文章编号: 1001-0920 (2013) 08-1130-08

基于拟随机序列与克隆选择的进化 V-detector 算法

金章赞[a], 廖明宏[b]

(厦门大学 a. 信息科学与技术学院，b. 软件学院，福建 厦门 361005)

摘 要: 阴性选择 (NS) 算法是人工免疫的核心方法, 检测器生成是其关键. 针对其经典 V-detector 算法中高维数据失效及随机生成初始检测器集过于集中而导致过早收敛等问题, 首先采用拟随机序列生成初始检测器; 然后通过克隆选择优化检测器集合, 以覆盖非自体空间大小及数量作为亲和力标准, 克服传统进化阴性选择 (ENS) 算法的局限性, 并采用新型进化算子使得算法生成最优检测器集合; 最后, 通过实验验证了该方法的有效性.

关键词: 进化阴性选择算法；拟随机系列；克隆选择；检测器生成

中图分类号: TP301 **文献标志码:** A

Evolutionary V-detector algorithms based on clone selection and quasi random sequence

JIN Zhang-zan[a], *LIAO Ming-hong*[b]

(a. School of Information Science and Technology；b. School of Software，Xiamen University，Xiamen 361005, China．Correspondent：JIN Zhang-zan，E-mail：sacula1010@163.com)

Abstract: Negative selection(NS) algorithm is the core algorithm of artificial immune system, in which the detector generate mechanism is the key. But the performance of V-detector algorithm becomes unfavorable on high-dimension data and the set of initial detectors randomly generated are too concentrated leading to the algorithm convergence prematurely. Quasi random sequence is used to generate the set of initial detectors. Then the detector set is optimized by using clone selection, and the coverage of non-self-space and the number of detectors are used as the standard of affinity which can over come the limitations of ENSA. A new selection, cloning and mutation operator is used to generate the optimal mature detector set. Finally, experiments verify the effectiveness of the proposed algorithm.

Key words: evolutionary negative selection algorithms；quasi random sequence；clone selection；detector generation

0 引 言

生物免疫系统是一个高度复杂、自组织、自适应的并行分布式系统, 能够区分自体与非自体, 抵御外界病菌的入侵和感染, 维持机体自身生理活动的稳定与平衡. 受生物免疫系统启发, 研究人员将生物免疫系统相关优秀特性应用于解决各类实际问题, 并由此形成了人工免疫系统 (AIS) 这一新学科. 近年来, 由于其强大的信息处理能力, 人工免疫系统得到了长足发展, 成为人工智能中继神经网络、进化算法之后的又一个研究热点. 目前, 人工免疫方法主要包括: 阴性选择算法 (NS)、克隆选择算法和免疫网络模型[1]. 其中作为核心的阴性选择算法由于其独有特性已发展成为人工免疫学的主要方法, 对整个系统具有重要意义, 因而如何生成少量、高效的成熟检测器便成为阴性选择算法的关键环节.

本文对传统进化阴性选择算法加以改进, 提出了基于拟随机序列和克隆选择的进化 V-detector 算法. 该算法采用拟随机序列来生成初始检测器集合, 避免了高维数据失效以及随机生成初始检测器集合过于集中而导致算法过早收敛等问题. 然后通过克隆选择来优化检测器集合, 以其覆盖非自体空间大小及检测器数量作为亲和力计算标准, 克服了传统 ENS 算法的局限性, 并采用新型进化算子来生成最优检测器集合. 大量实验结果验证了本文方法的有效性.

1 阴性选择算法简介

1.1 传统 NS 算法

自 Forrest 等[2]提出阴性选择算法以来, 由于其无

收稿日期: 2012-03-19；**修回日期:** 2012-07-15.

基金项目: 中央高校基本科研业务费项目(2010121070)；福建省自然科学基金项目(2010J01342).

作者简介: 金章赞(1984−), 男, 博士生, 从事人工免疫系统等研究；廖明宏(1966−), 男, 教授, 博士生导师, 从事网络智能、普适计算等研究.

需先验知识, 仅利用有限数量的自体便能检测出无限数量的非自体等优点, 使 NS 算法得到了迅速发展. Gonzalez 等[3]针对二进制 NS 算法所存在的问题, 提出了实值阴性选择算法 (RNS). 该方法采用实值表述, 不但接近原始问题空间, 而且使用计算几何的相关特性来加速算法. 然而, 由于其需要预先设定检测器数量且半径固定, 限制了算法的扩展性. Zhou 等[4]在此基础上, 提出了 V-detector 算法. 该方法采用半径可变的检测器, 使得大半径检测器可以覆盖大部分非自体, 减少了检测器数量, 不但使存储空间大为降低, 而且时间开销也随之减少; 而小半径检测器能够覆盖"漏洞", 进一步减少了"漏洞"数量. 但其存在以下问题: 若生成初始检测器过于集中, 则易导致算法过早收敛[5]、处理高维数据的低效性[6]、"边界困境"[7]和检测器集合为非最优集等问题.

