

## 超球体检测器覆盖问题的研究

李鑫鑫, 张凤斌, 王 涛

(哈尔滨理工大学 计算机学院, 黑龙江 哈尔滨 150080)

**摘 要:**研究表明实值否定选择算法在多维形状空间下呈现出很高的时间和空间复杂性。针对实值否定选择算法中最常采用的超球体检测器,在理论上研究了它的体积,以及体积随半径和维数变化的性质,以此分析了高复杂性出现的原因。针对检测器存在重叠的问题,基于蒙特卡罗方法提出了一个估计检测器覆盖率的算法,用于比较不同检测器生成算法。由于该算法基于随机分布和概率方法,它极大地简化了计算复杂性。

**关键词:**实值否定选择算法;超球体;检测器覆盖;蒙特卡罗方法

**中图分类号:**TP301

**文献标识码:**A

**文章编号:**1673-629X(2008)06-0131-03

## Research on Coverage of Hypersphere Detectors

LI Xin-xin, ZHANG Feng-bin, WANG Tao

(College of Computer Science and Technology, Harbin University  
of Science and Technology, Harbin 150080, China)

**Abstract:** The high time and space complexity of real-valued negative selection algorithm in high dimensions has been shown in recent research. Theoretically makes a study of the volume and several properties of hyperspheres detector in real-valued shape-space to analyse the reason of this high complexity. To solve the problem of detectors overlapping, based on Mento Carlo method, proposes an algorithm which estimates the total space (volume) covered by the hyperspheres to compare the coverage of different negative selection algorithms. Using the method of chance distribution and probability, the algorithm reduces the computational complexity.

**Key words:** real-valued negative selection algorithm; hypersphere; detector coverage; Mento Carlo method

## 0 引言

生物体的免疫系统负责抵御外部病原体的入侵,免疫系统具有增强学习、免疫记忆、分布式结构等特征。研究人员将免疫系统的这些特性应用于解决实际问题,形成了人工免疫系统这个新的研究领域<sup>[1]</sup>。

否定选择算法是人工免疫系统中研究最广的领域之一,它是由 Forrest 首先基于生物免疫原理提出的<sup>[2]</sup>。在算法过程中,检测器生成算法不同,所生成检测器集的异己检测效率也有很大的区别,所以针对不同编码和匹配方式出现了多种探测器生成算法。Forrest 的算法采用随机产生探测器,时间复杂度非常高, Kim 通过实例证明这些方法会带来严重的计算开销问题<sup>[3]</sup>,同时有些问题领域用二进制空间无法表达。于是引入了实值空间来表示<sup>[4]</sup>,紧接着各种实值否定选

择算法就被大量研究<sup>[5,6]</sup>。但是近来的研究表明,实值否定选择算法在多维空间下并没有显现出很好的识别功能和覆盖效果。文中研究的重点就是超球体(实值否定选择主要采用的检测器)在多维空间的性质。

## 1 实值否定选择算法

## 1.1 否定选择算法

Forrest 提出利用否定选择算法来处理各种异常检测问题。该算法主要应用在空间  $U$  上,包括自体集  $S$  和非自体集  $N$  满足:  $U = S \cup N$  and  $S \cap N = \emptyset$

否定选择算法可简述如下:

(1) 在空间  $U$  中定义自体集  $S$ ;

(2) 产生检测器集合  $D$ , 其中要产生的检测器与自体集  $S$  的每个自体都不能匹配;

(3) 用检测器集合  $D$  来实时检测自体集  $S$ , 如果发现匹配,说明自体集发生变化,有异常产生。

否定选择的基本观点是在自体空间外产生许多的检测器并应用这些生成的检测器来区分自体集和非自体集。

收稿日期:2007-09-14

基金项目:哈尔滨市学科基金(2003AFXXJ013)

