高技术通讯 2009 年 第 19 卷 第 3 期:273 ~ 278

doi:10.3772/j.issn.1002-0470.2009.03.011

种子检测器刺激-应答变异算法研究①

刘星宝②*** 蔡自兴*

(*中南大学信息科学与工程学院 长沙 410083)

(**湖南商学院现代技术教育中心 长沙 410025)

摘 要 为降低阴性选择算法(NSA)的时间复杂度,提出了一种应用种子个体连续位刺激变异的检测器生成策略:首先随机生成种子检测器集合,根据其与自体的亲和度选定变异个体和变异片段;其次在被选个体的特定基因片段发生刺激-应答变异(SRM),产生新的候选检测器个体;最后应用 r 位连续匹配准则筛选候选个体生成新的检测器。该策略的算法特点在于利用种子个体和自体集合的模式信息指导变异过程,降低候选检测器与自体的匹配成功率。实验表明,在保持高检测率的同时,种子检测器 SRM 算法比穷举算法、个体随机变异算法和检测器连续胞体超变异(CSH)算法的生成效率更高。

关键词 人工免疫系统,阴性选择算法,r-连续位匹配,亲和度测量,刺激-应答变异(SRM)

0 引 言

阴性选择算法(negative selection algorithm, NSA)是人工免疫系统的重要原则之一,降低其时间复杂度是目前该领域的热点问题[1]。Forrest S 最初给出了检测器生成的方法[2],该方法需要搜索的空间和系统自体状态空间成指数关系,致使生成算法需要耗费大量的时间和空间资源。Forrest S 等人又提出了针对特定部分匹配策略的改进方法,能够在线性时间内生成检测器,并从理论上探讨了这个问题[3,4],但是该方法实现的难度很大。郭振河引入进化算法的思想,提出探测器的概念[5],给出了位变异算法和余数生长两种算法,由于应用了个体变异,缩短了检测器的生成时间。何申用剪枝的方法生成检测器[6],该方法用高为编码长度的完全二叉树表示全体空间,通过剪枝减少搜索次数,以减少检测器的生成时间,明显降低了时间复杂度,但空间耗费和编码长度成指数关系。

以上工作除文献[6]外,采用的生成方法大多是随机搜索及其变种,变异操作的引入缩短了检测器的生成时间。但是在变异过程中可以进一步改进下面两点,以提高检测器的生成效率:(1) 传统的变异策略使用的候选检测器数目仍然非常大,可以尝试通过其它的变异策略进一步减少使用的候选检测器;(2) 变异过程没有利用历史信息和个体与自体集合之间的模式信息指导。基于以上考虑,本文提出应用刺激-应答变异(simulated-response mutation, SRM)的检测器生成策略,即:首先通过随机方法生成种子检测器集合,依据与自体集合的亲和度选定参加变异的种子个体;对种子的特定片段给予刺激,使该片段发生一系列变异反应,从而产生新的候选检测器个体,然后应用 r 位连续匹配生成新的检测器集合。该算法的特点在于通过特定检测器的刺激-应答变异产生新的个体,减少了候选检测器的数目。

1 问题定义

1.1 相关定义

在以下的研究中,我们采用{0,1}二进制编码。个体的编码长度为 L,检测器集合 *Detectors* 是异体集合的真子集。阴性选择算法的研究目标是在自体和异体之间建立一个匹配映射关系,快速地找出检测器集合 *Detectors*,使得检测器 *Detectors* 中的个体和自体集合不匹配,并且覆盖尽可能多的异体。我们首先定义相关的概念。

定义 1 若 X、Y 是长度相同的二进制字符串

① 国家自然科学基金(60404021,60234030)和国家基础研究(A1420060159)资助项目。

② 男,1977 年生,博士生;研究方向:人工免疫系统,进化计算; E-mail:liuxb0608@gmail.com

(收稿日期:2008-04-21)

