基于小生境技术的多样性抗体生成算法

凌 军[1],曹 阳[1,2],尹建华[1],徐国雄[1],黄天锡[1]
(1. 武汉大学电子信息学学院,湖北武汉 430079;2. 武汉大学软件工程国家重点实验室,湖北武汉 430072)

摘 要: 新的入侵方法以及网络计算环境的不断变化,使得入侵越来越难以防范.因此应用人体免疫机制构建下一代入侵检测系统成为一个新的研究热点.本文采用一种新的抗原 - 抗体编码方法,并在此基础上,提出一种基于共享函数的小生境遗传算法,用来产生多样性的抗体基因,最后给出多样性评价函数以验证算法的有效性.实验结果表明:该算法能够使抗体在演化过程中保持较好的多样性.

关键词: 免疫系统; 抗原-抗体; 小生境; 多样性
中图分类号: TP393 **文献标识码**: A **文章编号**: 0372-2112 (2003) 08-1130-03

An Algorithm for Generating Diverse Antibody Based on Niche Technology

LING Jun[1], CAO Yang[1,2], YING Jian-hua[1], XU Guo-xiong[1], HUANG Tian-xi[1]
(1. *College of Electronic and Information Wuhan University, Wuhan, Hubei* 430079, *China*;
2. *State Key Laboratory of Software Engineering Wuhan University, Wuhan, Hubei* 430072, *China*)

Abstract: The ever-rising complexity of operating system and communication networks has resulted in increased difficulty in detecting intrusions. So applying immune mechanism of human body to construct next generation intrusion detection becomes a new research focus. This paper presents a novel method to encode the antibody-antigen. And An generic algorithm based on niche technology of share function for generating diverse antibody is provided. In order to verify the validity of the algorithm, two evaluation function are suggested. The experimental results reveal that the algorithm has much to offer to keep population's diversity during evolution.

Key words: immune system; antigen-antibody; niche; diversity

1 引言

近年来,网络已应用到社会经济、人民生活的各个领域,人们在得益于信息革命所带来的巨大机遇的同时,不得不面对信息安全的严峻考验:近年来网上攻击事件层出不穷,给计算机网络安全造成极大威胁.因此,确保网络和计算机系统免受攻击、安全而稳定的运行已成为一个重要的研究课题.

入侵检测研究面临巨大的挑战主要表现在:(1)攻击目标复杂多变,攻击工具更加高级,攻击能力大大提高;(2)日益增加的网络业务类型;(3)入侵检测系统本身易受到攻击;(4)网络业务量过大,难以进行实时分析;(5)网络结构更加复杂.

过去二十年来,借鉴生物学的理论来解决计算问题越来越引起人们的关注.其基本原理是借用自然生物系统的概念和机理来构建一种解决方案,用以解决其他领域的计算问题.最为著名的有借鉴大脑工作原理的神经网络和借鉴达尔文进化论的演化计算.

人体免疫系统成功地保护身体免受众多病原体和外部组织的侵害.免疫系统的许多特性,如分布式、多层次保护、识别能力、学习和记忆能力,对于设计新一代网络安全系统具有重要的借鉴意义.许多计算机安全研究者和人工智能研究者对免疫系统进行深入研究,提出了多种计算机免疫模型[1~3].

在免疫系统中,当免疫细胞表面的受体分子(抗体)和抗原表面的抗原决定基之间发生化学结合,则说明抗体识别出抗原[4].为了识别大量的外部抗原,免疫系统必须生成大量抗体.但是,免疫系统的资源有限,不可能为每一种可能遇到的抗原产生一种抗体.因此,如何生成大量多样性的抗体是入侵检测免疫系统成功的关键.

本文首先提出一种新的抗原 - 抗体编码方法.在此基础上,提出了一种基于小生境技术的多样性抗体生成算法,并给出一种多样性评价函数.实验结果表明:该算法能够有效地保持抗体集的多样性.