作者简介:李鑫鑫(1983-),男,山东滨州人,硕士研究生,研究方向为网络安全;张凤斌,博士生导师,教授,研究方向为信息与网络安全。

## 1.2 实值否定选择算法(RNS)

实值否定选择算法通常操作在  $n$  维的 Euclidean 空间  $R^n$  上,通常研究其在超方体  $[0.0,1.0]^n$  上的性质。这种高层次的表示方法比二进制更能够准确地描述现实的各种知识,自体与检测器主要采取的匹配方法是它们之间的 Euclidean 距离。假设空间  $R^n$  有向量  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ ,  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ , 则该两个向量之间的 Euclidean 距离为:

$$\text{Euclidean}(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^n |x_i - y_i|^2 \right)^{\frac{1}{2}}$$

可以根据图 1 的算法来判断一个向量是否异常。

Detecte - Anomaly( $D, m, r$ )

$D$ : 检测器集合

$m$ : 要判断的向量

$r$ : 每个探测器的检测器的半径

Repeat:

For detector <sub>$i$</sub>  in  $D$

If Euclidean(detector <sub>$i$</sub> ,  $m$ )  $\leq r$

$M$  is anomaly; return

图 1 识别向量是否异常的算法

Gonzalez<sup>[4]</sup>提出了一种基于迭代的实值否定选择算法来处理连续数据,使其能够最大化覆盖率,但这种算法缺少一定的理论支持。后经过改进他又提出了一种随机 RNS 算法<sup>[5]</sup>,在随机产生的基础上采用模拟退火算法来分布检测器,这样使检测器的覆盖率在理论上有了支持。

Zhou Ji 提出了一种可变大小的 RNS 算法<sup>[6]</sup>,称为  $V$ -detector 算法。它能够产生很小的检测器来覆盖黑洞,这样就比固定半径的 RNS 算法有更大的覆盖率;另一方面,与以前采用检测器的数目作为控制参数不同,它采用估计的覆盖率来产生检测器,这样就能使检测器的产生更具自动化。

## 1.3 在多维空间下的 RNS 算法

虽然上面的各种算法在基于自己设计的数据来测试时都显示出较好的识别效果,但是它们更多是在二维空间下进行实验的。而实际在多维空间下,特别是在基于真实的数据集下,如 KDD 数据集, RNS 算法并不能显示出很好的性能。

其实不仅 RNS 算法,对多维数据进行分类的很多算法有同样的困难。随着维数的增加,搜索的时间和空间会呈指数级增加。对于 RNS 算法,当维数较低时,少量的检测器可以覆盖整个非自体空间;但当维数很高时,少量的检测器就不能覆盖整个非自体空间,需要产生指数级的检测器才能达到要求的覆盖率,这在现实中是无法达到的。假如产生 1000 个样本点,能确信在一维和二维空间下它能很准确地描述所需要的区

域。最少需要两个点来表示一维空间这个范围,在二维空间下至少需要 4 个点,而在 10 维空间下,由于需要的覆盖空间不是那么整齐,只用 1000 个样本点来描述就太少了。所以说现在研究多维空间下检测器性质显得尤为重要。

## 2 超球体的性质

在 RNS 算法中,最常用的一种检测器是超球体,也有用超方体作为检测器的,由于超球体检测器比较普遍,在这里主要研究超球体的性质。

### 2.1 超球体的体积

半径为  $r$  的  $n$  维超球体的体积为:

$$V(n, r) = r^n \cdot \frac{\pi^{n/2}}{\Gamma(n/2 + 1)}$$

其中,对  $n \in N$ , 有  $\Gamma(n+1) = n!$ ; 对半整数, 有  $\Gamma(n + \frac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdot 7 \cdots (2n-1)}{2^n} \sqrt{\pi}$ 。

