高技术通讯 2012 年 第22卷 第5期:449 ~ 454

doi:10.3772/j.issn.1002-0470.2012.05.001

V-detector 优化算法①

张凤斌② 席 亮③ 王胜文 岳 新

(哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)

摘 要 针对人工免疫系统中 V-detector 否定选择算法造成的检测器集合黑洞和检测器高重叠率等问题,借鉴生物免疫系统对免疫细胞的调节机制,提出了 V-detector 优化算法。该算法从父代产生候选检测器子代并通过检测器之间以及检测器与自体集合之间的亲和力对比更新检测器集合,使得检测器集合对非自体空间的覆盖更加合理。通过二维仿真实验和 KDD CUP 99 数据集实验测试,经优化后的检测器集合对非自体空间的覆盖性能有了显著提高,有效提高了系统的检测性能。

关键词 人工免疫系统,实值,可变阈值否定选择算法,优化算法

0 引 言

人工免疫理论是借鉴生物免疫系统(biological immunity system, BIS)的机理来解决现实问题的理论[1],是继人工神经网络、遗传算法之后,从生物系统中提取的又一智能化理论[2]。虽然起步较晚,但是经过国内外专家的不断努力,它已经具备了完整的理论体系、成熟的理论模型和简单的计算方法[3,4],而且其中许多理论和方法已经在相关领域得到了很好的应用[5,6]。目前的人工免疫算法主要有否定选择算法(negative selection algorithm, NSA)、免疫网络和克隆选择算法[7]。其中,否定选择算法作为最早的人工免疫方法之一,模仿生物免疫系统中自体/非自体区别方法,成为一种异常检测方法[8],通常在形态空间(shape-space)中讨论问题。作为否定选择算法簇的一员,实值否定选择(real-valued negative selection, RNS)算法的搜索空间是连续的,并且通常以超球体或超立方的形式表示自体/检测器[9]。

检测性能的好坏主要取决于检测器的质量,即检测器对非自体空间的覆盖率(识别区域)[10]。而基本 RNS 并不能保障检测器对非自体空间的覆盖率,容易造成黑洞问题。Zhou 对此提出了实值 V-detector 否定选择算法(RNS with variable-coverage detectors,简称 V-detector 算法)[11]。V-detector 算法以迭代的方式在形态空间中生成不同大小的超球体检测器,大的检测器覆盖较大区域空间,小的检测器覆盖自体/非自体边界的漏洞,同时检测器规模也得到了较好的控制。因其在一定程度上减少了黑洞问题,提高了 RNS 的精确性和时效性,从而成为该领域中的一个里程碑式的算法。但是它所生成的检测器相互覆盖的高重叠问题是其一大弊病。后来 Zhou 和李涛等分别对其进行了改进[12-14],但这一问题并没有得到很好的解决。洪征等针对高重叠问题对其进行了改进[15],虽然有效地减少了重叠问题,但却增加了检测器对自体空间覆盖的可能,一定程度上增加了系统的误报率。公茂果等利用多目标免疫算法对 V-detector 进行分布优化[16],在不增加检测器数量的前提下更好地覆盖了非自体空间,但算法复杂度较高。本文就以上问题进行分析,借鉴生物免疫系统对免疫细胞的调节机制,提出一种 V-detector 优化算法。该算法通过生成更优秀的子代检测器以及去除冗余检测器来构造较为合理的检测器集,从而达到提高系统整体性能的目的。

1 相关分析

1.1 实值形态空间与否定选择算法

实值空间把与 R^n 子集对应的"自体/非自体"

① 国家自然科学基金(60671049, 61172168)资助项目。

② 男,1965 年生,博士,教授,博士生导师;研究方向:网络与信息安全;E-mail: zhangfb@hrbust.edu.cn。

③ 通讯作者,E-mail: xljyp2002@yahoo.com.cn

(收稿日期:2011-01-25)

