基于多种群遗传算法的检测器生成算法研究

杨东勇 [1] 陈晋音 [2]

摘 要 有效的检测器生成算法是异常检测的核心问题, 针对现有算法存在检测率低、匹配阈值固定、检测器集合庞大等问题, 本文提出了基于多种群遗传算法的检测器生成算法, 根据形态学空间的分析和覆盖问题原理, 自体集根据特征进行划分, 各个种群根据划分独立按遗传算法进化, 最后求得所有检测器种群的并集得到成熟的检测器. 所提出的算法有效降低检测器的冗余度, 减少检测器规模, 保持检测器的多样性; 并利用 $maxSelf$ 实现匹配阈值 r 的自适应, 适用于多种匹配规则, 减小了阈值设置的局限性, 给出了算法的检测率高于传统算法的理论证明, 并通过实验验证了算法的有效性. 另外, 通过统计算法的时间复杂度, 证明算法时间复杂度没有明显增加.

关键词 人工免疫系统, 否定选择, 检测器, 多种群遗传算法, 自适应
中图分类号 TP18

Research on Detector Generation Algorithm Based on Multiple Populations GA

YANG Dong-Yong[1] CHEN Jin-Yin[2]

Abstract Efficient detector generation algorithm is the kernel of anomaly detection. Aiming at low true positive (TP) value, unhandy matching threshold value and large detector set size of existent algorithms, a novel detector generation algorithm based on multiple populations genetic algorithm is put forward in this paper. According to morphologic analysis of intrusion detection system and covering problem principle, self set is divided into several partitions on the basis of their characters. Each population evolves according to each self partition independently and their best populations will be combined as the final matured detector set, which decreases redundancy of detectors, minimizes the size of detector set, and maintains diversity of detectors. Matching threshold r is self-adaptive according to $maxSelf$ which enlarges application area of the algorithm by applying several matching rules. The TP value is improved compared with traditional algorithm through theoretical proof and efficiency of the algorithm is testified by simulation tests. Time complexity of the algorithm is analyzed and the algorithm does not have a significant time complexity increase.

Key words Artificial immune system, negative selection, detector, multiple populations genetic algorithm, self-adaptive

生物免疫系统是一个自适应、自组织和自学习系统, 具有很强的自我保护功能[1], 人工免疫系统模拟生物免疫系统并成功应用于网络入侵检测领域[2]. 检测器生成算法是检测系统的关键部分, 直接决定了检测器的检测性能和效率. 1995 年 Forrest 提出利用否定选择算法 (Negative selection algorithm) 实现检测器的耐受[3], 此后, 如何设计高效的智能检测器模型已逐渐成为一个研究热点.

否定选择算法是根据生物免疫系统中 T 细胞的产生和作用机理而提出的一种检测器生成算法, 在异常检测中起着重要作用[4]. 否定选择算法是在随机产生预检测器的基础上利用否定选择来完成检测器的自体耐受过程, 因此检测器的生成效率和检测效率都是至关重要的. Forrest 等最早提出的否定选择算法所使用的是穷举检测器生成算法[3]. 为了克服检测器产生算法过大的冗余度问题, D′haeseleer 等提出线性检测器生成算法和贪心检测器生成算法[5]. 前者所耗时间分别与自体集合大小和检测器集合大小呈线性关系, 但仍有冗余; 后者可消除冗余, 但不能使检测器生成时间最小化. 随后利用其他技术来提高检测器的生成效率, 利用匹配阈值矩阵来设置各个检测器不同的匹配阈值, 从而提高检测器的检测率[6], 检测器的否定选择依然需要预设阈值来实现自体耐受, 产生检测器的算法性能没有提高, 而只是针对不同检测器个体采用不同的匹配阈值. Kim 等分析了否定选择算法在检测器生成过程中的贡献[7], 认为否定选择算法可以有效过滤无效检测器, 但单纯基于否定选择算法生成检测器的效率会比较低, 主要是因为初始阶段随机产生预检测器, 随后没有一定的进化措施, 从而导致生成算法效率较低. 此外, 还有多种改进否定选择算法的研究[8−9]. 总体来说, 检测器生成算法面临以下几个问题:

1) 需要比较大的检测器集合来保证较高的检测

收稿日期 2008-01-08 收修改稿日期 2008-03-31
Received January 8, 2008; in revised form March 31, 2008
浙江省自然科学基金 (Y106735) 资助
Supported by the Natural Science Foundation of Zhejiang Province (Y106735)
1. 浙江工业大学软件学院 杭州 310032 2. 浙江工业大学信息工程学院 杭州 310032
1. College of Software, Zhejiang University of Technology, Hangzhou 310032 2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310032
DOI: 10.3724/SP.J.1004.2009.00425

率, 导致每次检测器生成过程和检测过程花费较多时间.

2) 否定选择算法中针对一定的匹配规则, 例如海明距离规则、r-contiguous 连续匹配规则、r-chunck 距离规则等, 都存在匹配阈值的概念, 目前都没有很好地解决阈值的问题, 预设常数阈值对不同的应用领域存在局限性.

3) 形态学空间中分析否定选择算法, 针对特定的匹配规则没有展开深入的研究, 有待进一步的理论与实验研究.