对于  $n$  维单元球的体积  $V(n)$  可以通过递归得出:

$$V(1) \rightarrow V(2) \rightarrow \cdots \rightarrow V(n)$$

对于一个二维的单元球  $C^2 = \{(x_1, x_2) \in R^2 \mid x_1^2 + x_2^2 \leq 1\}$ , 它的体积可计算如下:

$$\begin{aligned} V(C^2) &= 2 \cdot \int_{-1}^1 \sqrt{1-x_2^2} dx_2 \\ &= 2 \cdot \int_0^\pi \sqrt{1-\cos^2(t)} \sin(t) dt \\ &= 2 \cdot \int_0^\pi \sin^2(t) dt = \pi \end{aligned}$$

由二维的体积可以推得三维的体积:  $V(C^2) \rightarrow V(C^3)$

$$\begin{aligned} V(C^3) &= \int_{-1}^1 \pi(\sqrt{1-x_3^2})^2 dx_3 \\ &= \pi \int_{-1}^1 (1-x_3^2) dx_3 = \frac{4}{3} \pi \\ &\dots \end{aligned}$$

递归得出  $n$  维超球体的体积:  $V(C^{n-1}) \rightarrow V(C^n)$

$$\begin{aligned} V(C^n) &= V(C^{n-1}) \cdot \int_{-1}^1 (1-x_n^2)^{(n-1)/2} dx_n \\ &= \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} \end{aligned}$$

### 2.2 超球体的性质

在多维实值空间下,超球体就会显示出一些特别的性质。可以通过改变半径和维数,观察超球体的体积变化,研究超球体的性质对否定选择算法的影响。

推论 1: 当  $n$  趋于无穷时,超球体的体积趋于 0:  
 $\lim_{n \rightarrow \infty} V(n, r) = 0$

证明: 由  $\Gamma(\frac{n}{2} + 1) \approx \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}$ , 得:

$$\lim_{n \rightarrow \infty} \left( r^n \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} \right) = \frac{1}{\sqrt{2\pi}} \lim_{n \rightarrow \infty} \left( \frac{re\sqrt{\pi}}{n^{n+1/2}} \right)^n = \frac{1}{\sqrt{2\pi}}$$

$$\lim_{n \rightarrow \infty} \left( \frac{c_n}{n^{n+1/2}} \right) = 0$$

从推论 1 可以看出对于固定半径的超球体,随着维数的增加它的体积就会变小。当维数很高时,它的体积趋于 0。这样,就需要在超方体  $[0.0, 1.0]^n$  产生大量的检测器才能够比较完整地覆盖非自体空间,于是产生检测器的否定选择算法就有很高的时间和空间复杂度。

推论 2:超球体在半径  $r - \epsilon$  到  $r$  之间部分的体积占整个超球体体积的比例为

$$\alpha(n, r, \epsilon) = 1 - (1 - \frac{\epsilon}{r})^n, \text{ 其中 } 0 < \epsilon < r.$$

$$\begin{aligned} \text{证明: } 1 - \frac{V(n, r - \epsilon)}{V(n, r)} &= 1 - \frac{(r - \epsilon)^n \cdot \frac{\pi^{n/2}}{\Gamma(n/2 + 1)}}{r^n \cdot \frac{\pi^{n/2}}{\Gamma(n/2 + 1)}} \\ &= 1 - (1 - \frac{\epsilon}{r})^n \end{aligned}$$

由推论 2 可以看出,如果向量随机分布,那么大部分的向量就会分布超球体的表面附近。

假如在超方体  $[0.0, 1.0]^n$  中取  $r = 1, \epsilon = 0.1, n = 50$  的超球体,  $\alpha(n, r, \epsilon) = \alpha(50, 1, 0.1) = 1 - (1 - \frac{0.1}{1})^{50} \approx 99.48\%$ , 也就是说半径 0.9 到 1 之间的体积占整个超球体体积的比例约 99.48%, 如果向量随机分布,约 99.48% 的向量分布在超球体半径为 0.9 到 1 的范围之间,这样就给以 Euclidean 距离为衡量标准的检测方法增加了很大的困难。