形态空间归一化到 $[0,1]^n$ 矩形空间或直径为 1 的超球体空间 U。U 分为自体子空间 S 和非自体子空间 R:$U = S \cup R$。检测器集 $D \subseteq R$, 自体 s 和检测器 d 分别分布于 S 和 D, 一般通过否定选择算法进行区分。

否定选择算法是对免疫细胞的成熟过程的模拟。算法应用在两个阶段:训练检测器阶段和检测器检测阶段,如图 1 所示。训练阶段主要负责成熟检测器的生成,大体过程如下:将候选检测器进行耐受训练,删除与自体对抗的个体,并保留能检测非自体的个体作为成熟检测器,经历了耐受的检测器模拟成熟的免疫细胞;在检测阶段,事件依次与每个检测器进行亲和力匹配,亲和力高的事件被认为是异常。目前,RNS 算法和 V-detector 算法及其改进算法基本采用欧氏距离来衡量样本间的亲和力,距离越大则亲和力越小。

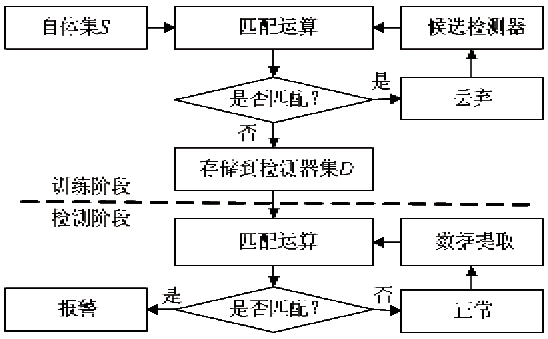


图 1 否定选择算法(训练阶段和检测阶段)

1.2 实值 V-detector 问题分析

1.2.1 检测器数量问题

无论是 RNS 算法还是 V-detector 算法及其改进算法,为了较好地覆盖非自体空间,系统需要生成大量的检测器。如果数量不够,则必然会出现某些非自体空间没被覆盖而造成检测器黑洞现象,如图 2 中的 d_1、d_2 间的黑洞。在检测阶段,此区域的异常将不会被检测出来,从而降低了系统的检测率;如果检测器数量过大,在检测阶段,事件与每个检测器进行亲和力计算也会增加系统负担,降低系统的时效性。

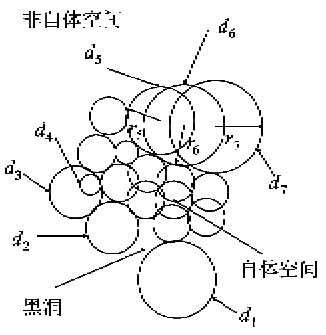


图 2 V-detector 存在的问题

1.2.2 边界重叠覆盖问题

原始的 RNS 还存在 S/R 边界黑洞问题,即 S/R 边界较难被检测器覆盖的问题。V-detector 算法以检测器 d 中心与其亲和力最大的自体 s 边界间的距离为检测半径,从而避免了这一问题。但却造成每个检测器都会覆盖 S/R 边界区域,重复覆盖现象严重。如图 2 中的 d_5、d_6、d_7, 其重叠覆盖非自体空间的比例很大; d_3 彻底覆盖了 d_4 所覆盖的非自体区域,显然 d_4 是冗余的,它的唯一作用就是增加系统在检测阶段的负担。

1.2.3 检测器数量 N_D 对系统检测性能的影响

为了便于计算,假设检测器相互独立,设每个检测器匹配异常的概率为 P_i, 则任一个异常没有被检测器集合匹配的概率 P_f 为

$$P_f = \prod_{i=1}^{N_D} (1 - P_i) \tag{1}$$

当 N_D 很大时, P_i 维持在一个较小的水平,为了简便表示,在此设定 P_i 为定值 P_m, 则式(1)可近似表示为

$$P_f \approx (1 - P_m)^{N_D} \approx e^{-N_D P_m} \tag{2}$$

从式(2)中可以看出,当 P_m 一定时,检测器集合规模与一次误报率成指数关系。而且在实际应用中,检测器相互独立不可能全部成立,在 V-detector 中,高度重叠的检测器相互独立性更小,使得检测失败概率增大。