1.2　进化阴性选择算法

针对 V-detector 算法存在的问题, 国内外学者对其进行了大量研究, 认为其问题主要是控制检测器产生机制, 该算法受随机搜索的限制, 不能保证完整地覆盖非自体空间, 并产生大量重叠. 而基于进化搜索的检测器生成机制, 借助于其优秀的搜索特性, 能够完备地覆盖非自体空间, 并减少检测器之间的重叠. 因而人们将生物进化机制与阴性选择机制相结合, 提出了进化阴性选择算法 (ENS), 将进化思想运用于检测器生成过程, 通过进化机制生成更加优秀的检测器. 如 Dasgupta[8-9]、Joseph [10]、Marek [11]、Gao[12]、Jorge [13]、Zhang[14]等分别利用遗传算子来优化检测器生成机制. 但遗传算法只考虑了全局搜索, 在一定情况下还会出现过早收敛、退化等现象. Gao 等[15]针对 NS 算法不适于动态环境的问题, 提出了基于神经网络的 NS 算法. Wang 等[16]采用粒子群优化检测器生成, 使得检测器集合能够尽可能地覆盖非自体. Liu 等[17]采用多目标优化进化检测器集合, 将检测器半径和重叠率作为亲和力标准, 但该算法易陷入局部最优而导致过早收敛.

本文将克隆选择算法用于检测器生成机制, 通过克隆选择操作使得算法快速收敛, 利用变异操作使得算法保持一定的多样性, 抑制了早熟现象, 并对其选择、克隆、变异操作加以改进, 提出了基于拟随机序列和克隆选择的进化 V-detector 算法.

2　克隆选择

2.1　克隆选择原理

克隆选择原理是免疫学基本理论之一, 它解释了抗体的形成机理并阐明了免疫应答的多样性机制. 1973 年, Jerne[18]提出了克隆选择原理. 该原理的主要内容是: 当外部细菌、病毒等抗原侵入机体后, B 细胞将对这些抗原进行识别, 识别后的 B 细胞将克隆扩增分化为浆细胞, 最终产生一种蛋白质分子即抗体细胞. 在细胞克隆过程中, 抗体细胞还经历了一个变异的过程, 其结果是产生对抗原具有导向性的抗体. 克隆选择的主要特征是: 克隆选择对应着一个亲和力成熟的过程, 即对抗原亲和力较低的抗体在克隆选择机制的作用下, 经历克隆和变异操作后, 其亲和力逐步提高而"成熟"的过程.

2.2　克隆选择算法

Castro 等[19]基于克隆选择原理提出了克隆选择算法, 这是一种模拟免疫系统学习过程的进化算法. 抗原被一些与之匹配的 B 细胞识别, 这些 B 细胞大量分裂, 产生新的 B 细胞并在原有 B 细胞的基础上发生变异, 以寻求与抗原更好匹配的 B 细胞, 那些与抗原匹配更好的 B 细胞将再次分裂. 如此循环往复, 最终找到与抗原完全匹配的 B 细胞, 这些 B 细胞最终一部分变成浆细胞产生抗体, 另一部分变成记忆性 B 细胞, 形成免疫记忆, 这一过程即为克隆选择过程. 克隆选择算法模拟了这一过程的优化, 其计算步骤如下:

1) 生成候选方案的一个集合 P, 它是记忆细胞 M 的子集与剩余群体 P_r 之和, 即 $P = P_r + M$.

2) 选择 n 个具有较高亲和力的抗体.

3) 克隆这 n 个抗体, 组成一个临时的克隆群体 C. 与抗原亲和力越高, 抗体在克隆时的规模越大.

4) 把克隆群体提交到高频变异, 根据亲和力的大小决定变异, 产生一个成熟的抗体群体 C^*.

5) 对 C^* 进行再选择, 组成记忆细胞集合 M. P 中的一些成员可以被 C^* 中的一些改进的成员替代.

6) 生成 d 个新的抗体并取代 P 中 d 个低亲和力的抗体以保持多样性.

克隆选择算法继承了生物免疫系统的众多属性, 并具有自组织、自学习、自识别、自记忆的能力, 因此它不仅避免了优秀抗体的丢失, 而且能够得到全局最优解, 是一种新的全局优化搜索算法. 其在算法实现上兼顾了全局和局部搜索, 通过克隆选择操作可使算法快速收敛, 通过变异操作使算法保持了一定的多样性, 抑制了早熟现象. 故本文对传统基于进化的 NS 算法进行改进, 将克隆选择引入检测器 (即抗体) 生成过程, 通过克隆选择生成最优成熟检测器集合.

