

文章编号:1001-9081(2007)09-2160-03

## 免疫入侵检测中基于数据场的动态识别算法

符海东,李 雪

(武汉科技大学 计算机科学与技术学院,武汉 430081)

(lx84112002@yahoo.com.cn)

**摘 要:**将数据场理论引入到计算机免疫的研究中,设计了一种识别器的构造方法及其动态识别算法。抗体的培育是建立在不完全自体集的基础上,算法可以识别出未知自体,降低自免疫反应发生的概率,并通过动态识别算法完善抗体集,克服了现有的入侵检测系统对自体集要求较高的局限性,简化了克隆变异以及记忆机制的实现方法。实验表明:新的免疫动态识别方法使入侵检测系统具有更高的动态平衡性和自适应性。

**关键词:**免疫;入侵检测;数据场;动态识别

**中图分类号:** TP393.08 **文献标志码:** A

### Dynamic recognition algorithm based on data field in immune intrusion detection

FU Hai-dong, LI Xue

(College of Computer Science and Technology,

Wuhan University of Science and Technology, Wuhan Hubei 430081, China)

**Abstract:** A construction method of detector and its relevant dynamic recognition algorithm were put forward by introducing the data field theory to computer immunology. Antibodies are brought up based on self set. By recognizing the unknown self set, the algorithm can decrease the rate of self-immunity, and also improve the antibody set dynamically and overcome the limitations of traditional IDs that have high requirement for self set, thus simplify the way to implement cloning, mutation and memory. The results of experiments show that the new dynamic recognition algorithm makes IDs possess a higher self adaptability and dynamic equilibrium capability.

**Key words:** immune; intrusion detection; data field; dynamic recognition

## 0 引言

生物免疫系统在抵御外界各种病毒和细菌等病原体的入侵方面担当着与入侵检测系统类似的任务。生物免疫系统所具备的免疫防御、免疫自稳和免疫监视等生理功能<sup>[1]</sup>,使生物体在生存期间能够战胜来自体外的各种病原体的侵袭以及体内细胞癌变等挑战。生物免疫系统能识别非己抗原,对其产生免疫应答并处理,对自身抗原则维持耐受。即能够正确识别自体(self)和非自体(nonself),这是生物免疫系统最重要的特点。同时,生物免疫系统在识别过程中所呈现出的动态平衡性、分布性、多样性、自组织和适应性等特性都是目前入侵检测系统所缺乏的。因此,生物免疫系统为入侵检测系统的设计开辟了新的途径。

## 1 相关工作

为了实现入侵检测的智能性、高效性,很多基于人工免疫的方法被提出。文献[2]提出了否定选择算法,用于生成检测器,完成了检测器的耐受过程。文献[3]提出了一种基于分子的人工免疫系统,用来模仿自然免疫系统,目的是保护计算机免受计算机病毒和其他因素的破坏。文献[4]建立了一套计算机免疫系统,用来抵御外来入侵,保障计算机系统的安全。文献[5]提出了动态克隆选择算法(DynamiCS),主要用

于网络入侵检测(NIDS)。

大部分基于免疫的入侵检测系统都建立在否定选择的识别模式上,这一识别模型存在固有的缺陷:

1)在否定选择算法中,抗体是随机生成并通过已知 self 集训练成熟的。在实际的环境中,对 self 和 nonself 认识存在局限性,依据否定选择算法培育的抗体在识别过程中很可能产生两种错误:错误地将未知的 self 当作 nonself 来识别;错误地将非法入侵当成合法行为。

2)入侵检测系统应该是一个动态平衡的系统。在一个平衡状态下,通过外部干扰(入侵、病毒等),系统应该能够自动达到一个新的平衡状态<sup>[6]</sup>。目前识别器的构造和检测所采用的方法有:基于距离的方法(如:Euclidean 距离、Manhattan 距离及 Hamming 距离等)、基于匹配度的方法(如:r\_连续位匹配规则)和基于结合强度的方法<sup>[7]</sup>。这些静态的构造和识别方法,对于外界的扰动不具备自组织和自适应性。

3)目前,识别器记忆机制是通过“优选”部分成熟识别器、人为延长它们的寿命、优化竞争策略和对其进行静态存储的方法来实现。这种记忆方法是静态的,对系统资源的依赖性较高。

针对以上不足,受数据场理论的启发,将数据场理论引入到计算机免疫入侵检测系统中,设计了一种识别器的构造方法及其动态识别算法。抗体的培育是建立在不完全自体集

收稿日期:2007-03-20;修回日期:2007-06-05。 基金项目:湖北省教育厅重点科研项目(2004D006)。

作者简介:符海东(1971-),男,湖北武汉人,副教授,博士,主要研究方向:人工免疫、人工智能、网络信息安全; 李雪(1984-),女,湖北襄樊人,硕士研究生,主要研究方向:计算机网络安全、人工免疫。