1.3 生物免疫调节机制

生物免疫调节机制是指在遗传基因控制下具有增强和抑制作用的免疫细胞和分子的相互制约与调节,控制各种免疫细胞的浓度和活性,来调控免疫应答的强度和正、负方向。各种免疫细胞在该机制的作用下呈现出一个显著特征:当其浓度达到一定水平时,其活性反而被抑制。这就很好地说明了免疫调节机制在控制着免疫应答的发生、发展和消退,保持生物免疫系统的高效性和自适应性[17]。

2 V-detector 优化算法

本文针对 V-detector 的特点及存在的问题,借鉴生物免疫系统对免疫细胞的调节机制,提出一种 V-detector 优化算法。该算法通过检测器间的亲和

力判定检测器重叠率的高低更新检测器集。定义自体 $s_i = (s_{i1}, s_{i2}, \cdots, s_{in}, r_i)$,检测器 $d_i = (d_{i1}, d_{i2}, \cdots, d_{in}, r_i)$。为了与 V-detector 算法进行对比,算法依然采用传统的欧氏距离方法衡量亲和力。算法优化的目的是用较合理的检测器较好地覆盖非自体子空间而不侵犯自体空间,最大化

$$V(D) = \text{Volume}\{x \in R \mid \exists d \in D, A(x,d) \leqslant \varpi\} \quad (3)$$

限制条件为

$$\{\forall d \in D \mid \exists s \in S, A(s, d) \leqslant \gamma\} = \varphi \quad (4)$$

且

$$\{\forall d_i, d_j \in D, A(d_i, d_j) \leqslant r_i \mid A(d_i, d_j) \leqslant r_j\} = \varphi \quad (5)$$

其中 $V(D)$ 为检测器集合所覆盖的区域,$A(d_i, d_j)$ 为两检测器的欧氏距离。φ 和 γ 为匹配阈值。式(4)要求检测器不侵犯自体空间,式(5)要求检测器间互不覆盖。

算法基本思想:计算检测器集各个体间的亲和力,亲和力越高说明个体间越相似,该个体及其相似个体的浓度越高。算法去除被其他个体完全覆盖的个体;将重叠部分在一定范围内的个体用子代替代;将黑洞用周边个体的子代覆盖。

算法设定:

(1)阈值 σ。使得两检测器距离在 $[\sigma, L]$ 时保留该两个个体而不做优化,其中 $L = r_i + r_j$。这是考虑到个体间重叠率在小于某一阈值的情况下对于系统整体的检测效率影响并不大而做出的减少计算代价的措施。σ 越小则保留的父代越多。

(2)阈值 δ。如果子代与其亲和力最大自体的边界间的距离小于 δ 则丢弃该子代而保留父代。这是为了防止子代的覆盖区域较父代的覆盖区域的并减少过多而造成的二次黑洞问题。δ 越小则删除的父代越多。

具体优化算法如下:

输入:S,D,优化后检测器集合 D'(初始为空):输出:D'。

步骤 1:对每个 $d_i \in D$,如果 $d_i \neq Null$,则找到与其亲和力最大的 $d_j (d_j \in D \cup D')$,转向步骤 2;否则转向步骤 8。

步骤 2:如果 $A(d_i, d_j) <= | r_i - r_j |$,转向步骤 3;否则转向步骤 4。

步骤 3:删除半径较小者并将另一个作为子代放入 D',转向步骤 1。

步骤 4:如果 $A(d_i, d_j) < \sigma$,转向步骤 5;否则转向步骤 6。

步骤 5:取二者的中点生成子代 d_{son} 并查询与其亲和力最大的 s,令 $r_{son} = A(d_{son}, s) - r_s$。如果 $r_{son} < \delta$ 则丢弃 d_{son} 并且 $d_i, d_j \rightarrow D'$,否则 $d_{son} \rightarrow D'$,删除 d_i, d_j。转向步骤 1。