集合，X 与 Y 的亲和度定义为对应位置相同的位数，记作 $affi(X,Y)=\max\{f(X,s)\mid s\in Y\}$。

定义 2 最大连续位匹配：如果 X 和 Y 从 p_1 位开始有连续的 k_1 位相同；从 p_2 位开始有连续的 k_2 位相同；……，从 p_t 位开始连续的 k_t 位相同，且正整数 k 满足 $k=\max\{k_1,k_2,\cdots,k_t\}$，$1\leqslant i\leqslant t$，则说 X 和 Y 从 p_i 开始最大 k 位连续匹配，记作 $M(X,Y)=<p_i,k>$。同样，字符串 X 与集合 S 最大连续位匹配定义为 $M(X,S)=\max\{k\mid M(X,s)=<t,k>,\ s\in S\}$。例如，当 $X=1000011011$ 和 $Y=1100111001$ 时，第 3、4 连续两位元素相同，第 6、7、8 连续三位元素相同，则 $k=3$，$M(X,Y)=<6,3>$。这个参数给出了两个字符串中相似密度最大的位置。

1.2 刺激 – 应答变异

受 B 细胞特定基因片段连续变异的启发，文献[7]提出连续胞体超变异（contiguous somatic hypermutation，CSH）。该变异方式随机确定变异起始点和变异长度，从变异点开始变异算子依照固定的变异压力持续作用于胞体，直至变异算子遍历整个变异长度为止，其变异方式如图 1 所示。

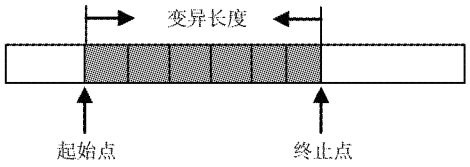

图 1 CSH 示意图

变异压力在特定连续区域上持续作用，使连续胞体超变异方式产生的新个体能更好地进行局部搜索[7,8]。但是 CSH 应该注意以下两个问题：

（1）随机选择变异个体，随机选择起始变异点（hotspot）和随机产生变异长度，在这些随机化的过程中，没有利用个体携带的部分历史信息，说明这是一种盲目的、没有导向性的变异方式。

（2）在 CSH 中，每个比特位都是以相同的、固定不变的概率 p 变异，这种固定概率的变异方式不能体现前面比特位变异对后续比特位的影响，忽略了变异片段的整体性。

基于对 CSH 算子和基因刺激运动的认识，在 CSH 操作的基础上提出刺激-应答变异（SRM）算子。该变异思想来源于基因的信号刺激-应答机制。DNA 分子的基因片段受到刺激会做出应答并单向传递刺激信号，引发后续基因产生“共鸣”，致使其功能阀门被打开或者关闭。该过程持续进行，直到这些基因小群所属的胞体达到新的、稳定的状态才会停止。在该机制中，受到刺激的基因打开或者关闭功能阀门，同时加强或削弱刺激信号，并将改变后的信号传送到后续基因使其打开或关闭功能阀门。受刺激-应答机制的启发，在二进制字符串的变异过程中，引入刺激 – 应答机制。对参加变异的个体，依据某种准则选定变异起始点和变异长度；用初始信号刺激变异起始点，使其产生应答，然后将加强或者减弱的刺激信号传递到后续比特位。后续比特位受到刺激后作出应答，并将强度发生变化的刺激信号向后传递直至变异终止点。该变异过程如图 2 所示。

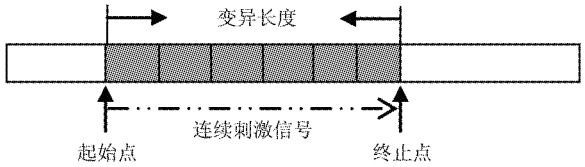

图 2 变异过程示意图

SRM 根据自体与种子检测器的内部模式结构选择变异点和变异长度，使得变异操作更具有针对性，提高了新个体的匹配成功率。变异概率采用奖励惩罚机制是 SRM 方法的又一个特点，该机制使前序比特位的变异结果影响后续比特位的变异。SRM 的主要特征是：（1）参与变异的个体不是随机选择的，而是根据个体和自体集合之间的亲和度确定的。个体与自体集合的亲和度越小，被选中的机会越大。（2）变异的起始位置和长度不是随机选择的，而是由变异个体和自体集合内部的模式结构决定的，与自体相似密度最大的个体片段变异。（3）采用奖励惩罚机制增加或削弱变异压力。当比特位发生变异时，传到后续比特位的变异压力将会减小，但不能低于压力下界；若比特位保持不变，传到后续比特位的变异压力将会增加，但不能超出最大压力上界。