的基础上,动态识别算法可以识别出未知自体、动态完善抗体集。克服了现有的入侵检测系统对自体集要求较高的局限性,实现了识别器记忆的动态性。

## 2 数据场

### 2.1 数据场理论<sup>[8]</sup>

在基础物理学中“场”用于描述物质粒子间的非接触相互作用。比如:电场、磁场、重力场和速度场等。处于场中的各对象受到场力的作用会运动,最终达到平衡的状态。

受物理学中场论的启发,我们尝试将物质粒子间的相互作用及其场描述方法引入抽象的数据域空间。已知空间  $\Omega \subseteq R^p$  中包含  $n$  个数据对象的数据集,将每个数据对象视为  $p$  维空间中具有一定质量的粒子,其周围存在一个虚拟作用场,位于场内的任何其他对象都将受到场力的作用,由此所有对象的联合作用就在空间上确定了一个数据场。如图1所示。根据场论知识,如果空间中存在多个对象且没有外力作用,对象由于相互作用会相向运动,最终聚集成簇,达到平衡状态。当向处于平衡状态的数据场中添加新的对象时(即数据场受到外界的扰动),原有的数据场中的对象会产生相应的运动,最终形成新的分布。

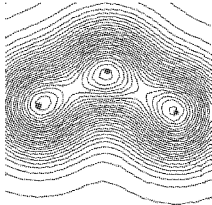


图1 多个数据对象形成的数据场

### 2.2 数据场相关定义

为了有效地描述数据场的空间分布规律,采用矢量强度函数和标量势函数进行表示,数据场中的标量势采用高斯函数定义。

**定义1** 已知空间  $\Omega$  中包含  $n$  个对象的数据集  $D = \{x_1, x_2, \dots, x_n\}$  及其产生的数据场,令数据对象的位置矢量分别为  $x_1, x_2, \dots, x_n$ ,则任一场点  $x$  处的势值和场强矢量可表示为:

$$\varphi(x) = \varphi_D(x) = \sum_{i=1}^n [m_i \times e^{-(\|x-x_i\|/\sigma)^2}] \quad (1)$$

$$F(x) = \sum_{i=1}^n [(x - x_i) \times m_i \times e^{-(\|x-x_i\|/\sigma)^2}] \quad (2)$$

其中:  $\|x - x_i\|$  为对象  $x_i$  到场点  $x$  的距离;  $m_i \geq 0 (i = 1, 2, \dots, n)$  为对象  $x_i$  的质量,满足归一化条件,即  $\sum_{i=1}^n m_i = 1$ ;  $\sigma \in (0, +\infty)$ ,用于控制对象间的相互作用力程,称为辐射因子。

**定义2** 通过最小化势函数与总体密度间的误差平方积分可优化估计对象的质量。其计算公式如下:

$$\min J = \min_{\{m_i\}} \left[ \frac{1}{2 \cdot (\sqrt{2})^d} \sum_{i=1}^n \sum_{j=1}^n m_i \cdot m_j \cdot e^{-\left(\frac{\|x_i-x_j\|}{\sqrt{2}\sigma}\right)^2} - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n m_i \cdot e^{-\left(\frac{\|x_i-x_j\|}{\sigma}\right)^2} \right] \quad (3)$$

**定义3** 已知代表对象  $x^* \in D$  及其质量  $m^*$ ,如果存在子集  $C \subseteq D$ ,使得  $\forall x \in C$  都存在一个点列  $x_0 = x, x_1, \dots, x_k \in \Omega$ ,使得  $x^*$  与  $x_k$  间的距离小于等于  $0.705\sigma \cdot m^*$  且  $x_i$  位于  $x_{i-1}$

的梯度方向 ( $0 < i < k$ ),则称  $C$  为以  $x^*$  为代表点的聚类。

## 3 基于数据场的动态识别算法

### 3.1 识别器的构造算法

系统的所有网络行为抽象为一个长度为  $p$  的属性字符串,该字符串表示IP地址、端口号和协议类型等网络事务特征。这些字符串组成集合  $D$ 。自体集合  $S$  为正常网络服务事务,其中  $S \subset D$ 。

基于数据场理论的识别器的构造算法,不同于Forrest提出的否定选择算法。在自体数据对象形成的数据场中,各个数据对象会在数据场中运动、聚集成簇,达到平衡状态。每个簇代表一类相似自体的集合。选取每个簇类的中心作为识别器,代表该类自体。

识别器集合  $D$  可表示为:  $D = \{d_1, d_2, \dots, d_i, \dots, d_n\}$ ,  $d_i = \langle m_i, r_i, y_i \rangle$ 。其中,  $d_i$  表示第  $i$  个抗体,  $m_i$  表示抗体  $d_i$  的质量,  $r_i$  表示抗体  $d_i$  的感知区(即数据  $d_i$  的场强作用范围),  $y_i$  表示抗体  $d_i$  的位置矢量。