步骤 6:如果 $A(d_i, d_j) > L$,转向步骤 7;否则转向步骤 1。

步骤 7:取二者的中点生成子代 d_{son},令 $r_{son} = A(d_i, d_j) - L$,判断 d_{son} 是否覆盖自体区域。如果是则丢弃,$d_i, d_j \rightarrow D'$;否则 $d_{son} \rightarrow D'$,并删除 d_i, d_j。转向步骤 1。

步骤 8:得到优化后检测器集 D',算法结束。

3 分析与实验

通过上面的算法可以看出,在寻找与 d_i 亲和力最大的 d_j 的时间代价和所有父代处理一遍的时间代价都是 $O(N_D)$,在查找与子代亲和力最大的自体的时间代价和判断子代是否覆盖自体空间的时间代价都是 $O(N_S)$,其中 N_S 为自体个数。所以优化算法总的时间代价约为 $O(\text{Max}(N_S, N_D) \cdot N_D)$。而在优化过程中增加的空间代价为 $O(1)$。在检测器集合优化阶段,这种 $O(n^2)$ 的时间代价和 $O(1)$ 的空间代价是可以被接受的。而且优化后的检测器集合去除了一些不必要的个体,新增加了具有更好效果的子代个体,所以优化算法是可行的。

为了进一步说明算法的可行性和有效性,本文通过仿真实验来验证检测器对非自体空间的覆盖效果,采用 KDD CUP 99 数据集[18]来检测优化后的检测器检测性能。实验均采用式

$$s'_{ij} = (s_{ij} - \min_j)/(\max_j - \min_j) \quad (6)$$

将数据各维属性归一化到 $[0,1]^n$,并设定自体半径均为 0.05。式中 $\max_j$ 和 $\min_j$ 分别表示该属性的最大值和最小值。

3.1 仿真实验

为了有直观的效果展示,本实验以二维空间为例进行仿真实验。以空间中心为中心的五角星区域为自体区域,其中的样本点作为自体样本(51 个)。选择五角星区域是因为它可以较好地体现自体区域的复杂性并能体现出 V-detector 在处理自体/非自体边界区域的优势以及优化算法对于自体/非自体边界覆盖的保持。自体区域如图 3(a)所示。

为了直观说明检测器对与非自体的覆盖率,本文采取文献[19]的方法计算自体集大小 $Q(S)$ 和检

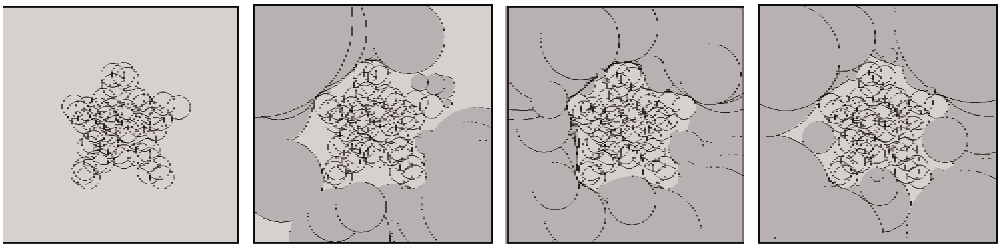
测器覆盖区域大小 $Q(\sum d)$ 以及检测器集总的重叠大小 overlapping (D)，从而得到 V-detector 优化前后检测器集合对非自体空间的覆盖率。

$$p(D) = \frac{Q(\sum d) - \text{overlapping}(D)}{1 - Q(S)} \times 100\% \tag{7}$$

其中

$$\text{overlapping}(D) = \sum_{i \neq j} \text{overlapping}(d_i, d_j)$$
$$i, j = 1, 2, \cdots, N_D \tag{8}$$