3　基于拟随机序列与克隆选择的进化 V-detector 算法

3.1　算法步骤

1) 本文采用 Sobol 拟随机序列生成 n 个初始检测器集合 $D = \{D_1, D_2, \cdots, D_n\}$, 并通过计算与自体

集之间的距离来确定各个检测器集合中检测器d的半径r_d. 若距离小于自体半径r_s, 则丢弃; 否则半径r_d为其到与其最近的自体之间的最短距离, 即

$$r_d = \mathrm{Dis}(d, \mathrm{nearest\_}s) - r_{\mathrm{nearest\_}s}.$$

2) 计算各个初始检测器集合D_i的亲和力$f(D_i)$, $i = 1, 2, \cdots, n$, 并按亲和力的大小对各个检测器集合从低到高排列. 亲和力大小的标准是检测器集合所覆盖非自体空间的大小及其个数, 覆盖非自体空间越大且数量越少的检测器集合, 其亲和力值越大. 如果具有最高亲和力的检测器集合满足收敛条件, 则算法退出; 否则继续.

3) 选择m个高亲和力检测器集合.

4) 对这m个检测器集合进行克隆, 形成一个克隆检测器集合群P, 亲和力越高的检测器集合, 其克隆次数也相应越多.

5) 对检测器集合群P进行有导向性的变异, 即亲和力越高其变异程度越低, 亲和力越低其变异程度越高. 生成一个成熟的检测器集合群体P^*, 计算检测器集合群体P^*的亲和力并排序.

6) 生成k个新检测器集合并取代P^*中的k个低亲和力检测器集合以保持抗体的多样性. 检测P^*中的各个检测器是否与自体重叠, 若重叠, 则重新计算其半径. 如此反复循环, 直到生成满足收敛条件的成熟检测器集合.

上述算法的流程如图1所示.

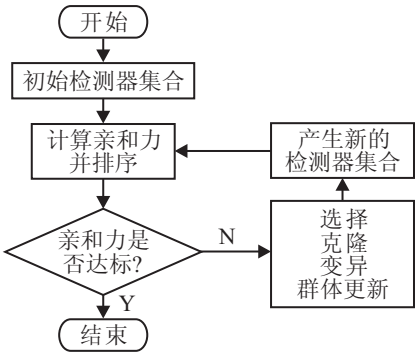

图 1 算法流程

3.2 初 始 化

传统V-detector算法采用伪随机数来生成初始检测器集合. 其特点是占用内存少、产生速度快、便于重复产生且不受限制. 但该方法存在高维不均匀性、长周期相关性等缺点, 若生成的初始检测器过于集中, 则容易导致算法过早收敛. 拟随机序列, 也称低差异序列, 一般产生于单位n维超立方体内并均匀分布. 在许多情况下, 相比伪随机数, 拟随机序列在多维空间中的超均匀分布性能更好地覆盖领域空间, 拟随机序列可以产生具有低差异度的个体, 它以牺牲随机性为代价, 换取均匀性的提高, 两者的分布性如图2、图3所示. 对于相同数量和大小的检测器集合, 采用拟随机序列生成的检测器集合, 其覆盖体积要大于伪随机序列生成的检测器集合[13], 故采用拟随机序列生成的检测器集合将有效地克服传统V-detector算法中高维数据失效以及随机生成初始检测器集合过于集中而导致算法过早收敛等问题.

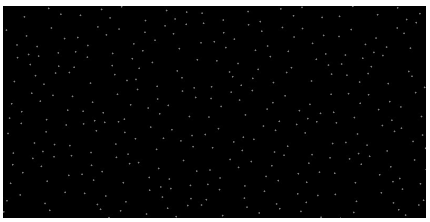

图 2 二维拟随机序列

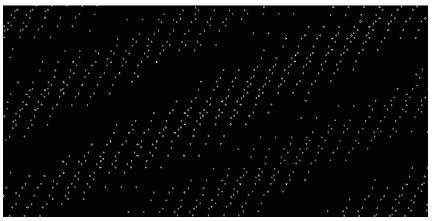

图 3 二维伪随机序列

目前已提出的拟随机序列主要有Faure、Sobol、Halton序列. 其中Faure和Halton序列在高维呈现强相关性, 且生成时间长于Sobol序列, 因此本文采用Sobol序列来生成初始检测器集合. Sobol是半随机序列的一种, 可以扩展到任意维度, 在随机采样中可以达到高质量的采样效果, 相比其他拟随机序列, 其采样更均匀, 是一个稳定的高覆盖率的随机序列[20]. Sobol序列是基于一组称为“直接数”的数v_i而构造的[21]. 设m_i是小于2^i的正奇数, 则有

$$v_i = m_i/2^i. \tag{1}$$

数v_i(同时m_i)的生成借助于系数只为0或1的简单多项式. 多项式可表示成

$$f(z) = z^p + c_1 z^{p-1} + \cdots + c_{p-1}z + c_p. \tag{2}$$

对于$i > p$, 有如下递归公式:

$$v_i = c_1 v_{i-1} \oplus c_2 v_{i-2} \oplus \cdots \oplus c_p v_{i-p} \oplus \lfloor v_{i-p}/2^p \rfloor, \tag{3}$$

其中$\oplus$表示二进制按位异或. 对于m_i, 对等的递归公式为

$$m_i = 2c_1 m_{i-1} \oplus 2^2 c_2 m_{i-2} \oplus \cdots \oplus 2^p c_p m_{i-p} \oplus m_{i-p}. \tag{4}$$

3.3 计算亲和力

免疫系统中, 亲和力值最大的抗体即为“成熟”抗体. NS算法的关键是生成高效的成熟检测器集合, 而优秀的检测器集合是那些能以较少数量尽可能地覆盖更大非自体区域的检测器集合, 且生成漏洞最少.