基于检测器生成算法的研究现状, 本文采用形态学空间方法分析了否定选择算法, 提出了利用多种群遗传算法进化检测器集合 (Multiple population T-cell maturation algorithm, MPTMA), 并利用覆盖圆环的内外半径来实现匹配阈值的自适应调节, 其核心思想主要如下: 1) 根据自体集中个体的相似程度, 利用集合划分的概念分成若干个自体子集合, 利用相应个数的检测器子种群分别进化特征明显的检测器子集合, 从而降低检测器的冗余, 减少检测器个数, 同时保持检测器特征多样性; 2) 提出利用二进制串的覆盖问题来解决检测器最小化的问题, 利用尽可能小的检测器集合尽可能覆盖所有的自体集; 3) 检测器进化过程中设置 $r = maxSelf + 1$ 实现匹配阈值的自适应, 适用于多种匹配规则. 本文的主要贡献在于: 1) 在形态学空间分析检测器中抗体和抗原, 利用覆盖问题来模拟检测器生成过程, 即利用圆环模拟检测器覆盖分散在圆内的无规则抗原; 2) 提出了基于多种群遗传算法的检测器生成的定理, 给出了理论证明, 通过仿真实验验证了提出算法的有效性, 并和现有几种高效的检测器生成算法进行了比较; 3) 讨论了多种群遗传算法适用的匹配规则, 探讨其普适性.

1 否定选择算法及其分析

1.1 否定选择算法简介

否定选择算法分两步[10]: 1) 产生预检测器. 根据正常数据生成自体集, 随机产生预检测器串, 如果预检测器个体与自体集匹配则删除, 否则保留下来作为成熟的检测器, 重复以上步骤直至产生足够大的检测器集合为止. 2) 利用成熟的检测器集合检测网络采集的数据, 如果与检测器中个体匹配则认为该数据是异常数据. 免疫事件都在形态空间 (Shape-space) S 中发生[11], 数学上, 这种形态被描述成 L 维字符串或者向量. 因此在形态空间内有一个体积是 V 的区域, 含有抗体和抗原形状互补区域, 抗体识别抗原的过程就是与抗原匹配并结合的过程, 由于存在匹配阈值的概念, 所以这种匹配不一定要求二者完全精确匹配, 只要这种匹配所导致的亲和力大于某一固定的阈值即可. 从而可以得出结论, 利用抗体及其覆盖区域来匹配尽可能多的抗原, 使得检测率最高.

图 1 中 X 表示抗原, 整个大圆表示空间 V, 而里面的小圆表示每个抗体及其匹配范围, 半径是 e, n 表示抗体个数, 用 V 表示整个空间 S, 用 Ve 来表示抗原覆盖空间, 则每一个抗原一般由 nVe/V 个不同的抗体匹配识别, 而不能被识别的概率 P 为

$$P = \left(1 - \frac{Ve}{V}\right)^n \approx e^{-\frac{nVe}{V}} \tag{1}$$

因此, 在 V 一定的情况下要降低 P 就需要增大 Ve, 但是检测器集合的大小又不可能无限制的增大, 因此有必要提出一种方法利用尽可能小的检测器集合来覆盖尽可能大的抗原空间.

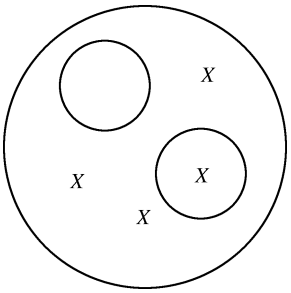


图 1 形态空间

Fig. 1 Morphological space

1.2 检测器串覆盖问题分析

基于对形态空间的分析, 本文提出利用检测器串覆盖问题的思想来生成高效的检测器集合. 在文献 [12] 中提到利用区域匹配规则来进化检测器, 如图 2 所示, 其遗传算法中用到的适应度函数定义为: $fitness = maxSelf - minSelf$, 其中 $maxSelf$ 和 $minSelf$ 是检测器与自体集的最大最小亲和度, 即将圆环的内外半径差作为优化目标. 实际上如果将检测器生成问题归结为二进制串的覆盖问题时, 这样定义目标函数存在不合理因素, 要求覆盖圆环的

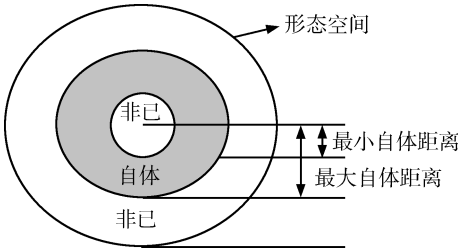


图 2 形态空间中检测器集合用圆环模拟[12]

Fig. 2 Detectors in morphological space are simulated as circles[12]

最小面积应该通过移动圆环圆心和改变内外半径来实现, 因此有必要寻求更加合理的解决方案.

检测器集合需要极大覆盖自体集, 同时最小化检测器集合, 本文将此问题归结为基于二进制串的覆盖问题. 二进制串看作圆中的点, 则自体集个体就是分散在圆内的若干分散点, 检测器集合是圆环, 利用圆环尽可能去覆盖所有的自体集分散点, 同时需要最小化圆环的面积. 集合和几何空间中的覆盖问题是一个 NP 完全问题, 只能得到近似解.

通常的最小化覆盖圆环面积的方法是通过移动圆环圆心和改变圆环的内外半径来求近似最优解, 本文中提出的思路是: 首先根据自体集中个体的相似度, 将它们进行划分, 在 S 空间中表现出来的则是将距离较近的自体作为一个划分; 然后针对每一个划分利用一个检测器子集的圆环去覆盖, 最后将圆环的并集作为最优的检测器集合, 如图 3 所示.