2 抗原-抗体的表示

2.1 特征集

设计抗体-抗原特征结构和表示是构建入侵检测免疫模型的关键.特征集是对网络行为的抽象.特征集的选取直接影响入侵检测免疫模型的性能.特征集的选取必须遵循以下原则:

收稿日期:2002-12-11;修回日期:2003-04-10
基金项目:国家自然科学基金(No. 69983005, No. 60132030);国家教育部博士点基金(No. RFDP1999048602)

(1)特征集必须能够精确描述网络行为;

(2)特征集能够区分本体(self,表示正常的网络行为)和异体(nonself,表示异常的网络行为)的行为;

(3)特征集能够容易编码.

因为本文研究的是基于网络的入侵检测系统.而网络流量的基本单元是数据包,所以所使用的抗体-抗原特征根据数据包的特征来构建.在众多文献中,网络入侵检测系统所使用的数据包特征各不相同,例如,新墨西哥大学的 LISYS 使用了三种特征:源 IP 地址、目的 IP 地址和 TCP 端口[4].而 snort 系统使用了超过 30 种网络数据包特征来检测攻击.

以上系统使用单个数据包特征,也取得很好的成效.但是,网络数据包却往往表现出以下特性:(1)随机性:由于网络业务量的复杂多变以及用户访问网络的随意性,使得这些数据呈现很强的随机性;(2)相关性:单纯的网络事件(例如 telnet、ftp、www 访问等)往往不能完整地反映网络状况,但在较长的时间范围内表现出较强的相关性.因此,对网络数据的处理,不能仅仅孤立地对网络事件进行处理,而必须综合考虑一个时间段范围内的特征,才能真正反映它们的具体属性.因此,本文采用基于时间窗口的形状空间表示抗体-抗原.下面给出抗体-抗原形状空间的一些定义.

定义 1　可测度集:设 F 为一组属性值集合,包含所有审计数据或连接记录中可能出现的属性值,例如源/目的地址、连接标识等,称之为可测度集.设 $V=\{filed_1, field_2, \cdots filed_n\}$ 为 F 的子集,称之为记录属性.

可测度集表明一个数据包或一条连接的属性集合,例如网络连接的可测度集为 $(timestamp, service, src\_host, dst\_host, src\_port, dst\_port, flag)$,分别表示连接的时间戳、服务类型、源主机、目的主机、源端口、目的端口,标志等属性.其中有一些属性对于准确描述连接的性质很重要,例如源主机,还有一些属性只是辅助信息.网络连接可以由一个五元组组成:$\langle timestamp, src\_host, src\_port, dst\_host, dst\_port\rangle$,我们将重要的信息作为记录属性.

定义 2　事件 $E:V$ 的实例化.形式为:$B=\{(field_1, value_1), (filed_2, value_2)\cdots(field_n, value_n)\}$.其中 $value_i(1\leq i\leq n)$ 为 $field_i$ 类型的值.记录属性的实例化称之为事件,表 1 给出一些事件的实例.

表 1　网络事件

Timestamp	*Src_host*	*Src_port*	*Dst_host*	*Dst_port*
11:41:40	192.168.0.1	4561	192.168.0.10	80
11:41:45	192.168.0.7	4442	192.168.0.1	23
…	…	…	…	…
12:45:40	192.168.0.4	5634	192.168.0.10	80

网络数据包形成连接记录后,连接信息主要由时戳、源/目的地址(主机)、源/目的端口(服务类型)组成,这些是网络数据的必要属性.为了挖掘属性之间的关联规则,并尽可能地减少规则冗余,引入参考量集的概念,用以表示网络数据中重要的属性.参考量集一般由源/目的地址(主机)、源/目的端口(服务类型)组成.例如对于一种拒绝攻击,可以设定端口为参考量,然后计算其他属性(例如源地址)与参考量的关系,如果这种关系满足一定的规则(例如源地址数多于最大值),则可认为是发生拒绝攻击.由此可见,参考量集的选择十分重要.

定义 3　参考量集,设 R 为可测度集 F 的子集,其中包含的元素为可测度集的重要特征,称之为参考量集.