### 3 检测器覆盖范围的估计算法

在第 2 节介绍了单个超球体的覆盖问题,但是当用否定选择算法来生成多个检测器时,一般都有检测器存在重叠问题,即整个空间内检测器覆盖的体积并不是所有检测器的体积之和。

下面就介绍一种普遍的算法来估计检测器的覆盖率。

#### 3.1 蒙特卡罗方法

由概率定义知,某事件的概率可以用大量试验中该事件发生的频率来估算,当样本容量足够大时,可以认为该事件的发生频率即为其概率。蒙特卡罗方法的基本思想是,首先建立一个概率模型,使它的参数等于问题的解,然后通过对模型或过程在计算机上生成随机数来计算所求参数,最后给出所求解的近似值,而解的精度可由参数估值的标准差来表示。该方法回避了结构可靠度分析中的数学困难(如维数  $n$ ),只要模拟

的次数足够多,就可得到一个比较精确的失效概率和可靠度指标<sup>[7]</sup>。

给定可信度  $1 - \delta$ , 可以利用切比雪夫不等式得出积分误差不大于  $\epsilon$  时所需的样本点数目  $N = \lceil 1/4\delta\epsilon^2 \rceil$ 。

#### 3.2 用蒙特卡罗方法估计超球体覆盖范围

Detector - Coverage(RNS,  $\epsilon, \delta$ )

RNS: 采用的否定选择算法

$\epsilon$ : 估计的体积的绝对误差

$\delta$ : 可靠度

$[H, r]$ : 超球体检测器集

$[H, r] \leftarrow \text{RNS}$  // 由否定选择算法产生检测器集

inside  $\leftarrow 0$

$N \leftarrow \lceil 1/4\delta\epsilon^2 \rceil$

P1.  $X \leftarrow$  从  $[0, 1]^n$  随机选择的向量

If Detecte - Anomaly( $H, X, r$ ) //  $X$  是否能被  $[H, r]$  检测到

Inside  $\leftarrow$  inside + 1

Goto P1

Return (inside/ $N$ )

end

图 2 估计超球体覆盖范围的算法

由图 2 可以看到算法的过程,随机产生  $N$  个样本点,判断它们是否能被实值否定算法(RNS)产生的检测器集检测,得到能被检测到的样本点数 inside,则该实值否定算法的覆盖率为 inside/ $N$ 。

### 4 结束语

讨论否定选择算法在实值形状空间下的各种检测器生成算法、包括基于迭代的 RNS 算法、随机 RNS 算法、V - detector 算法等。发现在多维空间下无论是哪种算法都呈现很高的时间和空间复杂度,于是分析了最常采用的超球体检测器在多维空间下的体积和它的各种性质,以研究高复杂度出现的原因。对于另一种检测器——超方体和超球体检测器的不同和相同的性质,以及性质对覆盖率的影响是下一步研究的工作。

最后利用蒙特卡罗方法设计了一个估算多个检测器覆盖空间的算法,用于比较已经研究得到的各种无论是二维还是多维空间下的否定选择算法。

#### 参考文献:

- [1] Dasgupta D. Artificial Immune Systems and Their Applications[M]. New York: Springer - Verlag, 1999.
- [2] Forrest S, Perelson S, Cherukuri R. Self - nonself discrimination in a computer[C]//Proceedings of IEEE Symposium on Research in Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1994: 202 - 212.