Dasgupta[10]、Gao[12]、Liu [17]、陶新民[22]、Zhang[14]等利用检测器半径大小来评价检测器优劣, 然而该标准只保留大半径检测器, 抛弃小半径检测器, 从而导致检测器无法覆盖小漏洞. Ostaszewski[11]、Wang[16]和 Liu 等[17]将优秀抗体定义为覆盖非自体区域广且与其他检测器重叠少的检测器, 虽然考虑了检测器之间重叠的问题, 但小漏洞问题仍未得到有效解决.

本文不同于传统基于单个检测器覆盖率的亲和力计算标准, 而是将整个检测器集合的空间覆盖率及其数量作为衡量亲和力大小的标准, 使得检测器集合既有大半径检测器, 又包含小半径检测器. 通过大半径检测器既能覆盖更多的非自体空间, 又能减少检测器数量; 而小半径检测器可以覆盖那些大半径检测器无法覆盖的微小漏洞, 从而降低了漏洞数量. 其亲和力公式可以表示为

$$\min\{|N(D_i)| : p(D_i) \geqslant p_0\}. \tag{5}$$

其中: $N(D_i)$ 为第 i 个检测器集合中检测器的个数, $p(D_i)$ 为其非自体空间覆盖率, p_0 为收敛标准.

检测器集合的空间覆盖率无法直接求解, 可以用 Monte Carlo 方法计算求得. 在非我空间内随机采集 n 个样本点, 采用文献 [23] 中假设检验的方法来判断检测器集合是否达到了非我空间覆盖率. 假设检验覆盖率估计方法来源于概率论中的假设检验理论和中心极限定理. 首先设假设检验理论中的原假设 H_0 为: 继续加入新的检测器, 增大覆盖率; 备择假设 H_1 为: 覆盖率已满足, 不再继续生成新的检测器. 因为只有检测器生成这一过程满足对称分布, 中心极限定理才能正确应用, 所以必须保证假设条件 $np \geqslant 5$ 和 $nq = n(1-p) \geqslant 5$ 成立. 此时必须保证样本大小, 即候选检测器生成数目满足 $n > \max(5/p, 5/(1-p))$. 其中: p 为预先设定的成熟检测器集合的非自体空间覆盖率, x 为被检测器集合覆盖的样本数量, z_a 为显著性水平, 可由正态分布表查得. 于是由中心极限定理可得

$$z = \frac{x - np}{\sqrt{np(1-p)}}. \tag{6}$$

若 $z \geqslant z_a$, 则算法将拒绝原假设 H_0, 说明检测器集合的覆盖率已经满足需要; 反之, 将接受原假设 H_0 的概率为 $P(x < z_\alpha) = 1 - \alpha$. 因此亲和力计算公式转变为

$$\min\{|N(D_i)| : z \geqslant z_a\}. \tag{7}$$

3.4　检测器选择

为了产生优秀检测器集合, 规范变异方向, 使检测器集合向着亲和力值更高的方向变异, 需对检测器集合进行选择, 亲和力值越高, 选择概率越大. 但是, 当群体中某种个体的数量占据了相当规模, 而此检测器集合又不是最优检测器集合时, 依据上述规则易导致抗体过早收敛. 为此, 应对某些达到一定规模的检测器集合进行抑制, 同时增加小规模检测器集合的产生以提高多样性. 文献 [24] 在传统基于亲和力大小选择机制的基础上, 增加了小生境浓度调节因子.

小生境原理是指: 如果某个物种有较多的个体, 则该物种的个体应以较大幅度降低, 鼓励较少个体的繁衍. 通过抑制高浓度检测器来增加低浓度检测器的被选择概率, 从而保持了个体的多样性. 小生境浓度调节因子是基于检测器集合的浓度进行的, 检测器集合浓度越大, 则被选中的概率越小. 检测器集合 D_i 的浓度 C_i 定义为

$$C_i = N(i)/N(D). \tag{8}$$

其中: λ 为亲和力阈值, $N(i)$ 为与检测器集合 i 亲和力值大于 λ 的检测器集合个数, $N(D)$ 为检测器集合总数. 计算出每个检测器集合的浓度之后, 便可通过选择机制进行检测器集合的促进和抑制调节. 于是检测器集合 D_i 的选择概率 S_i 由亲和力概率 p_z 和浓度概率 p_c 两部分组成, 即

$$S_i = p_z \cdot p_c = \left(z_i \Big/ \sum_{i=0}^{N} z_i\right) \cdot \frac{1}{\alpha} \mathrm{e}^{-\beta \cdot C_i}. \tag{9}$$

其中: α 和 β 为常数, 文中均取为 1; z_i 为检测器集合 i 的亲和力值.