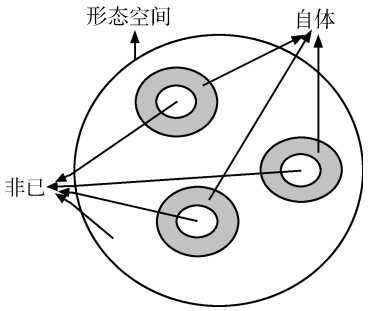


图 3　基于自体集划分的检测器圆环

Fig. 3　Detector circles are divided by their self sets

2　多种群检测器生成算法

2.1　算法简介

本文算法的基本思想是利用自体集各个划分不同特征作为各个检测器子集的进化目标, 从而保持了检测器的多样性特征, 降低了检测器的冗余度.

算法设计思想主要体现在下述几个方面: 1) 有针对性地生成各个检测器子圆环覆盖, 减少遗漏的自体个体; 2) 各个子圆环的并覆盖小于等于所有圆环的覆盖之和, 可以得到近似最优解. 算法的流程图如图 4 所示. 图 4 中阴影框是两个重要的操作, "自体分类" 完成自体集的划分过程, 分成若干个子集合, 然后每个子集合作为独立的自体集, 基于遗传算法的 T 细胞生成算法 (Genetic algorithm based T-call maturation algorithm, GATMA) 生成检测器, 虚框内是遗传算法的操作, 最后将生成的最优检测器个体求并, 即完成 "结合" 操作, 生成最终完整检测器集合.

算法具体步骤如下:

步骤 1. 准备. a) 首先确定自体集/非自体集的描述形式; b) 定义自体, 也称自体集获取, 确定自体集的范围, 构造自体集; c) 定义亲和度的计算方法.

步骤 2. 初始化. 产生初始的预检测器集合, 每个检测器都是一个字符串.

步骤 3. 自体集划分. 根据一定划分规则 (见第 3.3 节中的具体分析) 划分成 N 个自体子集分别作为 N 个检测器种群的自体集.

步骤 4. 每个子种群实现基于遗传算法的检测器生成算法的过程. 将检测器集合作为种群, 检测器个体就是种群中的个体, 按照 SGA 的步骤, 初始化种群、选择、交叉和变异, 直至达到优化精度要求或者达到最大进化代数.

步骤 5. 合并最优检测器种群. 将各个检测器种群的最优种群的并集作为成熟的检测器集合.

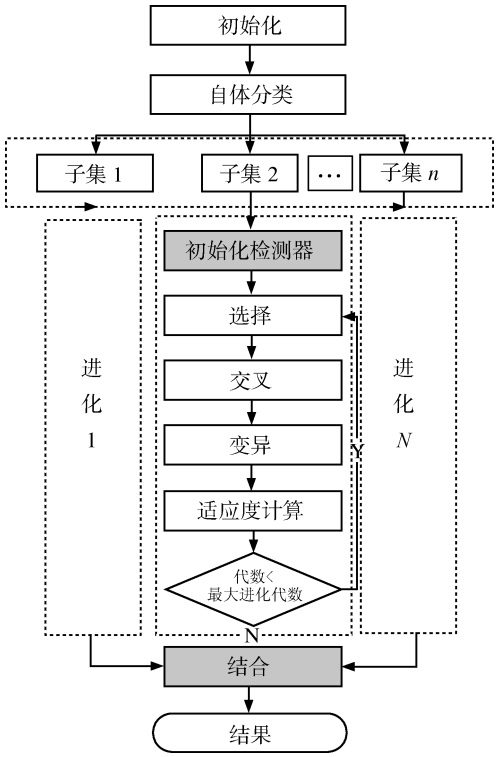


图 4　基于多种群遗传算法的检测器生成算法流程图

Fig. 4　Flow chart of detector generation algorithm based on multiple population GA

2.2　算法的理论基础

根据以上设计的检测器生成算法, 分析其算法的性能, 根据文献 [10] 中的定义 1～3, 得出定理 1, 并给出定理 1 的证明.

定义 1. 自然免疫系统表示为 $\sum_{\mathrm{NIS}} = (X_{\mathrm{NIS}}, \Omega_{\mathrm{NIS}}, \Upsilon_{\mathrm{NIS}}, G_{\mathrm{NIS}})$, 其中 X_{NIS} 为自然免疫系统的输入, 包括抗体和抗原, 令 E 表示抗原全体, 则整个抗原全体包括两个互斥的集合, 即自身蛋白集合和病原体集合, 用 S 来表示自体蛋白集合, NS 表示病

原体集合, 则有 $S \cup NS = E$, $S \cap NS = \emptyset$, 而输入 $X_{\rm NIS} \in E$, $\Upsilon_{\rm NIS}$ 为自然免疫系统的输出, 若输入为自身, 则自然免疫系统输出 $\Upsilon_{\rm NIS} = 0$, 否则 $\Upsilon_{\rm NIS} = 1$. $G_{\rm NIS}$ 表示自然免疫系统输入与输出关系的非线性函数, 有

$$\Upsilon_{\rm NIS}=G_{\rm NIS}(X_{\rm NIS})=\begin{cases}1, & 若\ X_{\rm NIS}\in NS\\ 0, & 若\ X_{\rm NIS}\in S\end{cases} \tag{2}$$