定义 4　时间窗口上下文 TC,结构定义为 $(R, F, \mathrm{Rel}(R, F), Length)$,其中 R 为参考量集,$R=\{r_1, r_2, \cdots, r_m\}$,$F$ 为可测度集,$F=\{f_1, f_2, \cdots, f_m\}$.$\mathrm{Rel}(R, F)=\{R(r_i, f_j) \mid 1\leq i\leq n, r_i\in R, 1\leq j\leq n, f_j\in F\}$ 描述 r_i 和 f_j 之间的定量关系.

根据定义 4 可知,时间窗口上下文的 $\mathrm{Rel}(R, F)$ 具有 $m\times n$ 个不同的项,如果 R 和 F 确定的话,$\mathrm{Rel}(R, F)$ 的项数也为常数,因此,抗原-抗体可用时间窗口上下文表示,表征在一个时间范围内的网络特征.

2.2　基因编码

根据 2.1 节,抗原-抗体由时间窗口上下文表示,而时间窗口上下文的 $\mathrm{Rel}(R, F)$ 具有 $m\times n$ 个不同的项,因此,可将每一项视为组成抗原-抗体的基因.基因以二进制串的形式编码.在本文中,选取二进制基因编码的长度为 10.时间窗口中上下文的各项为十进制数,可以通过公式 1 将二进制基因转化为十进制的时间窗口上下文各项:

$$R=\sum_{i=1}^{L} bit(i)\times 2^i \qquad (1)$$

其中,R 表示时间窗口上下文某一项,$bit(i)$ 为二进制基因的第 i 个二进制位的值.

3　基因多样性的生成

在入侵检测免疫模型中,为了以较小的资源检测更多的异体,必须确保在免疫模型进化过程中保持抗体的多样性,如何产生多样性的抗体是入侵检测免疫模型的关键.Oprea 等人提出采用遗传算法生成多样性的抗体[5],但是该算法需要多样性的抗原作为训练集.本文提出一种基于共享函数的小生境遗传算法,用来产生多样性的抗体基因.基于共享函数的小生境遗传算法基本思想是:在解空间中,一些相邻的个体组成小生境.通过计算种群中各个体之间的距离,确定某个个体周围相邻个体的数目(即小生境个体数),将该小生境所有个体的适应值按照小生境规模以一定的方式降低,显然,如果某个小生境中有较多的个体,那么该小生境中所有个体的适应值以较大的幅度降低.因此,小规模生境的被选择的概率会有所提高,从而维持群体的多样性[6].根据上述分析,本算法根据种群内部个体之间的相互关系指导演化趋势,不需要多样性的抗原作为训练集合.

3.1　基本原理

设 $d(i, j)$ 为个体 i 和 j 之间的距离,L 为个体二进制位数,其计算公式为:

$$d(i, j)=\sum_{k=1}^{L} |bit(i, k)-bit(j, k)| \qquad (2)$$

其中, $bit(i,k)$ 表示第 i 个个体第 k 个二进制位的值.

个体 i 和个体 j 之间的共享函数为:

$$sh(d(i,j)) = \begin{cases} 1 - \dfrac{d(i,j)}{\delta} & , d(i,j) < \delta \\ 0 & , d(i,j) > \delta \end{cases} \tag{3}$$

其中, δ 为事先指定的峰半径. α 为控制共享函数的形状的参数,通常 $\alpha = 1$,即为线性共享函数,如果 $\alpha > 1$,为凹函数;如果 $\alpha < 1$,为凸函数. 得到所有个体的共享值之后,可以由以下公式计算个体的小生境个体数.

$$m_i = \sum_{j=1}^{N} sh(d(i,j)), \qquad i = 1,2,3,\cdots,N \tag{4}$$

其中, N 为种群中个体的数目. 显然,个体的小生境个体数越大,聚集在该个体周围的个体就越多. 然后,计算共享后个体的适应值:

$$f_i' = f_i / m_i \tag{5}$$

其中, f_i 为共享前的个体适应值. 演化过程中,使用共享后的个体适应值,如果某个物种有较多的个体,那么该物种中个体将以较大的幅度降低,从而鼓励较少个体的物种繁衍.