基于数据场的识别器构造算法描述如下:

输入:自体集合  $S$ ,影响因子  $\sigma$ ,抗体集合的个数  $n$

输出:识别器集合  $D$

算法步骤:

- 1)  $MS = Estimate\_Mass(S, \sigma)$ ;  
//估计自体对象的质量,见定义2;
- 2)  $D' = Cluster\_Center(S, \sigma)$ ;  
//在自体数据形成数据场中,计算各聚簇中心的集合,公式见//定义1,计算方法参见定义3;
- 3)  $MD = Mass\_Range(D', S, \sigma)$ ;  
//估计候选识别器的质量及其场强的作用范围,见定义1、//定义2;
- 4)  $D = Max(MD, n)$ ;  
//从候选识别器集合中选取质量最大的  $n$  个个体;

### 3.2 动态识别算法

#### 3.2.1 算法思想

由数据场的聚类思想可知,如果未知数据对象  $b$  与识别器(聚类中心)具有相似的性质,则会落入该识别器的感知区,称该识别器识别了对象  $b$ ,对象  $b$  为自体数据。

如果所有的识别器都不能识别对象  $b$ (即对象  $b$  不在抗体集的感知区内),对象  $b$  就成为一个孤立点。对象  $b$  可能是未知的自体数据,也可能是异常数据。如果  $b$  是未知的自体数据,则将其作为新的抗体加入抗体集合中。这样,根据识别出的未知自体生成的新抗体可以在识别过程中动态加入,从而实现了抗体集合的不断完善。

判断孤立点是自体还是非自体的方法如下:

对于每个识别器,设置正常活动阈值  $Nr$  属性,在识别器集合形成的数据场中,当新的数据对象进入时,原有的数据场会发生变化,场中的数据对象会运动,形成新的分布。数据场中的数据对象的运动应该在一定的范围内,超过范围的运动,被认为是不正常的。由于造成场中数据对象异常运动的原因是由引入的新数据对象造成的,因此,该数据对象被认为是异常的。识别器的活动阈值  $Nr$  的设置公式如下:

$$Nr_i = \frac{r_i}{r_{\max}} \cdot m_i \cdot Dis\_min(d_i, d_j) \quad (4)$$

其中:  $Nr_i$  表示识别器  $i$  的正常活动范围,  $r_i$  表示识别器  $i$  的感知范围,  $r_{\max}$  为识别器集合中感知范围的最大值,  $d_j$  表示与识别器  $i$  距离最近的识别器,  $Dis\_min(d_i, d_j)$  表示识别器  $i$  与识别器  $j$  的距离,  $m_i$  表示识别器  $i$  的质量。

类似于抗体的克隆繁殖,在一定时间内,如果抗体识别了一个未知对象,其质量会增加。由式(1)可知识别器的识别范围与其质量成正比,当质量增加时其识别范围增加,对该类自体的识别能力增强,记忆时间相应的延长。如果一定时间内,抗体未识别任何未知对象,其质量相应的会衰减,识别范围会减小,记忆能力相应的会减弱。质量小于一定的阈值,该抗体被删除。质量越大,对数据场的空间分布影响越大,识别范围越大,对该类抗原的记忆能力越强,寿命越长。由此,通过改变识别器的质量,实现了抗体记忆的动态性和生命周期的动态改变。相比于通过“优选”部分成熟识别器、人为延长它们的寿命、优化竞争策略和对其进行静态存储的传统的静态记忆方法,该方法具有显著的优越性。

### 3.2.2 算法描述及说明

基于上述思想,动态识别算法如下:

1) 初始抗体集合的生成采用3.1节中“识别器的构造算法”进行实现。抗体表示为: $D = \{d_1, d_2, \dots, d_i, \dots, d_n\}$ ,  $d_i = \langle m_i, r_i, y_i, Nr_i \rangle$ 。其中, $d_i$ 表示第*i*个抗体, $m_i$ 表示抗体 $d_i$ 的质量, $r_i$ 表示抗体 $d_i$ 的感知区(即数据 $d_i$ 的场强的作用范围), $y_i$ 表示抗体 $d_i$ 的位置矢量, $Nr_i$ 表示抗体的移动阈值。

2) 抗原**b**进入后,对其进行识别。识别过程如下:寻找与抗原**b**距离最近的抗体 $d_i$ ,计算**b**与 $d_i$ 的距离 $L$ ,判断 $L$ 是否在 $d_i$ 的感知区内,如果在其作用范围内,则认为该对象是自体数据,转第5步。否则,将其作为一孤立点,转入第3步进行再次判断。