定义 2. 人工免疫系统表示为 $\sum_{\rm AIDS}=(X_{\rm AIDS}, \Omega_{\rm AIDS}, \Upsilon_{\rm AIDS}, G_{\rm AIDS})$, 其中 $X_{\rm AIDS}$ 为 $\sum_{\rm AIDS}$ 的输入, 它可能为操作系统日志, 也可能为网络数据包. 输入的论域与一般的入侵监测系统同为 W, 整个论域划分为两个互斥的集合, 即入侵集合 I 和自体集合 S, 且有 $X_{\rm AIDS} \in W$. 其中, $\Upsilon_{\rm AIDS}$ 为 $\sum_{\rm AIDS}$ 的输出, $\Upsilon_{\rm AIDS}$ 取 0 或 1 分别表示 $\sum_{\rm AIDS}$ 不报警或报警.

定义 3. 入侵检测系统的检测率可以表示为 $R_{\rm TD}(\sum_{\rm AIDS}) = P(\Upsilon_{\rm AIDS} = 1/X_{\rm AIDS} \in I)$, 漏测率表示为

$$R_{\rm MD}(\Sigma_{\rm AIDS}) = P\left(\Upsilon_{\rm AIDS} = \frac{0}{X_{\rm AIDS}\in I}\right) = 1 - R_{\rm TD}(\Sigma_{\rm AIDS}) \tag{3}$$

误测率表示为

$$R_{\rm FD}(\Sigma_{\rm AIDS}) = P\left(\Upsilon_{\rm AIDS} = \frac{1}{X_{\rm AIDS}\in \bar{I}}\right) \tag{4}$$

根据上述定义, 对本文提出的检测器生成算法进行分析, 得到以下结论: 基于多种群遗传算法进化的检测器模型采用多个圆环覆盖自体集的思想, 比传统的检测器具有更好的检测性能, 由此归纳得到定理 1.

定理 1. 单个圆环覆盖自体集生成检测器的检测率是 $R_{\rm TD}(D_i)$, 多个圆环覆盖自体集生成检测器的检测率是 $R_{\rm TD}(\sum_{\rm AIDS})$, 则有 $R_{\rm TD}(\sum_{\rm AIDS}) > R_{\rm TD}(D_i)$.

证明. 因为 $\sum_{\rm AIDS}$ 的 $(\Omega_{\rm AIDS}) = D_1, D_2, \cdots, D_n$, 对于不同检测器种群, 各自分别有

$$R_{\rm TD}(D_i) = P\left(\Upsilon_{D_i} = \frac{1}{X_{\rm AIDS}\in I}\right), \quad i = 1, 2, \cdots, n \tag{5}$$

$$R_{\rm FD}(D_i) = P\left(\Upsilon_{D_i} = \frac{1}{X_{\rm AIDS}\in \bar{I}}\right), \quad i = 1, 2, \cdots, n \tag{6}$$

因为各个检测器在入侵检测的时候是并联互补的, 即

$$R_{\rm TD}(\Sigma_{\rm AIDS}) = P\left(\Upsilon_{\rm AIDS} = \frac{1}{X_{\rm AIDS}\in I}\right) = P\left(\frac{\{\Upsilon_{D_1}=1\}\cup\{\Upsilon_{D_2}=1\}\cup\cdots\cup\{\Upsilon_{D_n}=1\}}{X_{\rm AIDS}}\right)\in I, \quad i = 1, 2, \cdots, n \tag{7}$$

且 $\{\Upsilon_{D_i} = 1\} \subset \{\Upsilon_{D_1} = 1\} \cup \{\Upsilon_{D_2} = 1\} \cup \cdots \cup \{\Upsilon_{D_n} = 1\}$, 所以

$$R_{\rm TD}(\Sigma_{\rm AIDS}) > R_{\rm TD}(D_i), \quad i = 1, 2, \cdots, n \tag{8}$$

所以基于多种群遗传算法进化的检测器模型比传统的基于遗传算法生成的检测器具有更好的检测率. □

2.3 算法关键技术分析

2.3.1 自体集划分

借鉴形态学空间的概念, 以自体集合中各个自体间的距离来决定它们是否属于一个划分. 划分有两种方法: 1) 阈值法. 设定距离阈值 R, 个体间的距离根据匹配规则计算, 在集合中随机选择一个个体 E 作为一个划分中的一个元素, 遍历剩余的元素, 如果与 E 的距离小于 R, 则将其作为该划分, 否则排除; 依此类推. 其中划分的个数与阈值 R 相关, R 越大则划分的个数越小, 反之亦然. 2) 等划分法. 设定划分个数为 N, 则每个划分中的元素个数 n 是固定的, 即 $n =$ 自体个数$/N$, 在集合中随机选择一个个体 E 作为一个划分中的一个元素, 遍历剩下的元素, 分别计算它们与 E 的距离, 选取距离最小的 $n-1$ 个作为一个划分, 依此类推.

2.3.2 匹配规则和匹配阈值自适应

本文提出的算法适用于多种匹配规则的检测器生成, 例如海明距离、r-contiguous 连续匹配规则、r-chunck 距离规则等, 它们同属于计算抗原和抗体的对应位相似度匹配, 因此都可以等同于形态学的距离概念.