3.2 算法步骤

基于小生境技术的多样性抗体生成算法的出发点是在选择过程中采取一定的策略保持种群的多样性,本算法的基本内容是在标准演化算法的基础上,利用适应值共享的思想对适应值进行调整. 算法步骤如下:

(1) 初始化种群 $X = \{Xt(0), X_t(1), Xt(2), \cdots\cdots, X_t(N)\}$, $t = 0$;

(2) 根据公式(2)计算个体之间的距离 $d(i,j)$;

(3) 根据公式(3)计算个体之间的共享函数值 $sh(d(i,j))$;

(4) 根据公式(4)计算每一个体所在小生境的个体数 m_i;

(5) 根据公式(5)计算并设定每一个体共享后的适应值 f_i';

(6) 通过选择、交叉、变异操作,产生新的种群 X_t;

(7) 若满足停止条件,则停止;否则,转第(2)步.

4 结果分析

为了对上述算法的有效性进行验证,本文提出两种多样性度量方法.

4.1 基于基因的度量方法

根据 2.2,设第 t 代种群的个体 $X_t(I) = (x_t(I_1), x_t(I_2), x_t(I_3), \cdots, x_t(I_L))$ (串长为 L 的二进制编码,其中, $I = 1,2,3,\cdots,N$. N 个个体组成的矩阵为:

$$P_{X\times L} = \begin{bmatrix} x_t(11) & x_t(12) & \cdots & x_t(1L) \\ x_t(21) & x_t(22) & \cdots & x_t(2L) \\ \cdots & \cdots & \cdots & \cdots \\ x_t(N1) & x_t(N2) & \cdots & x_t(NL) \end{bmatrix}$$

如果上述矩阵每列中 0 和 1 各占一半,在交叉操作中一方面可以避免有效基因的缺失,另一方面可以以较大的概率产生新的个体,这样,若在矩阵 $P_{X\times L}$ 中每列的 0 和 1 都趋向于 $N/2$,则种群的多样性就越好.

根据上述分析,种群多样性度量函数 $d(p)$ 可定义如下:

$$d(p) = \frac{1}{L\times N}\sum_{j=1}^{L}\max\left\{\sum_{i=1}^{N}(1 - a_{i,j}), \sum_{i=1}^{N}a_{i,j}\right\} \in [0.5,1] \tag{6}$$

其中, N 为种群规模, L 为个体的二进制编码长度, $a_{i,j}$ 表示种群中第 i 个个体第 j 个二进制位的值. $d(p) \in [0.5,1]$ 表示每个等位基因位置上居较多数量的二进制位的平均百分率. 如果每列中 0 和 1 各占一半,为 $N/2$, $d(p) = 0.5$,表明种群的多样性就越好;如果每列各等位基因位置上二进制位都相同,则 $d(p) = 1$,种群的多样性最差. 显然, $d(p)$ 越小,种群的多样性就越好.

为了对多样性抗体算法进行测试,我们设计并开发了一种网络入侵免疫系统(NIIS)原型,在该系统中进行多样性抗体生成算法的测试,基本步骤如 3.2 节所示,算法的参数为:群体个体数为 100,演化代数为 200,变异概率为 0.001,交叉概率 0.9. 根据公式(6)计算演化每一代的 $d(p)$ 值,结果如图 1 所示:

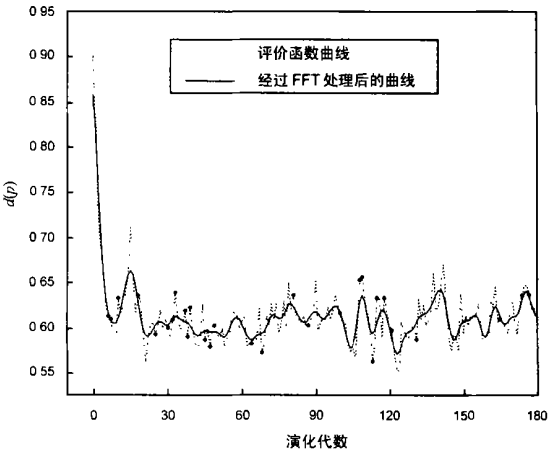

图 1 多样性评价函数曲线

在图 1 中,虚线为评价函数曲线,实线为经过 FFT 平滑处理的曲线,由图 1 可知,开始演化后,多样性度量函数值明显降低(多样性明显提高),随后维持比较稳定的趋势.