按 V-detector 算法分别生成检测器 40 个和 120 个，对非自体空间的覆盖率分别为57.2%和98.4%，而重叠率分别为 37.26% 和 71.65%。结果如图 3(b)、图 3(c)所示。可以看出，检测器在数量很少的情况下很容易出现黑洞现象，在检测器数量较多的情况下，虽然可以很好地覆盖非自体区域，但是检测器重叠现象很严重。按本文算法优化含 120 个检测器的检测器集(为了简单直观地说明，在此设定 $\sigma = \delta = L/2$，对于 σ 和 δ 对优化算法的影响将在下面的实验中讨论)，优化后的检测器数为 33 个，对非自体空间的覆盖率变为91.7%，而重叠率为 28.61%。结果如图 3(d)所示。可以看出，优化后的检测器不仅在数量上大大降低，而且对非自体区域覆盖的效果依然保持很好。



(a) 自体集　(b) V-detectors (40个)　(c) V-detectors (120个)　(d) 优化后检测器集

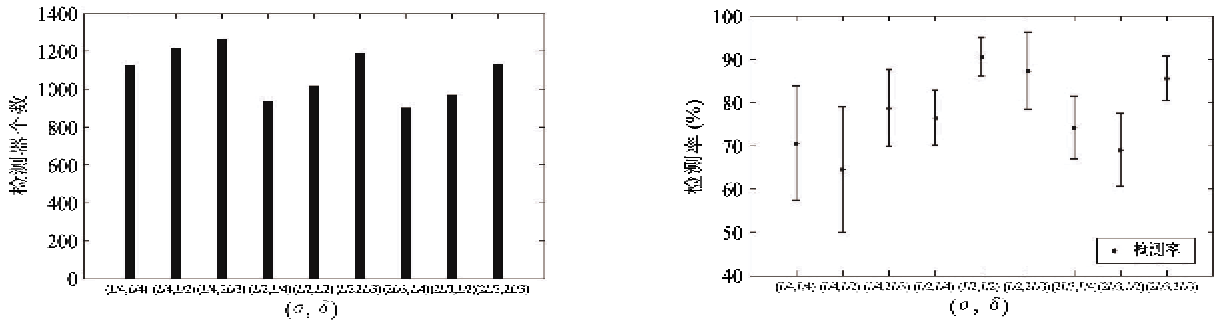
图 3　优化前后检测器对比

3.2　KDD CUP 99 数据集实验

为了探讨 σ 和 δ 对优化算法的影响，以及对比经优化算法优化后 V-detectors 与原始 V-detectors 的检测效果，本文选用本领域的权威数据集 KDD CPP 99 数据集进行实验。该数据集每条记录有一个标志位(正常/异常)。本实验选用其中一个 10% 的子集，以其中正常记录为自体训练检测器并使用这些检测器检测数据集中的异常记录。计算 TP(正确肯定次数)、TN(正确否定次数)、FP(错误肯定次数)、FN(错误否定次数)，并使用这 4 个值计算检测率(异常事件被检测为异常的概率 $P = TP/(TP + FN) \times 100\%$)和误报率(正确的事件被误检测为异常的概率 $P_f = FP/(TN + FP) \times 100\%$)来判定检测器对非自体空间的覆盖效果。

(1) σ 和 δ 对优化算法性能的影响

首先由 V-detector 算法生成检测器 3000 个，然后分别设定 σ 和 δ 为 $L/4$，$L/2$ 和 $2L/3$ 并进行不同组合 (σ, δ) 对其进行优化并分别检测数据集中的异常，结果如图 4 所示。从图 4(a)可以看出，无论是那种组合，优化后的检测器数量都大大降低。从图 4(b)显示的各组合检测率及其标准差(图中上下横线纵坐标)可以看出，(σ, δ) 为 $(L/2, L/2)$ 时检测率最高且最稳定。