3.5　检测器克隆

为避免优秀检测器集合丢失, 同时增加优秀检测器集合数量, 提高全局收敛效率, 需对检测器集合进行适当的克隆. 亲和力值越高的检测器集合, 其克隆的数量也越多. 这里采用简单的复制操作来表示克隆, 即对经过选择操作的 n 个按亲和力从小到大排序的检测器集合 $D = [D_1, D_2, \cdots, D_n]$ 进行复制, 每个检测器集合有 m 个检测器 d, 设 $D_j = [d_{j1}, d_{j2}, \cdots, d_{jm}]$ 是其中第 j 个检测器集合, 可将其复制为

$$\begin{aligned}\mathrm{Clone}(D_j) = I_j D_j = \\ \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{(n+1-j)*1} \times [d_{j1}, d_{j2}, \cdots, d_{jm}]_{1*m} = D_{cj} = \\ \begin{bmatrix} d_{j1}, d_{j2}, \cdots, d_{jm} \\ d_{j1}, d_{j2}, \cdots, d_{jm} \\ \vdots \\ d_{j1}, d_{j2}, \cdots, d_{jm} \end{bmatrix}_{(n+1-j)*m}, \ j = 1, 2, \cdots, J.\end{aligned} \tag{10}$$

其中 I_j 为 $n+1-j$ 维单位列向量, 使得检测器集合复制数目正比于其亲和力.

3.6　检测器变异

为了获得检测器集合的多样性并使检测器集合

快速地成熟, 需对检测器进行适当的变异. 检测器的变异不仅有利于快速搜索, 而且能使检测器集合跳出局部最优, 得到全局最优. 为了让检测器集合朝着亲和力值高的方向变异, 应使亲和力值高的检测器集合的变异小一些, 甚至抑制其变异; 而亲和力低的检测器集合, 其相应的变异应大一些. 对于单个检测器, 若其与其他检测器之间的重叠率越大, 则其变异越剧烈.

这里先计算集合中每个检测器的重叠率, 若重叠率大于该检测器集合的阈值, 则进行变异, 其中重叠率越高、亲和力越低的检测器变异越大. 此外, 各个检测器集合的阈值各不相同, 其大小正比于检测器集合的亲和力值, 这意味着亲和力越低的检测器集合其阈值越小, 而相应的变异越大.

例如, 对第 j 个检测器集合 D_j 中的第 i 个检测器 $d_i^j=(d_1^i,d_2^i,\cdots,d_n^i)$ 进行变异, 该检测器与该集合中的检测器 $d_a^j=(d_1^a,d_2^a,\cdots,d_n^a)$, $d_b^j=(d_1^b,d_2^b,\cdots,d_n^b)$, $d_c^j=(d_1^c,d_2^c,\cdots,d_n^c)$ 重叠. 若其重叠率 $W(d_i^j)$ 大于阈值 th_j, 则其变异后的检测器 $d_{i'}^j=(d_1^{i'},d_2^{i'},\cdots,d_n^{i'})$. 其中

$$\begin{aligned} d_1^{i'}&=d_1^i+t_1\alpha_i^j,\\ d_2^{i'}&=d_2^i+t_2\alpha_i^j,\\ &\vdots\\ d_n^{i'}&=d_n^i+t_n\alpha_i^j. \end{aligned}\tag{11}$$

t 为检测器移动方向, 其计算公式如下:

$$\begin{aligned} &\text{if }\ [(d_1^i-d_1^a)+(d_1^i-d_1^b)+(d_1^i-d_1^c)]>0,\\ &\text{then }\ t_1=1;\ \text{else}\ t_1=-1.\\ &\text{if }\ [(d_2^i-d_2^a)+(d_2^i-d_2^b)+(d_2^i-d_2^c)]>0,\\ &\text{then }\ t_2=1;\ \text{else}\ t_2=-1.\\ &\quad\vdots\\ &\text{if }\ [(d_n^i-d_n^a)+(d_n^i-d_n^b)+(d_n^i-d_n^c)]>0,\\ &\text{then }\ t_n=1;\ \text{else}\ t_n=-1. \end{aligned}\tag{12}$$

$$\alpha_i^j=kW(d_i^j)/z_j.\tag{13}$$

其中: α_i^j 为变异因子, k 为变异步长, z_j 为亲和力. 重叠率越高、亲和力越小, 变异因子将越大, 其重叠率计算公式如下:

$$\text{Overlap}(d,d')=\begin{cases} 0,\ \|r_d-r_{d'}\|\geqslant r_d+r_{d'};\\ \left(\exp\left(\dfrac{r_d+r_{d'}-\|r_d-r_{d'}\|}{r_d+r_{d'}}\right)-1\right)^m,\\ \quad \|r_d-r_{d'}\|<r_d+r_{d'}. \end{cases}\tag{14}$$

$$\text{Overlap}(d)=\sum_{i\neq j}\text{Overlap}(d,d').\tag{15}$$

其中: m 为特征空间维数, $||r_d-r_{d'}||$ 为检测器之间距离. 其检测器阈值计算公式为

$$th_j=\beta\cdot z,\tag{16}$$

其中 β 为比例系数, 检测器阈值与其亲和力的大小成正比.