4.2 基于种群的度量方法

基于基因的多样性度量方法根据基因的内部结构评价种群的多样性,没有考虑种群种个体的关系,本节提出基于种群的度量方法.

根据公式(4), m_i 表示种群中每一个体的小生境数, m_i 越大,表明和该个体相似的个体数就越多. 因此,可以根据种群中最大小生境数来度量种群的多样性. 同样,在 NIIS 中,对每一代种群,计算最大生境数,结果如图 2 所示. 表明开始演化后,最大小生境数明显降低(多样性明显提高),随后维持比较稳定的趋势.

根据图 1 和图 2 的试验结果,可以看出,种群经过演化之后,多样性明显提高,并保持较稳定状态.

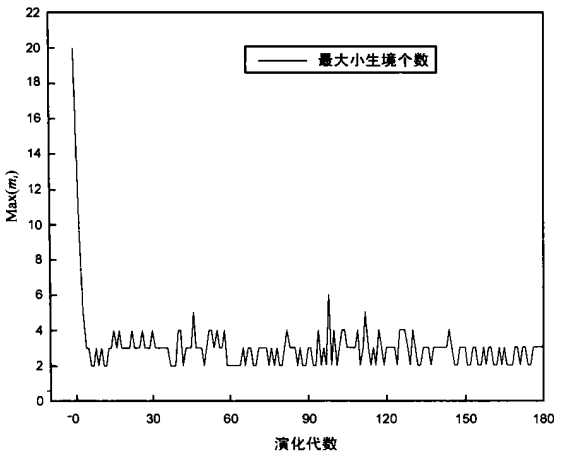

图 2　最大生境个数

5　结论

人工免疫系统为开发下一代入侵检测系统提供新的思路和方法,在入侵检测免疫模型中,关键技术是抗体 - 抗原的表示以及多样性抗体的生成,本文在深刻分析网络特征的基础上,提出一种新的抗体 - 抗原表示方法,并在此基础上,提出基于小生境技术的多样性抗体生成算法. 为了对算法进行评价,提出基于基因和基于种群的多样性度量方法,实验结果表明该算法能够有效地保持免疫模型演化过程中抗体集合的多样性. 本文的研究成果为实验室自主开发的网络入侵免疫系统(NIIS)奠定了基础.

参考文献:

[1] Kim J Bentley P. The artificial immune model for network intrusion detection [A]. Lotfi A. Zadeh. 7th European Conference on Intelligent Techniques and Soft Computing (EUFIT'99) [C]. Aachen, Germany: Verlag Mainz Press, 1999.

[2] Stephanie Forrest, Steven A. Hofmeyr. Immunology as information processing [A]. In Design Principles for the Immune Systems and Other Distributed Autonomous System [C]. Oxford: Oxford University Press, 2001. 361 - 388.

[3] Steven Hofmeyr, Stephanie Forrest. Architecture for an artificial immune system [J]. Evolutionary Computation, 1999, 7 (1): 1289 - 1296.

[4] Hofmeyr S A, Forrest S. Immunity by design: An artificial immune system [A]. Wolfgang Banzhaf. Genetic and Evolutionary Computation Conference (GECCO'99) [C]. Florida USA: Morgan Kaufmann Publishers, July 1999.

[5] M Oprea, S Forrest. How the immune system generates diversity: Pathogen space coverage with random and evolved antibody libraries [A]. Wolfgang Banzhaf. Genetic and Evolutionary Computation Conference (GECCO99) [C]. Florida USA: Morgan Kaufmann Publishers, July 1999.

[6] K Deb, D E Goldberg. An investigation of niche and species formation in genetic function optimization [A]. Proceedings Third ICGA [C]. San Mateo, CA: Morgan Kaufmann Publishers, 1989. 42 - 50.

作者简介:

凌　军　男,1976 年 3 月生于湖北,博士研究生,主要研究方向为智能网络管理、网络安全. Email: lingjun1976@163.com.

曹　阳　男,1943 年 9 月生于湖南,教授,博士生导师,主要研究领域为智能网络管理、网络安全、网络性能评价.