2 检测器变异算法

在生物的免疫系统中，抗原表位和免疫细胞表面受体匹配的紧密程度决定了免疫细胞是否被激活以及激活的程度。从信息处理的角度来看，两个个体 r 连续位匹配意味着它们在某些位置上存在着相同的“基因片段”，生成检测器的过程就是寻找和整个自体集合没有相同“基因片段”（满足一定长度）的个体。检测器集合和自体集合没有相同的基因片段，并且两者的亲和度平均值也低于随机字符串集

合与自体集合的亲和度均值。因此通过检测器变异产生的新个体成为合格检测器的概率大。

基于检测器变异的检测器的生成算法设计思想主要体现在以下几个方面:(1)亲和度变异的基本原理是“激活-增殖”,因此应当定义一个“好”的亲和度,准确地表达两个字符串之间的相似程度,以此为依据选择变异的检测器。(2)已有的检测器变异产生新的检测器,变异方法要继承历史信息,又要和父辈检测器保持一定的距离。(3)体现进化的思想,新生成的检测器进入下一次选择过程,父辈检测器在一定次数的变异后不参与下一代的亲和度比较。(4)检测器距离自体集合比较远,因此从概率的角度来看,检测器的变异更容易产生新的检测器。算法的基本步骤如下:

(1) 采用随机法生成个种子检测器集合 D_0。

(2) 重复下面的过程,置 *Detectors* = D_0。

(3) 通过 *max _ match _ affinity*( )函数计算 *Detectors* 中个体与自体集合 *Self* 的亲和度,获得参加变异的种子个体 d、变异起始点和变异长度。

(4) 对 d 的特定片段应用刺激 - 应答变异,产生候选检测器 d'。

(5) d' *Self* 做 r 连续位匹配测试:

(*a*) d' 与 *Self* 不匹配且 $d' \notin$ *Detectors*,置 *Detectors* = *Detectors* + $\{d'\}$。

(*b*) d' 与 *Self* 匹配,且 d' 没有达到最大变异次数 *max _ mutation*,转到(4);否则转到步骤(3)。

(6) 若达到结束条件,程序停止运行。

算法说明:(a) 种子检测器 D_0 可以用目前已知的各种方法生成,比如随机法、位变异法等。D_0 集合所含元素比较少,生成这部分检测器所用的时间可以忽略不计,但是 D_0 的规模对算法的生成效率有很大的影响。(*b*) 设置最大变异次数 *max _ mutation* 可以限制算法在某个点过度搜索。(*c*) 最大位连续匹配函数 *max _ match _ affinity*( )是算法的核心,它确定了个体变异的特定片段和变异的起始点,该函数伪码如下:

```
Function max _ match _ affinity( )
1.  Input two binary strings A and B with same length
2.  affinity = 0; start _ position = 0; max _ match = 0;
3.  C = xor(A,B); //对 A、B 做二值逻辑异或操作
4.  affinity = length(A) - sum(C) // 计算 A、B 的亲和度
5.  While (pointer < length(C)) // 计算 A、B 的最大连续位匹配
6.      flag = pointer;
7.      while (C(pointer) = = 0); pointer ++ ; //遍历连续相同的基因位
8.      if(pointer-flag > max _ match){ //判断是否为最大连续位
9.          max _ match = pointer-flag; start _ position = flag;
10.         pointer ++ ;
11.     else
12.         pointer ++ ;
13.         } // if 循环结束
14. }//主循环 while.结束
```

3 实验观察与分析

本节优化 SRM 算法的重要参数、模拟算法性能,并与其它几种生成算法做了比较。实验环境采用 Intel Pentium 4,双核 CPU(1.5G 和 1.6G),1G 内存,WinXP2 操作系统,采用 Matlab2007b 运行平台。实验数据采用二进制编码。

3.1 实验过程

实验过程由参数优化和性能比较两个环节构成,前者通过种子检测器规模、变异概率的变化确定这两个参数的最优值,后者比较不同算法在生成速度和检测率两个方面的表现。实验过程如下:

(1) 随机生成均匀分布、个体长度为 16 的自体集。

(2) 在具体的实验环境中计算理论检测器规模 TN_R 和理论候选检测器规模 TN_{R0},计算公式为[1]

$$Pm = m^{-r}[(L - r)(m - 1)/m + 1] \quad (1)$$

$$TN_R = -\log P_f / P_m \quad (2)$$

$$TN_{R0} = -\frac{\log P_f}{p_m(1 - P_m)^{N_s}} \quad (3)$$

(3) 实验重复 100 次,统计相关实验数据。

3.2 参数优化实验

实验一:种子规模对候选检测规模的影响

在实验中通过穷举法生成种子检测器,由于种子检测器的规模 SN_R 很小,其生成时间对算法影响不大,但其生成过程中使用的候选检测器被统计到 SRM 算法中。设自体集合规模 $N_s = 200$,匹配阀值 $r = 10$,漏检率 $P_f = 0.01$,变异概率 $p = 0.1$,理论检测器数目 $TN_{R0} = 2579$,理论检测器规模 $TN_R = 1179$。从实验中可以看出,当 $TN_R/SN_R = 93$ 时,算法所需要的候选检测器最少。种子检测器和候选检测器之间的关系如图 3 所示。

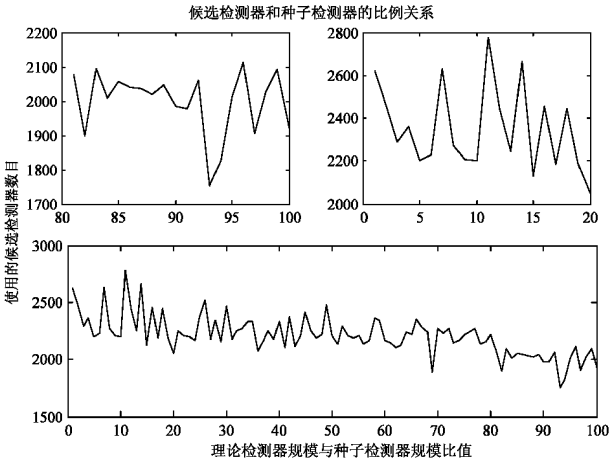

图 3 种子检测器规模对候选检测器的影响

实验二:变异概率对候选检测器规模的影响

在算法中检测器变异采用 SRM,该方法的特点是某一位的变异会引发后面多位的雪崩效应,使得特定的基因片段发生振荡。触发振荡的变异概率决定了算法的搜索能力。设自体集合规模 $N_s = 200$, $r = 10$, $P_f = 0.01$, $TN_R/SN_R = 93$,变异概率 p 以步长 0.01 遍历[0,1]区间。实验结果表明当变异概率 p 增大时,算法所使用的候选检测器总体上呈现增加的趋势。但在区间[0.10, 0.23]上,算法使用的候选检测器数目都比较少,因此取变异概率 $p = 0.2$。如图 4 所示。

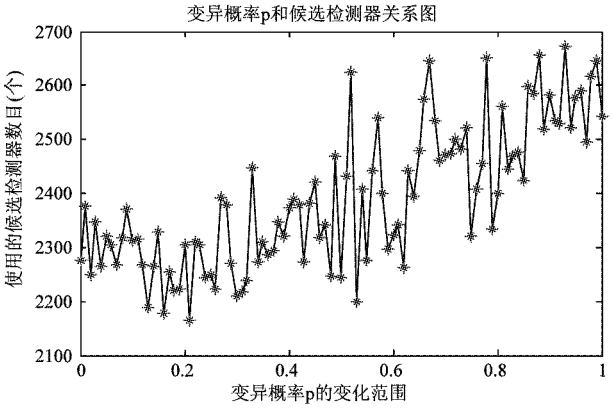

图 4 变异对候选检测器的影响

实验三:种子检测器和变异概率共同作用下的候选检测器规模

种子检测器 SN_R 和变异概率 p 作为算法的关键参数,对算法有很大的影响,本次实验测试两者共同作用下,算法使用的候选检测器规模的变化。设定 $N_s = 200$, $r = 10$, $P_f = 0.01$, p 以步长 0.05 遍历[0,1],理论检测器和种子检测器比值从 1 到 100 之间变化。实验表明当 $TN_R/SN_R \in [90, 95]$、$p \in$ [0.003,0.3]时算法使用的候选检测器规模最小,如图 5 所示。