匹配阈值 r 通过检测器进化过程中 $maxSelf$ 的改变而调整, 计算公式是

$$r = maxSelf + 1 \tag{9}$$

这样就克服了由匹配规则和预设阈值 r 带来的局限性, 扩大了算法的应用领域.

2.4 检测器生成算法特例

基于多种群遗传算法的检测器生成算法是基于遗传算法生成检测器的推广. 从原理上分析, 基于遗传算法的形态学原理是单圆环覆盖自体集, 是多圆环覆盖自体集情况下圆环个数等于 1 的特例; 另一方面, 若多种群遗传算法的种群个数 $popsize = 1$, 则其退化为基于遗传算法的检测器生成算法. 因此, 多种已有检测器生成算法是本文算法的特例.

3 仿真实验

在实际应用时, 可以根据不同的需要, 按不同的方式具体实现算法的各个步骤, 本文给出了一个基于二进制串来描述算法的具体实现范例以验证算法的性能, 并给出了实际测试结果.

3.1 实验设计

本文算法的关键是自体集的划分和各个种群基于遗传算法的检测器进化过程, 为了测试本文提出算法的性能, 将其与其他若干种现有的典型检测器生成算法进行比较. 比较的算法包括[12]: 否定选择算法 (NSA)、基于遗传算法的 T 细胞生成算法 (GATMA)、改进的基于遗传算法的 T 细胞生成算法 (IGATMA)、在线进化的 T 细胞生成算法 (LGATMA) 和基于匹配区域的 T 细胞生成算法 (MRMTMA).

实验中, 测试用自体集如表 1 所示, 算法的参数如表 2 所示.

表 1 测试用自体集 $S1$ 和 $S2$

Table 1 Self sets of $S1$ and $S2$

Pattern	$S1$	$S2$
1111**********	2	4
****1111********	2	4
********1111****	2	4
**********1111	2	4

自体集分别是 $S1$, $S2$, 第三种情况是先将 $S1$ 作为自体集, 然后在进化到 $gen = 1\,000$ 时动态加入 $S2$, 使得 $S1 + S2$ 作为自体集, 测试算法的动态检测能力, 对应动态网络的鲁棒性能.

3.2 实验结果

综合评价本文提出算法的性能, 从以下几个方面进行测试, 独立运行每种算法 20 次, 统计平均值和方差.

1) 检测率和漏测率

检测率是利用已知入侵攻击的实验数据集合来测试系统的一个指标. 几种检测器生成算法的检测率的平均最优值如表 3 所示. MRMTMA 1 把新个体直接作为下一代个体, 而 MRMTMA 2 用父子混合选产生新一代个体. 各种算法的检测率 TP 如图 5 (见下页) 所示. 图 5 (a) 中最底下的两条线分别是 NSA 中阈值 r 取 7 和 8 时的结果, 可以看到 MPTMA 的检测率接近 1.0, 远远优于其他算法. 当自体集是图 4 (b) 情况时, MPTMA 的检测率非常好, 接近于 1.0; 当自体集是图 4 (c) 情况时, 是一个动态改变自体集的情况, 在 $gen = 1\,000$ 的时候加入了 Set 2, 将 Set 1 + Set 2 作为自体集, 使得检测器集合的适应度改变, 从而改变了进化的方向, 因此检测率发生跌落, 但是从图中可以看到 MPTMA 很快恢复到接近 0.95 的检测率, 而其他生成算法则无法达到.

漏测率 Pm (Missing positive) 是检测率的互补量, 通常可以根据检测率计算得到. 漏测率的实验结果如表 4 (见下页) 所示, 与理论计算相吻合.

表 2 各种算法参数设置 (r 是预设匹配阈值)

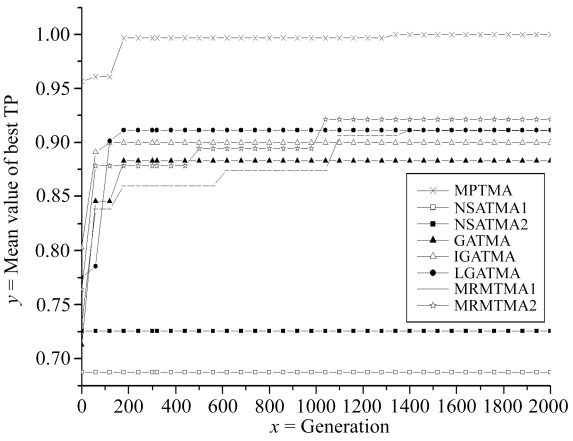
Table 2 Parameters settings (r is the pre-set matching threshold.)

Parameters	r	Detector size	Crossover probability	Missing positive	maxGen	popNum
NSA	7, 8	5	—	—	—	—
GATMA	—	6	0.3	0.05	2 000	—
IGATMA	—	6	0.3	0.05	2 000	—
LGATMA	—	6	0.3	0.05	2 000	—
MRMTMA	—	6	0.3	0.05	2 000	—
MPTMA	—	4	0.3	0.05	2 000	2