3) 如果**b**为孤立点,待其稳定后,计算抗体的移动距离 $Y$ ,判断 $Y$ 是否在其移动阈值内。如果在,则抗原**b**是自体,转入第4步。否则,转入第6步。

4) 如果**b**是自体,则将其作为新的抗体,计算其相应参数。加入抗体集合中,转第6步。

5) 在时间 $\tau$ 内,如果抗体 $d_i$ 识别抗原**b**,将其质量增加 $\Delta m$ ,重新计算抗体 $d_i$ 的相应参数。否则,减小其质量。

6) 抗原**b**识别结束。转第2步继续识别。

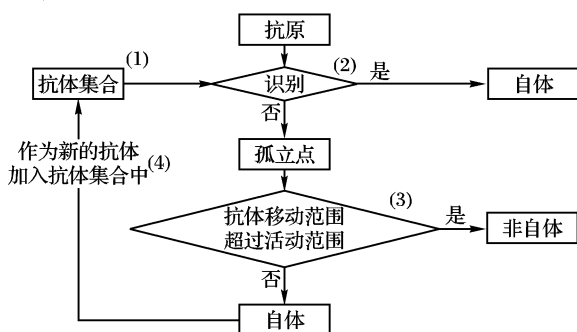


图2 基于数据场的动态识别算法

## 4 实验分析

实验数据采用 KDD Cup 1999 Data [KDD99] 中的数据。该数据集包含了网络中仿真的各种入侵。实验参数 $\sigma$ 设置可参照文献[9]所给的基于熵的优先方法,通过最小势熵来获取最优的 $\sigma$ 值。

实验过程中,随机从自体数据集中选取7000个数据点进行抗体的生成。初始自体数据对象在数据场中分布情况如图3。选取每个聚类中心作为抗体集,抗体个数设置为7个,设置每个抗体的感知区,如图4。抗体识别过程中抗体集的动态增加及感知区的变化如图5。

算法的执行时间与数据集的大小近似成线性关系。动态

识别算法在检测性能上,检测率达到了62.3%,误报率为1.2%。在不完全自体的集合的基础上,达到较高的检测率,说明新的免疫识别方法使入侵检测过程具有更高的动态平衡性和自适应性。

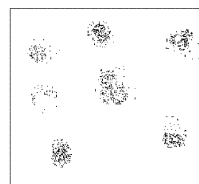


图3 自体数据初始分布

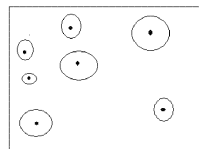


图4 抗体对象选取及感知区

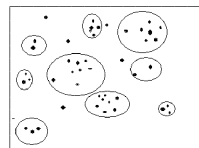


图5 抗体识别过程

## 5 结语

受数据场的启发,利用数据场的动态性和聚类性,将数据场的相关原理和理论应用到免疫入侵检测中,设计了一种识别器的构造和动态识别算法。该算法不同于传统的“自我—非我”识别方法,动态识别方法在不完全自我集的基础上生成不完全抗体集合,在识别过程中可以识别出未知的自我数据,动态完善识别器集合,实现了自我集合和抗体集的动态更新,克服了现有的入侵检测系统对自体集要求较高的局限性。此外,新的动态识别算法实现了抗体记忆的动态性和生命周期的动态改变。相比于通过“优选”部分成熟识别器、人为延长它们的寿命、优化竞争策略和对其进行静态存储的传统的静态记忆方法,该方法具有显著的优越性。

### 参考文献:

- [1] 李涛. 计算机免疫学[M]. 北京: 电子工业出版社, 2004.
- [2] FORREST S, PERELSON A S, ALLEN L, et al. Self/nonself discrimination in a computer[C]// Proceedings of IEEE Symposium on Research in Security and Privacy. [S. l.]: IEEE Press, 1994: 202-212.
- [3] DEATON R, GARZON M, ROSE J A, et al. DNA based artificial immune system for self/nonself discrimination[C]// Proceedings of the 1997 IEEE International Conference on Systems. [S. l.]: IEEE Press, 1997.
- [4] DASGUPTA D. Immune-based intrusion detection system: a general framework[C]// Proceedings of the 22nd national information systems security conference (NISSC). [S. l.]: IEEE Press, 1999.
- [5] BENTLEY K. Immune memory in the dynamic clonal selection algorithm[C]// 1st International Conference on Artificial Immune Systems (ICARIS-2002). Kent: University of Kent, 2002.
- [6] 梁意文. 网络信息安全的免疫模型[D]. 武汉: 武汉大学, 2002.
- [7] 莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨: 哈尔滨工业大学出版社, 2002.
- [8] 淦文燕, 李德毅, 王建民. 一种基于数据场的层次聚类方法[J]. 北京. 电子学报, 2006, 34(2): 258-262.
- [9] 淦文燕. 聚类—数据挖掘中的基础问题研究[D]. 南京: 解放军理工大学, 2003.