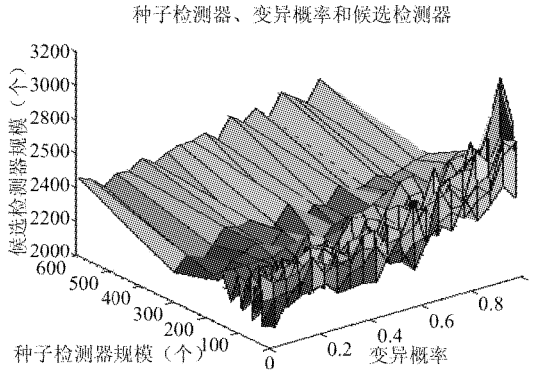

图 5 种子检测器规模和变异概率对候选检测器的影响

3.3 算法性能实验

我们采用穷举算法、个体经典变异算法、检测器经典变异算法、检测器 CSH 算法与检测器 SRM 算法作比较。本次实验环境为如下: $N_s = 200$, $L = 16$, $r = 10$, $P_f = 0.01$, $SN_R = fix(TN_R/93)$, $p = 0.2$。

3.3.1 不同算法的普适性比较

在本次实验中,让自体规模 N_s 和 r 分别变化,观察不同算法的适应能力。如图 6 所示,当自体规模从 100 递增到 500 时,各种算法使用的候选检测器都在增加,但检测器 SRM 算法使用的候选检测器普遍较少。图 7 表明了各种算法对 r 变化的适应能

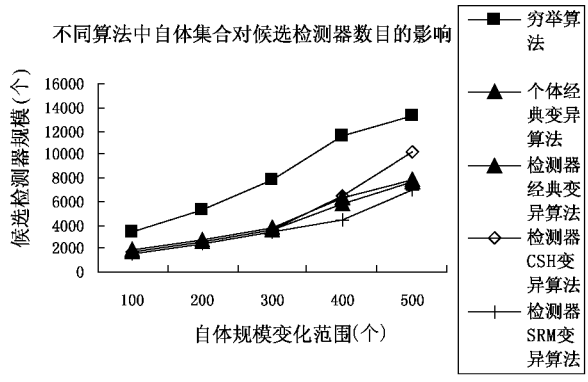

图 6 不同算法对自体集合变化的适应性比较

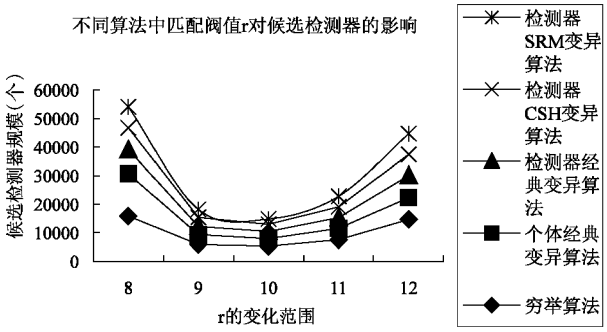

图 7 不同算法对匹配阀值变化的适应性比较

力。实验表明在 r 发生变化时,这 5 种算法都受较大的影响,但在每个确定的 r 下,检测器 SRM 算法使用的候选检测器最少。

3.3.2 算法的生成效率比较

检测器的生成效率是衡量算法优劣的重要标准之一。在生成检测器时,阴性选择算法对候选检测器的筛选过程耗费了大量的时间,因此缩小候选检测器规模是阴性选择算法的重要研究目标。本次实验比较了不同算法在相同的实验条件下,生成合格检测器所需的候选检测器数目。相关实验参数为 $N_s=200, L=16, r=10, P_f=0.01, p=0.2, SN_R=fix(TN_R/93)$,经典变异算法的变异概率为 0.1,CSH 概率设置为 0.9。实验结果表明,在相同的实验环境下,在参与比较的 5 中算法中,随机算法使用的候选检测器最多,检测器 SRM 算法需要的检测数目是 5 种算法中最少,低于理论值 2579。100 次实验均值的比较如图 8 所示。