3.7 检测器更新

对检测器集合进行更新, 保持检测器集合的多样性, 防止检测器集合过早收敛. 该操作将检测器集合中亲和力值较小的部分检测器集合用随机产生的等量新检测器集合取代.

4 实验结果与分析

为验证本文方法较传统方法的改进之处, 分别采用本文方法、V-detector 算法[4]和 GA-NSA[12]对二维平面五角星数据和鸢尾花数据进行异常检测.

实验中, 检测率 D 和误检率 F 分别定义为

$$D=\text{TP}/(\text{TP}+\text{FN}),\ F=\text{FP}/(\text{TN}+\text{FP}).\tag{17}$$

其中: TP、FN、FP、TN 分别代表事件“非自体样本被判定为非自体”、“非自体样本被判定为自体”、“自体样本被判定为非自体”、“自体样本被判定为自体”发生的次数.

4.1 二维平面数据

为了直观地比较 3 种方法的性能, 采用平面五角星图像作为自体. 图 4 ~ 图 7 为平面五角星图像, 图中样本空间为 $[0,1]^2$, 图中的五角星图案表示自体样本, 黑色圆圈表示检测器. 图 4 为以五角星为自体的空间分布图, 图 5 为 V-detector 算法效果图, 图 6 为 GA-NSA 算法效果图, 图 7 为本文方法效果图.

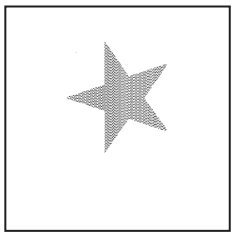

图 4 五角星为自体空间

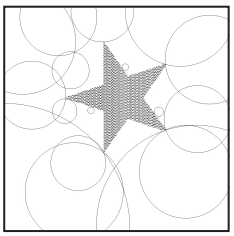

图 5 V-detector 算法

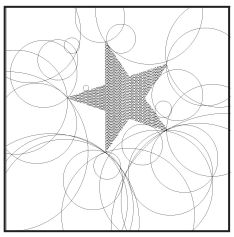

图 6 GA-NSA 算法

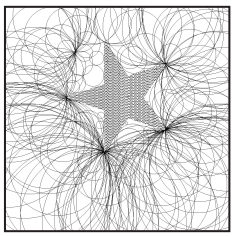

图 7 本文方法

GA-NSA 算法中各参数设置如下: 检测器数量 $N=10$, $r=0.1$, 迭代次数为 100 次. V-detector 算法中各参数取值为: $T_{\max}=200$, maximum_self_ coverage $=0.999\,9$, $r_s=0.1$, $c_0=0.99$. 本文方法参数设置如

下: 自体半径 $r_s = 0.1$, 检测器覆盖率 $p = 0.99$, 自体个数 $S_n = 7\,000$, 检测器个数 $T_{\max} = 200$, 显著性水平 $z_a = 2.245\,08$, maximum_self_coverage = 0.999 9.

从图 4～图 7 中可知, 本文方法对非自体空间的覆盖范围较其他两种算法有了极大的提高. 另外, 虽然 V-detector 算法的最大检测器数量设置为 200, 但其最终生成的成熟检测器数量却很少, 这主要是因为随机生成初始检测器集合时过于集中而导致算法过早收敛; 同样, GA-NAS 方法采用遗传操作使得检测器分布有了一定的优化, 但由于其只考虑了全局搜索, 易出现过早收敛现象. 而本文方法采用拟随机序列, 具有均匀分布性, 避免了过早收敛, 而且由于采用克隆选择优化检测器集合, 使得检测器集合能尽可能地覆盖非自体区域.

图 8 是 3 种方法关于检测率与自体半径的参数关系图, 图 9 是他们自体半径与误检率的参数关系图. 从图中可知, 检测率和误检率都随着自体半径的增大而减少. 而本文方法在自体半径一致的前提下其检测率优于其他方法.