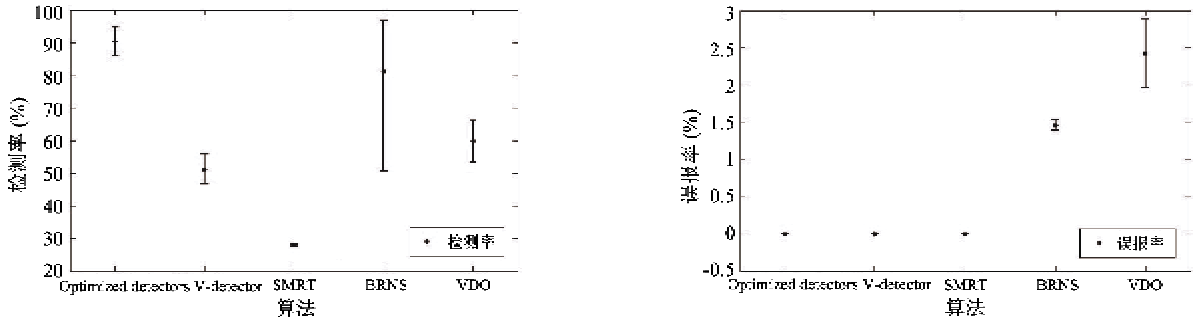


(a) 检测器个数对比　(b) 检测率对比

图 4　不同 (σ, δ) 组合对应检测器集合检测结果对比

(2)与 V-detector 及其改进算法的性能比较

选用几种 V-detector 的改进算法进行对比试验:文献[8]的 Statistically more reliable termination 算法(记为 SMRT,改进了 V-detector 的收敛方法),文献[9]的 Boundary-aware 算法(记为 BRNS,将部分自体区域转变成检测器自身的适应度)和文献[11]的优化算法(记为 VDO,通过移动检测器位置来减少检测器间的重叠率)。采用同样参数各生成检测器 3000 个,与上面的原始 V-detector 生成的检测器集和 $\sigma = \delta = L/2$ 的优化后检测器集进行检测比较。检测结果如图 5 所示,SMRT 检测率最差,VDO 误报率最高,BRNS 虽然比 VDO 效果好,但其检测率的标准差远大于优化后的 V-detectors,而且误报率较 V-detector 和优化后的 V-detectors 都偏高,只比 VDO 好一些。总之,通过本文算法改进的检测器集效果最好。



(a) 检测率对比　　(b) 误报率对比

图5　五种算法对应检测器集检测结果对比

4　结　论

本文讨论了实值 V-detector 的弊端,从提升检测器对非自体空间的覆盖性能角度出发,借鉴生物免疫系统对免疫细胞的调节机制,提出了 V-detector 优化算法。实验测试结果表明,与原始 V-detector 及相关改进算法相比,V-detector 优化算法在降低了检测器间相互覆盖率的基础上,保证了检测器对非自体空间的覆盖率,使检测器集合在规模上更加合理,从而有效提高了系统的检测性能。同时,V-detector 优化算法使 V-detector 算法更加实用化,这也为人工免疫系统及其相关领域提供一个很好的方法。另外,本文提出的优化算法还有许多地方需要进一步改进和分析,比如加入动态更新机制和更广泛的实验论证等。

参考文献

[1] Dasgupta D, Zhou J, Gonzalez F. Artificial immune system (AIS) research in the last five years. In: Proceedings of the IEEE Congress on Evolutionary Computation 2003, Canberra, Australia, 2003. 123-130

[2] Dal D, Abraham S, Abraham A, et al. Evolutionary induced secondary immunity: an artificial immune systems based intrusion detection systems. In: Proceedings of the 7th Computer Information Systems and Industrial Management Applications 2008, Ostrava Czech Republic, 2008. 65-70

[3] Dasgupta D, Yu S, Nino F. Recent Advances in artificial immune systems: models and applications. *Applied Soft Computing*, 2011, 11(2): 1574-1587