表 3 各种不同检测器生成算法的 TP 比较

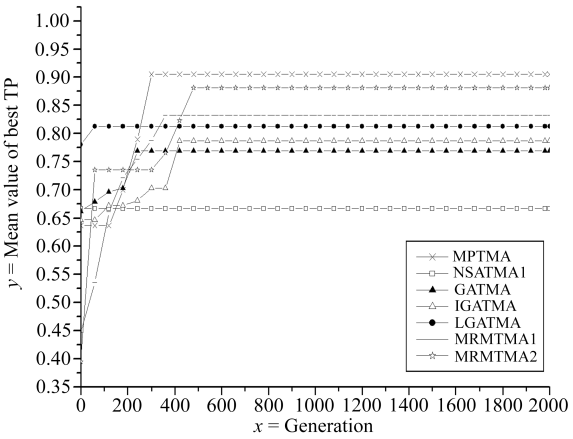
Table 3 TP value comparison of different detector generation algorithms

TMAs	Set 1	Set 2	Set 1 + Set 2
MPTMA	1.0 ± 1.20E-6	0.9052 ± 1.05E-5	0.92603 ± 2.36E-3
NSA ($r = 8$)	0.6875 ± 1.83E-2	0.6662 ± 2.65E-2	0.6129 ± 3.69E-2
NSA ($r = 7$)	0.7256 ± 1.81E-3	—	—
GATMA	0.8829 ± 2.34E-3	0.7688 ± 2.30E-3	0.5708 ± 3.34E-2
IGATMA	0.9123 ± 1.30E-3	0.7868 ± 3.33E-3	0.6253 ± 4.66E-2
LGATMA	0.9112 ± 1.43E-3	0.8125 ± 1.52E-3	0.6951 ± 3.35E-2
MRMTMA 1	0.9110 ± 1.01E-3	0.8322 ± 2.11E-3	0.7023 ± 2.54E-2
MRMTMA 2	0.9213 ± 1.02E-3	0.8812 ± 2.32E-3	0.6952 ± 1.05E-2



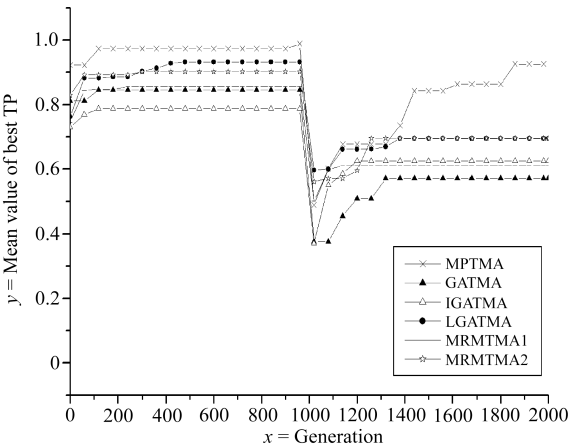
(a) 自体集是 Set 1 时, 八种检测器生成算法的 TP 比较图

(a) TP value comparison of eight detector generation algorithms when self set is Set 1



(b) 自体集是 Set 2 时, 八种检测器生成算法的 TP 比较图

(b) TP value comparison of eight detector generation algorithms when self set is Set 2



(c) 自体集是 Set 1 + Set 2, 八种检测器生成算法的 TP 比较图

(c) TP value comparion of eight detector generation algorithms when self set is Set 1 + Set 2

图 5 检测率的比较

Fig. 5 TP value comparison of all algorithms

2) 误测率

误测率 FP (False positive) 是把正常行为作为入侵报错的概率, 是检测系统的容错能力和准确性指标之一. 几种算法的误测率见表 5 (见下页).

3) 匹配阈值 r 的自适应调节

由于本文提出的 MPTMA 算法中利用式 (9) 实现了阈值 r 的自适应调节, 实验过程中随着自体集从 Set 1 到 Set 1 + Set 2 的改变, r 随着 $maxSelf$ 的调整也相应调整, 图 6 显示了检测器对应 $maxSelf$ 和 r 在不同进化阶段的适应过程.

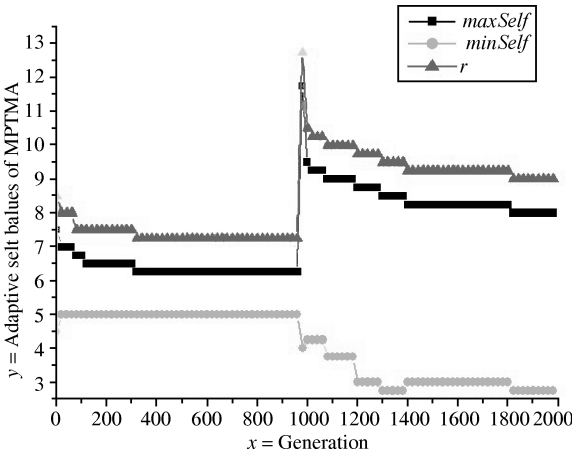


图 6 r 和 $maxSelf$ 在 MPTMA 中实现自适应

Fig. 6 r and $maxSelf$ are self-adaptive in MPTMA

从图中看到阈值 r 不用预设, 它根据 $maxSelf$ 的改变自动调整, $maxSelf$ 表示检测器圆环的外半径逐渐减小的趋势, 符合覆盖问题的原理, 同时满足尽可能覆盖所有的自体个体, 实现了检测器的自体耐受, 这样基本克服了现存检测器中阈值局限性的问题.

4) 算法时间复杂度

已有的各种否定选择算法虽然在检测率上有所提高, 但是时间复杂度大大增加, MPTMA 和其他各种算法的时间复杂度比较如图 7 (见下页) 所示. 从图中可以看到 MPTMA 除了比 NSA 的时间消耗多, 比其他几种进化算法的时间复杂度都低.