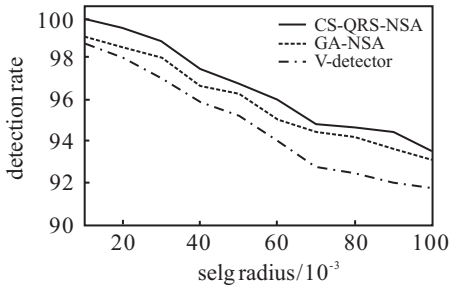

图 8　检测率与自体半径关系图

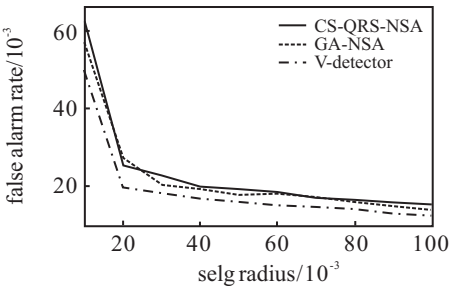

图 9　误检率与自体半径关系图

4.2　鸢尾花 (Iris) 数据

Iris 数据是常用的异常检测、数据挖掘和机器学习数据集, 该数据包含 150 种鸢尾花的信息, 每 50 种取自 3 个鸢尾花种, 分别为 Setosa、Versicolor 和 Virginica. 每种花有 4 种特征: 萼片长度(sepal length)、萼片宽度 (sepal width)、花瓣长度 (petal length)、花瓣宽度 (petal width). 图 10、图 11 分别是鸢尾花萼片长度-宽度、鸢尾花花瓣长度-宽度的平面分布图.

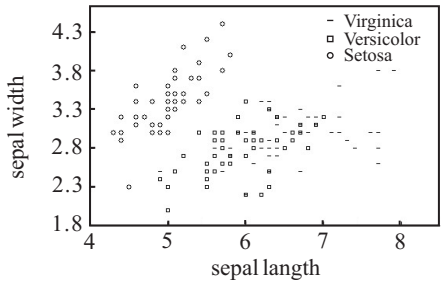

图 10　鸢尾花萼片长度-宽度的分布图

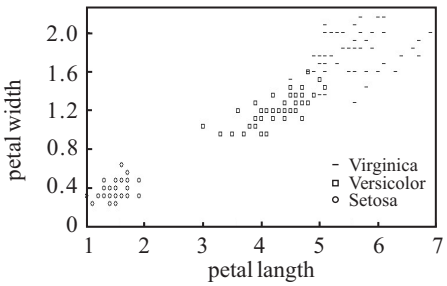

图 11　鸢尾花花瓣长度-宽度的分布图

对该数据集进行检测时, 将其中一个花种数据作为自我集, 另外两种作为非我集进行检测, 且利用自我集进行训练时采用 100%自我集以及 50%自我集分别进行训练. 检测结果如表 1 所示 (对各种数据取 100 次独立实验的平均值). GA-NSA 算法中各参数设置为: $r = 0.1$, 迭代次数为 100 次. V-detector 算法中各参数取值为: $T_{\max} = 1\,000$, $r_s = 0.1$, $c_0 = 0.99$, maximum_self_coverage = 0.999 9. 本文方法参数设置如下: 自体半径 $r_s = 0.1$, 检测器覆盖率 $p = 0.99$, 自体个数 $S_n = 7\,000$, 检测器个数 $T_{\max} = 200$, 显著性水平 $z_a = 2.245\,08$, maximum_self_coverage = 0.999 9.

表 1　3 种算法实验结果对比图

训练数据	算法类型	检测率 $D/\%$	误检率 $F/\%$	检测器个数
Setosa 100%	V-detector	99.98	0	20
	GA-NSA	99.99	0	16
	本文算法	99.99	0	15
Setosa 50%	V-detector	99.97	1.32	16
	GA-NSA	99.99	1.84	19
	本文算法	100.00	2.01	20
Versicolor 100%	V-detector	85.95	0	153
	GA-NSA	87.01	0	134
	本文算法	92.13	0	118
Versicolor 50%	V-detector	88.3	8.42	110
	GA-NSA	90.19	9.40	106
	本文算法	93.10	10.21	103
Virginica 100%	V-detector	81.87	0	218
	GA-NSA	84.97	0	170
	本文算法	90.43	0	120
Virginica 50%	V-detector	93.58	13.18	108
	GA-NSA	94.71	16.87	81
	本文算法	97.86	15.21	63

从实验结果可知, 本文方法具有较好的检测率且误检率也保持在较低水平. 这主要是因为本文方法采用了基于拟随机序列和克隆选择相结合的成熟检测器生成机制, 较好地抑制了算法过早收敛, 产生出质量更高的检测器集合. 此外, 由于本文以检测器数量为亲和力计算标准之一, 使得检测器个数进一步减少.