3.3.3 算法的生成质量比较

为了计算不同算法生成检测器的质量,将 $U_L=\{0,1\}^L$ 空间中所有个体都通过检测器检测,然后统计检测率。参数设置为 $N_s=200, L=16, r=10, P_f=0.01, p=0.2, SN_R=fix(TN_R/93)$,经典变异算法的变异概率为 0.1,CSH 概率设置为 0.9。实验结果表明 5 种算法检测率的均值为 93.25%,标准方差为 0.0065,考虑到这是在随机环境下得到的数据,因此各种算法的检测率相当,如表 1 所示。

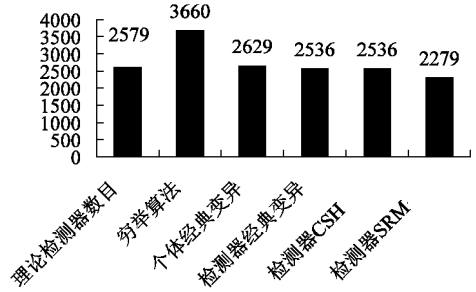

图 8 不同算法需要的检测器规模比较

表 1 相同环境下不同算法的检测率

	穷举生成算法	个体经典变异算法	检测器经典变异算法	检测器 CSH 算法	检测器 SRH 算法
检测率	93.21%	94.27%	92.46%	93.07%	93.25%

3.4 实验分析

检测器 SRM 算法利用了检测器群体与自体集合亲和度普遍偏低的特点,使种子检测器发生 SRM 产生新的候选个体,从而快速地生成检测器。在该算法中,种子检测器规模对算法的效率有很大影响,是非常关键的参数。小规模种子检测器对算法效率的影响主要在两个方面:首先是种子个体携带的历史信息太少,不能充分利用检测器和自体间的模式关系;其次是种子在问题空间中密度太低,种子覆盖的区域有限,空间搜索能力差;基于以上两个方面的检测器生成策略类似于随机个体的位变异生成算法,生成检测器的效率不高。若种子规模接近 TN_R 时,虽然种子个体携带丰富的历史信息,但是其在问题空间中的密度太大,致使变异算法效率不高。影响生成算法效率的第二个因素是“触发”SRM 的初始概率 p,该参数的大小决定了个体特定片段变异的位数和算法的局部搜索效率。在各种应用变异的算法中,变异概率值大都是实验数值,而没有更好的解决方案,实验表明当变异概率在 0.1 ~ 0.2 时,SRM 算法效率比较高。图 5 是上述两个参数共同作用下算法使用的候选检测器数目图像,可以看出,当 $SN_R=fix(TN_R/95)$, $p\in[0.1,0.2]$ 时,算法使用的候选检测器数目最少。

实验采用穷举算法、个体经典变异算法、检测器经典变异算法、检测器 CSH 算法作为检测器 SRM 算法的比较对象,实验结果表明,检测器 SRM 算法在提高生成效率的同时,保持了较高的检测率。穷举算法和个体经典变异算法由于候选个体携带的历史信息太少,致使生成效率不高。检测器经典变异和检测器 CSH 算法都利用成熟检测器作为变异对象,变异个体和自体集合之间不满足匹配规则,并且亲和度普遍较低,变异得到的新个体成为合格检测器的几率较大,但是在这两种生成方法中变异都是随机操作,没有利用个体和自体集合之间的模式关系。检测器 SRM 生成算法在选择变异个体、变异点和变异长度时,充分考虑了变异个体和自体集合的模式信息:选择与自体集合亲和度最小的检测器作为变异个体、选择个体与自体集合相似密度最大的片段确定变异区域和变异点。在变异中采用 SRM,该变异方式采用奖励惩罚机制,使比特位的变异操作影响到后续比特位的变异,这种变异操作将特定片段作为单向反馈整体,使得变异前后的片段与自体集合间的亲和度不致太高或者太低,降低了匹配成功概率,提高了算法的局部搜索效率。