[4] Barreira A F, Teixeira O N, De J G, et al. Evolutionary artificial immune system optimization. In: Proceedings of the 12th Annual Genetic and Evolutionary Computation Conference, Portland, USA, 2010. 2065-2066

[5] Mabu S, Chen C, Lu N N, et al. An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 2011, 41(1): 130-139

[6] Dasgupta D, Saha S. Password security through negative filtering. In: Proceedings of 2010 International Conference on Emerging Security Technologies, Canterbury, UK, 2010. 83-89

[7] Castro L N D, Timmis J. Artificial immune systems as novel soft computing paradigm. *Soft Computing Journal*, 2003, 7(7): 526-544

[8] Chmielewski A, Wierzchon S T. Simple method of increasing the coverage of nonself region for negative selection algorithms. In: Proceedings of the 6th International Conference on Computer Information Systems and Industrial Management Applications, Elk, Poland, 2007. 155-160

[9] Gonzalez F, Dasgupta D. Anomaly detection using real-valued negative selection. *Genetic Programming and Evolvable Machines*, 2003, 4(4): 383-403

[10] Zhou J, Dasgupta D. Applicability issues of the real-valued negative selection algorithms. In: Proceedings of the Conference on Genetic and Evolutionary Computation 2006, Seattle, USA, 2006. 111-118

[11] Zhou J, Dasgupta D. Real-valued negative selection algorithm with variable-sized detectors. In: Proceedings of the Conference on Genetic and Evolutionary Computation 2004, Seattle, USA, 2004. 287-298

[12] Zhou J, Dasgupta D. V-Detector: an efficient negative selection algorithm with "probably adequate" detector coverage. *Information Science*, 2009, 179(10): 1390-1406

[13] Gui-yang LI, Tao Li, Jie Zeng, et al. An improved v-detector algorithm of identifying boundary. In: Proceedings of the 8th International Conference on Machine Learning and Cybernetics, Baoding, China, 2009. 3209-3214

[14] Li G Y, Li T, Zeng J, et al. Negative selection algorithm based on immune suppression. In: Proceedings of the 8th International Conference on Machine Learning and Cybernetics, Baoding, China, 2009. 3227-3232

[15] 洪征，吴礼发，王元元. 应用改进的 V-detector 算法检测蠕虫. 北京邮电大学学报，2007，30(2)：98-101

[16] Liu F, Gong M G, Ma J J, et al. Optimizing detector distribution in v-detector negative selection using a constrained multiobjective immune algorithm. In: Proceedings of the Congress on Evolutionary computation 2010, Barcelona, Spain, 2010. 1-8

[17] Tew J, Phipps P, Mandel T. The maintenance and regulation of the human immune response: persisting antigen and the role of follicular antigen-binding dendrtic cells. *Immunological Review*, 1980, 53: 175-211

[18] KDD Cup 99 Data. http://kdd.ics.uci.edu/databases/kddcup99/ kddcup99.html

[19] 王大伟. 基于生物免疫的检测器分布策略研究:[博士学位论文]. 哈尔滨:哈尔滨理工大学计算机科学与技术学院，2008. 45-54

A V-detector optimization algorithm

Zhang Fengbin, Xi Liang, Wang Shengwen, Yue Xin

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

Abstract

An optimization algorithm for variable-coverage detectors (V-detectors) was designed based on the immune-cell regulation mechanism in biology immune systems to solve the problems of V-detector hole and high V-detector overlapping of the V-detector algorithm, a real-valued negative selection algorithm with V-detectors. The algorithm updates the detector set by the candidates generated from their parents and the affinity comparison to improve detectors' distribution performance. It was tested by the synthetic data and the KDD CUP 99 data set. The results show that the optimized detectors can increase the efficiency of detectors' distribution and improve the system's detection performance.

Key words: artificial immune system, real-valued, negative selection algorithm with variable- coverage detectors, optimization algorithm