3.3 结果分析

MPTMA利用多种群并列遗传算法进化, 从表 1 中知道 MPTMA 的检测器集合是最小的, 克服了文献 [8] 中指出的检测器规模庞大的问题, 另外由于存在自体集划分的操作, 因此需要分析自体集划分的稳定性和单一性. 本文实验分别基于三种不同自体集情况各运行 MPTMA 算法 20 次, 然后统计划分情况.

1) 自体集是 Set 1. 划分有三种: c11, c12, c13. 运行 20 次统计得到 c11, c12, c13 的划分次数比例

表 4　各种不同检测器生成算法的 Pm 比较

Table 4　Pm value comparison of different detector generation algorithms

TMAs	Set 1	Set 2	Set 1 + Set 2
MPTMA	1.5261E-6 ± 1.2E-7	0.0948 ± 1.05E-5	0.087 ± 3.63E-4
NSA ($r = 8$)	0.3124 ± 1.83E-2	0.3335 ± 2.65E-2	0.3871 ± 4.30E-2
NSA ($r = 7$)	0.2743 ± 1.87E-3	—	—
GATMA	0.1172 ± 2.23E-3	0.2312 ± 2.30E-3	0.4292 ± 1.65E-1
IGATMA	0.1001 ± 1.33E-3	0.2132 ± 3.02E-3	0.3747 ± 4.66E-2
LGATMA	0.0887 ± 1.32E-3	0.1875 ± 1.52E-3	0.3049 ± 3.65E-2
MRMTMA 1	0.0888 ± 1.01E-3	0.1678 ± 3.20E-3	0.2977 ± 3.35E-2
MRMTMA 2	0.0787 ± 1.02E-3	0.1188 ± 2.31E-3	0.3048 ± 2.98E-2

表 5　各种不同检测器生成算法的 FP 比较

Table 5　FP value comparison of different detector generation algorithms

TMAs	Set 1	Set 2	Set 1 + Set 2
MPTMA	1.22E-5 ± 1.53E-6	4.58E-5 ± 1.50E-6	1.67E-4 ± 5.33E-5
NSA ($r = 8$)	0 ± 0.0	0 ± 0.0	0 ± 0.0
NSA($r = 7$)	0 ± 0.0	—	—
GATMA	2.25E-4 ± 1.02E-5	1.23E-4 ± 1.55E-5	2.39E-4 ± 5.64E-5
IGATMA	2.14E-4 ± 2.12E-5	5.22E-4 ± 1.66E-5	1.57E-4 ± 2.31E-5
LGATMA	1.03E-4 ± 1.22E-5	1.00E-4 ± 2.33E-5	2.56E-4 ± 5.32E-5
MRMTMA 1	1.56E-4 ± 2.13E-5	2.11E-4 ± 2.02E-5	1.42E-4 ± 3.21E-5
MRMTMA 2	2.01E-4 ± 1.34E-5	2.18E-4 ± 1.36E-5	2.41E-4 ± 3.35E-5

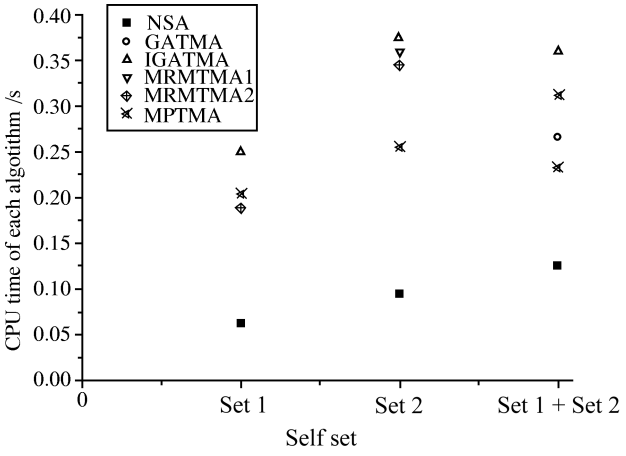


图 7　算法时间复杂度比较

Fig. 7　Time complexity comparison

为 15 : 4 : 1, 划分是 c11 的概率为 0.75.

2) 自体集是 Set 2. 划分有五种 c21, c22, c23, c24, c25. 得到划分的比例为 12 : 2 : 2 : 3 : 1, 划分是 c21 的概率为 0.60.

3) 自体集是 Set 1 + Set 2. 划分有五种 c31, c32, c33, c34, c35. 得到划分的比例为 13 : 1 : 4 : 2 : 1, 划分是 c31 的概率为 0.65.

从情况 1) ∼ 3) 的概率分布来看, 自体集的划分是不唯一的, 划分的可能性和自体集有关, 自体集越大可能性越多, 自体集的划分基本是稳定的, 可能性较大的一种划分可能发生概率在 0.60 以上, 基本可以将这种划分作为估计的划分可能. 根据分析, 现有的大部分检测器生成算法的时间复杂度与检测器大小成指数级增长, 这是一个不可忽视的问题[8], MPTMA 的时间复杂度与检测器大小的关系如图 8 所示, 由图 8 可以看到在检测器大小增加到 16 以后, 时间复杂度没有明显变化, 趋于稳定, 克服了检测器规模庞大问题.

