l)), \tag{6}$$

其中 $T(l)$ 为:

$$T(l) = \begin{cases} 2^l - 2^{l-r} - (l - r) \cdot 2^{l-r-1} & \text{当 } r \leqslant l \leqslant 2r, \\ 2T(l-1) - T(l-r-1) & \text{当 } l > 2r. \end{cases} \tag{7}$$

对于给定的漏检率,为了足以覆盖所能检测到的“非己”空间,则 B_S 应满足 $B_S \geqslant P_S \cdot (N_U - N_S)$.

3.2 不同匹配阈值的检测元分布

由于该算法匹配阈值有多个,使得检测元的数目增加,为了分析不同匹配阈值的检测元分布情况,随机生成 10 000 个模式串,在不同 r_c 下作仿真实验,结果如表 1 所示.

表 1 不同匹配阈值的检测元的分布(占成熟检测元的比例)

Table 1 The distribution of the detector with different threshold matching (The proportion of total mature detector)

	r = 14	r = 15	r = 16	r = 17	r = 18	r = 19	r = 20	r = 21
r_c = 14	4 630 46%							
r_c = 15	4 552 45%	799 7.9%						
r_c = 16	4 534 45%	785 7.8%	646 6.4%					
r_c = 17	4 561 45%	741 7.4%	649 6.4%	504 5.0%				
r_c = 18	4 613 46%	755 7.5%	613 6.1%	512 5.1%	421 4.2%			
r_c = 19	4 604 46%	741 7.4%	635 6.3%	552 5.5%	387 3.8%	369 3.6%		
r_c = 20	4 520 45%	728 7.2%	649 6.4%	543 5.4%	398 3.9%	332 3.3%	248 2.4%	
r_c = 21	4 552 45%	799 7.9%	680 6.8%	482 4.8%	382 3.8%	349 3.4%	274 2.7%	196 1.9%

从表 1 看出,不同匹配阈值的检测元占未成熟检测元的比例基本固定,匹配阈值小的检测元数目较多,因而检测到的“非己”字符串数目就越多,它的专一性就越差; 匹配阈值大的检测元所占比例小,因而

它能够检测到的“非己”字符串数目较少,专一性较强. 当 $r = 32$ 即模式串的长度时,检测元只能检测到一个“非己”字符串,专一性最强. 因此该算法实现了以较小的检测元集合,检测到较大范围的“非己”行为.

3.3 黑洞数目的变化情况

在具有最小检测元集的变阈值免疫算法中,黑洞的数量取决于最大匹配阈值 r_{max} 的值. 根据 D'haesseleer[8] 给出的基于连续位匹配规则的计算黑洞的算法,计算了不同最大匹配阈值下黑洞数量,如图3所示.

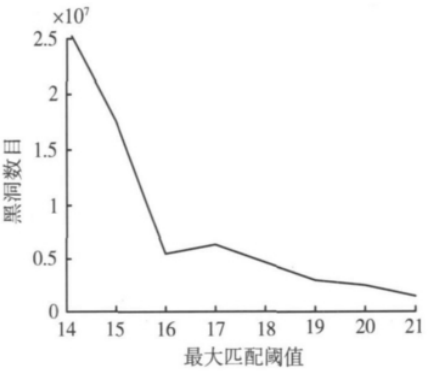

图3 不同最大匹配阈值的黑洞数

Fig.3 The number of black holes with different maximum matching threshold

从图3可以看出,黑洞数量随着最大匹配阈值的增加而迅速下降,这是由于较大匹配阈值的检测元的加入,而较大匹配阈值的检测元检测范围缩小,使得一些原本是黑洞的模式也可以被检测到. 在图3中,当 $r_c = 17$ 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.

4 结语

本文通过对人工免疫系统中阴性选择算法机理的分析,利用模糊思想,提出了一种生成最有效检测元集的变阈值阴性选择免疫算法. 算法借助深度优先搜索,有效地提高了待检测的检测元成为成熟检测元的概率,产生的有效检测元集中不仅有检测元,还包括其对应的匹配阈值 r,且匹配阈值 r 可调,检测元的数目增加,检测的范围扩大,成功地解决了传统阴性选择算法不可避免的“检测黑洞”数量问题. 仿真结果表明,该算法与普通阴性选择算法相比,具有较高的检测率和较少的黑洞数量,空间覆盖率明显提高,实现了以较少的检测元集合,检测到较大范围的“非己”行为.

[参考文献](References)

[1] Hofmeyr S, Forrest S. Immunity by design: An artificial immune system [C] // Wolfgan B, Jason M D, et al, eds. Proc of the Genetic and Evolutionary Computation Conf. San Francisco: Morgan Kaufman Publishers, 1999.

[2] De Castro L N, Timmis J. Artificial Immune Systems: A New Computational Intelligence Approach [M]. Heidelberg: Springer-Verlag, 2002.

[3] Kelsey J, Hender S B, Seymour R M. A stochastic model of the interleukin (IL) 21 B network [C] //Proceeding of the 7th International Conference on Artificial Immune Systems. Phuket: Springer, 2008: 1 211.

[4] Andrews P S, Timm I S J. Adaptable lymphocytes for artificial immune systems [C] //Proceeding of the 7th International Conference on Artificial Immune Systems. Phuket: Springer, 2008: 3 762 386.

[5] 张宇. 人工免疫系统中阴性选择算法的研究 [D]. 杭州: 浙江大学电气工程学院, 2007.
Zhang Yu. Research on negative selection algorithm of artificial immune system [D]. Hangzhou: School of Electrical Engineering, Zhejiang University, 2007. (in Chinese)

[6] 周建国. 网络入侵检测的免疫学建模及其仿真研究 [D]. 北京: 北京航空航天大学计算机学院, 2002.
Zhou Jianguo. Immunological moding of network intrusion detection & its simulate research [D]. Beijing: School of Computer, Beijing University of Aeronautics and Astronautics, 2002. (in Chinese)

[7] Hofmeyr S A. An immunological model of distributed detection and its application to computer security [D]. Albuquerque, NM: Computer Science Department, University of New Mexico, 1999.

[8] D'haesseleer P. Further efficient algorithms for generating antibody string, Technical Report CS95-03 [R]. New Mexico: The University of New Mexico, 1995.

[责任编辑: 严海琳]