4 结 论

快速生成检测器是阴性选择算法的一个核心问

题,目前很多方法都是基于随机搜索策略及其变种,这些变种大都应用了进化算法中的个体变异思想,由于变异个体和变异点的选取没有充分利用自体信息,致使算法改进不明显。本文提出了基于检测器 SRM 的检测器生成算法,它的主要特点是采用了特殊的变异个体和变异策略。算法以初始的检测器为变异对象,由于种子检测器和自体集合的亲和度普遍偏低,所以以检测器为变异对象生成合格检测器的概率较高。但是种子检测器的规模对算法的影响比较大,就像前面实验所示,当种子检测器的比值在 93 左右时,需要的候选检测器数目最少。在另外一个方面,受生物基因片段刺激应答过程的启发,文章提出了新的变异形式:SRM 在自体信息的引导下该变异发生在特定检测器的特定片段,这样有目的的变异,比盲目的随机位变异更有效率。下一步的工作主要集中在两个方面:一是探讨检测器变异这种随机过程的理论基础,对以遗传算法为代表的各种进化算法提供可供参考的理论基础;二是参考分子生物学中细胞刺激 - 应答的过程,优化算法的部分参数。

致谢:国家自然科学基金项目(项目号:60404021,60234030)、国家基础研究项目对本研究课题提供了资助,中南大学计算智能研究所的各位同事对本文提出了有价值的意见和建议,作者表示衷心的感谢。

参考文献

[1] Zhou J, Dasgupta D. Revisiting negative selection algorithms. *Evolutionary Computation*, 2007, 15(2): 223-251

[2] Forrest S, Perelson A S, Allen L, et al. Self-nonself discrimination in a computer. In: Proceedings of IEEE Symposium on Research in Security and Privacy. Oakland: IEEE Press, 1994.202-212

[3] D'haeseleer P. An immunological approach to change detection: theoretical results. In: Proceedings of the 9th IEEE Computer Security Foundation Workshop. Washington D C: IEEE Computer Society Press, 1996. 132-143

[4] D'haeseleer P, Forrest S, Helman P. An immunological approach to change detection: algorithm, analysis and implications. In: Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1996. 110-119

[5] 郭振河,谭营,刘政凯.基于阴性选择原则的 Non-self 探测器生成算法. 小型微型计算机, 2005, 26(6): 959-964

[6] 何申, 罗文坚, 王煦法.一种检测器长度可变的非选择算法. 软件学报, 2007, 18(6): 1361-1368

[7] Kelsey J, Timmis J. Immune inspired somatic contiguous hypermutation for function optimization. In: Proceedings of Genetic and Evolutionary Computation Conference. Berlin Heidelberg: Springer-Verlag, 2003. 207-218

[8] Timmis J, Edmonds C, Kelsey J. Assessing the performance of two immune inspired algorithms and a hybrid genetic algorithm for function optimisation. In: Proceedings of the Congress on Evolutionary Computation, Potland, Oregon, USA, 2004. 1044-1051

Study on the simulated-response mutation algorithm based on seed detectors

Liu Xingbao[*,**], Cai Zixing[*]

([*] School of Information Science and Engineering, Central South University, Changsha 410083)

([**] Center of Modern Education Technology, Hunan Business College, Changsha 410025)

Abstract

A new detector generation strategy, based on seed individuals and contiguous somatic simulating mutation, was proposed to reduce the time complexity of the negative selection algorithm (NSA). The strategy produces seed detectors and determines the special detectors and gene segment by measuring the affinity between the seed set and the self set, and then a stimulated-response mutation (SRM) occurs in a special gene fragment and the candidate individuals emerge, and finally selects the new competent detectors according to the r-contiguous bit matching rule. The characteristic of the algorithm is that it uses the pattern information to guide the mutation process for reducing the matching rate of candidate individuals. The experimental results show that the algorithm outperforms several similar algorithms based on mutation operator in term of time complexity and coverage.

Key words: artificial immune system, negative selection algorithm, r-contiguous bit matching rule, affinity measure, stimulated-response mutation (SRM)