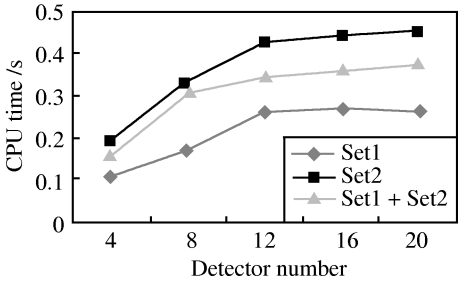


图 8　MPTMA 的时间复杂度统计

Fig. 8　Time complexity of MPTMA

基于多种群遗传算法的检测器生成算法与文献

[8] 提出的自适应检测器生成算法相比, 两者都是基于否定选择算法, 本文提出的算法能自适应调节匹配阈值, 并利用多种群最小化检测器集合; 而文献 [8] 是利用亲和度变异的否定选择算法实现的, 虽然两者均实现了检测器集合在满足高检测率条件下的最小化, 但所采用的策略完全不同. 本文在研究了文献 [8] 的工作后, 从一个全新的角度, 即形态学空间的角度进行了分析. 实验也证明了算法的有效性以及在时间复杂度上的优势.

4 结束语

检测器的高效生成问题是当前基于人工免疫的异常检测研究中的一个重点和难点. 本文提出了基于多种群遗传算法的检测器生成算法, 并给出了算法性能的理论证明, 该算法具有检测器集合较小, 检测率高, 时间复杂度低, 匹配阈值 r 自适应等特点, 实验结果也表明了算法的有效性, 适用于多种应用领域. 下一步工作将是进行本文算法的具体应用研究, 并对其进行更深入的理论分析.

References

1 Kim J, Bentley P J. The human immune system and network intrusion detection. In: Proceedings of the 7th European Congress on Intelligent Techniques and Soft Computing. Aachen, Germany: EUFIT, 1999. 1120−1125

2 Luther K, Bye R, Alpcan T, Muller A, Albayrak S. A cooperative AIS framework for intrusion detection. In: Proceedings of IEEE International Conference on Communications. Washington D. C., USA: IEEE, 2007. 1409−1416

3 Forrest S, Perelson A S, Allen L, Cherukuri R. Self-nonself discrimination in a computer. In: Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy. Oakland, USA: IEEE, 1994. 202−212

4 Kim J, Bentley P J, Aickelin U, Greensmith J, Tedesco G, Twycross J. Immune system approaches to intrusion detection – a review. In: Proceedings of the 3rd International Conference on Artificial Immune Systems. Catania, Australia: Kluwer Academic Publishers, 2004. 316−329

5 D′haeseleer P, Forrest S, Helman P. An immunological approach to change detection: algorithms, analysisand implications. In: Proceedings of IEEE Symposium on Security and Privacy. Oakland, USA: IEEE, 1996. 110−119

6 Luo W J, Wang X, Tan Y, Wang X F. A novel negative selection algorithm with an array of partial matching lengths for each detector. In: Proceedings of the 9th International Conference on Parallel Problem Solving from Nature. Reykjavik, Iceland: Springer, 2006. 112−121

7 Kim J, Bentley P J. An evolutionary of negative selection in an artificial immune system for network intrusion detection. In: Proceedings of the Genetic and Evolutionary Computation Conference. San Francisco, USA: Morgan Kaufmann Publishers, 2001. 1330−1337

8 Luo Wen-Jian, Cao Xian-Bin, Wang Xu-Fa. Research on adaptively generating detector algorithm. *Acta Automatica Sinica*, 2005, **31**(6): 907−916
(罗文坚, 曹先彬, 王煦法. 检测器自适应生成算法研究. 自动化学报, 2005, **31**(6): 907−916)

9 Li T. An immune based dynamic intrusion detection model. *Chinese Science Bulletin*, 2005, **50**(22): 2650−2657

10 Jiao Li-Cheng, Du Hai-Feng, Liu Fang, Gong Mao-Guo. *Immune Optimizing Algorithm, Learning and Detection*. Beijing: Science Press, 2006. 350−363
(焦李成, 杜海峰, 刘芳, 公茂果. 免疫优化计算、学习与识别. 北京: 科学出版社, 2006. 350−363)

11 Li Tao. *Computer Immune Algorithms*. Beijing: Publishing House of Electronics Industry, 2004. 39−41
(李涛. 计算机免疫学. 北京: 电子工业出版社, 2004. 39−41)

12 Chen Jun-Gan. Intrusion Detector Maturation Algorithm Based on Artificial Immune System [Master dissertation], Zhejiang University of Technology, China, 2005
(陈军敢. 基于人工免疫系统的入侵检测器生成算法研究 [硕士学位论文], 浙江工业大学, 中国, 2005)

杨东勇 浙江工业大学软件学院教授. 主要研究方向为人工智能, 网络安全, 全方位视觉系统和多智能体系统.
E-mail: ydy@zjut.edu.cn
(**YANG Dong-Yong** Professor in the College of Software, Zhejiang University of Technology. His research interest covers artificial intelligence, network security, ominidirectional vision system, and multi-agents system.)

陈晋音 浙江工业大学信息工程学院控制理论与控制工程专业博士研究生. 主要研究方向为多智能体系统, 机器人路径规划, 智能算法和网络安全. 本文通信作者. E-mail: chenjinyin@163.com
(**CHEN Jin-Yin** Ph. D. candidate in control theory and control engineering at the College of Information Engineering, Zhejiang University. Her research interest covers multiple agent system, robot path planning, intelligent algorithms, and network security. Corresponding author of this paper.)