(下转第 137 页)

```

}
}
...
}
sleep(1);
}
pthread_exit((void *)0);
}

```

图4 do\_rules 主要过程

### 3.4 防火墙规则集的优化

Netfilter代码主要是循环执行对应协议及调用点的钩子函数,其中最主要的是调用 ipt\_do\_table 函数。在该函数调用中循环执行规则中的每个 match,如果成功则执行该规则的 target,否则继续下一条规则的匹配直至最后一条缺省的安全策略,最后向协议栈返回处理结果,匹配过程如图5所示。因此可知 Netfilter 一个重要的性质:规则集的顺序敏感性。它采用标准的、自顶向下地使用规则遍历<sup>[4]</sup>,直到发现一条匹配规则。因此建立规则的数量、规则链的长度和数据包在匹配之前的比较次数将显著影响防火墙的效率。并且从本质上来说,防火墙的性能并非由规则的数量决定而是由比较的次数决定的。因此从以下3个方面进行优化:

```

do {
    .....
    if (match)
        .....//如果匹配,则执行规定动作;
    else {
        no_match: //不匹配
        e = (void *)e + e->next_offset;
    }
    .....
} while (!hotdrop);

```

图5 ipt\_do\_ipable 主要过程

(1) 在包过滤模块中使用状态机制可以让正在进

行中的、已经得到认可的、已经被接受的交换绕过防火墙规则。交换被初始化和接受之后,数据包被当作是已经建立交换的一部分,剩余的防火墙过滤就可以被绕过去。因此,动态与静态过滤规则需结合使用。

(2) 将规则组织成层次分明的树型结构,根据数据包的特征,通过有效的组织优化,数据包的匹配能够有选择地进行,而不必按顺序自顶向下遍历。一般可根据传输层协议,分为 TCP 规则、UDP 规则、ICMP 规则及其他规则。这样允许从临界比较点划分比较关系,能使最终匹配的数据包进行比较的次数减少为原来的一半<sup>[5]</sup>。

(3) 传输层协议的顺序因素:UDP 规则放到 TCP 规则之后,将 ICMP 规则放到规则链的后端。

## 4 结 语

本系统不仅可以实时地检测攻击行为,还能根据攻击行为危害程度采取不同的动作,包括日志和报警阻断。其中日志记录入侵活动,报警阻断则修改防火墙的规则策略。通过对防火墙规则集的优化提高了系统对防御的灵敏性。

### 参考文献:

- [1] Lindstrom P. Understanding Intrusion Prevention[EB/OL]. 2003. [http://www.networkassociates.co-m/us/\\_local/promos/\\_media/wpspire.pdf](http://www.networkassociates.co-m/us/_local/promos/_media/wpspire.pdf).
- [2] Sturges S. Snort Users Manual[EB/OL]. 2007. [http://www.snort.org/docs/snort\\_hmanuals/htmanual\\_280/](http://www.snort.org/docs/snort_hmanuals/htmanual_280/).
- [3] 王丽辉,李 涛,张晓平,等.一种联动防火墙的网络入侵检测系统[J]. 计算机应用研究,2006,23(3):95-97.
- [4] Suehring S, Ziegler R L. LINUX 防火墙[M]. 北京:机械工业出版社,2006.
- [5] 朱立才,杨寿保,宋舜宏. Netfilter/iptables 防火墙性能优化方案与实现[J]. 计算机工程与应用,2006,42(15):117-120.

(上接第133页)

- [3] Kim J, Bentley P. Negative selection and nicking by an artificial immune system for network intrusion detection[C]//Proc of GECCO'99. Orlando, Florida, USA: Morgan Kaufmann, 1999:149-158.
- [4] Gonzalez F, Dasgupta D, Gomez J. The effect of binary matching rules in negative selection[C]//In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO). Chicago: Springer-Verlag, 2003:198-209.
- [5] Gonzalez F, Dasgupta D, Nino L F. A Randomized Real -

- Valued Negative Selection Algorithm[C]//In Proceedings of the 2nd International Conference on Artificial Immune Systems. Edinburgh, UK: Napier University, 2003:261-272.
- [6] Zou Ji, Dasgupta D. Real - valued negative selection algorithm with variable - sized detectors[C]//Proceedings of GECCO. Seattle, Washington: Springer, 2004:287-298.
- [7] 徐钟济.蒙特卡罗方法[M]. 上海:上海科学技术出版社, 1985.