4.3 KDD99 数据

为了验证在高维数据上的有效性, 分别将 3 种算法对高维 KDD99 数据[25]进行测试. 首先, 把 KDD99 提供的训练库格式化成 $[0,1]^{41}$ 空间的数据; 其次, 从格式化完的库里随机选取 3 个子集 S_1, S_2, S_3; 再次, 用数据集 S_1 的正常记录训练产生检测器集; 最后分别把检测器集置于 S_2、S_3 上进行测试. GA-NSA 算法中各参数设置如下: $r = 0.1$, 迭代次数为 100 次. V-detector 算法中各参数取值为: $T_{\max} = 1\,000$, $r_s = 0.1$, $c_0 = 0.99$, maximum_ self_coverage $= 0.999\,9$. 本文方法参数设置如下: 自体半径 $r_s = 0.1$, 检测器覆盖率 $p = 0.99$,自体个数 $S_n = 1\,000$, 检测器个数 $T_{\max} = 1\,000$. 实验结果取各个算法运行 50 次的平均值, 具体见表 2.

表 2 算法实验效果 (用 S_1 训练, 用 S_2、S_3 测试)

数据集	算法类型	检测率 $D/\%$	误检率 $F/\%$
S_2	V-detector	5.54	0.23
	GA-NSA	34.98	0.98
	本文算法	99.13	3.12
S_3	V-detector	6.31	0.12
	GA-NSA	37.25	1.02
	本文算法	99.67	2.87

从表 2 可见, 本文方法在保持误检率低的情况下实现了对高维数据的有效检测.

5 结 论

本文将克隆选择原理应用于阴性选择算法, 将检测器集合的覆盖率及其数量作为亲和力标准, 利用克隆选择极佳的寻优性能, 生成了最优的成熟检测器集合, 探索出一种新的阴性选择方法. 另外, 采用拟随机序列生成初始检测器, 避免了算法的过早收敛. 实验结果表明, 该方法能有效地对异常数据进行快速准确的检测, 具有很高的检测率, 是一种高效可靠的异常检测方法.

参考文献(References)

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

[2] Forrest S, Perelson A S, Allen L, et al. Self-nonself discrimination in a computer[C]. Proc of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos: IEEE, 1994: 221-231.

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

[4] Zhou Ji, Dasgupta D. Real-valued negative selection algorithm with variable-sized detectors[C]. Proc of GECCO. Washington: Springer, 2004: 287-298.

[5] 金章赞, 肖刚, 陈久军. 基于视觉感知与 V-detector 的水质异常检测方法[J]. 信息与控制, 2011, 40(1): 130-136. (Jin Z Z, Xiao G, Chen J J. Anomaly detection of water quality based on visual perception and V-detector[J]. Information and Control, 2011, 40(1): 130-136.)

[6] Stibor T, Timmis J, Eckert C. A comparative study of real-valued negative selection to statistical anomaly detection techniques[C]. Proc of the 4th Int Conf on Artificial Immune Systems. Berlin: Springer, 2005: 262-275.

[7] Zhou Ji, Dasgupta D. Augmented negative selection algorithm with variable-size detectors[C]. IEEE Congress of Evolutionary Computation. Washington: IEEE Press, 2004, 1: 1081-1088.

[8] Dasgupta D, Gonzalez F. An immunity-based technique to characterize intrusions in computer networks[J]. IEEE Trans on Evolutionary Computation, 2002, 6(3): 1081-1088.

[9] Joseph M Shapiro, Gary B. An evolutionary algorithm to generate hyper-ellipsoid detectors for negative selection[C]. GECCO2005. Washington: ACM, 2005: 337-344.

[10] Dasgupta D, Krishna K, Kumar D Wong, et al. Negative selection algorithm for aircraft fault detection[C]. Proc of the 3rd Int Conf on Artificial Immune Systems. Catania: Springer, 2004: 1-13.

[11] Marek Ostaszewski, Marek Ostaszewski, Pascal Bouvry. Immune anomaly detection enhanced with evolutionary paradigms[C]. Proc of the 8th Annual Conf on Genetic and Evolutionary Computation. Seattle, 2006: 119-126.

[12] Gao X Z, Ovaska S J, Wang X. Genetic algorithms-based detector generation in negative selection algorithm[C]. Proc of the IEEE Mountain Workshop on Adaptive and Learning Systems. Logan, 2006: 133-137.

[13] Jorge L, Amaral M, Jose F A. Real-valued negative selection algorithm with a quasi-Monte Carlo genetic detector generation[C]. Proc of the 6th Int Conf on Artificial Immune System. Barcelona, 2007: 156-167.

[14] Zhang Jie, Luo Wenjian, Baoliang X. Generating an approximately optimal detector set by evolving random seeds[C]. Proc of the 8th Int Conf on Dependable, Autonomic and Secure Computing. Chengdu, 2009: 162-168.

[15] Gao X Z, Ovaska S J, Wang X. A neural networks-based negative selection algorithm in fault diagnosis[J]. Neural Computer & Application, 2008, 17(1): 91-98.

[16] Wang H M, Gao X Z, Huang X L. PSO-optimized negative selection algorithm for anomaly detection[J]. Applications of Soft Computing, 2009, 52: 13-21.

[17] Liu F, Gong M G, Ma J J. Optimizing detector distribution in V-detector negative selection using a constrained multi-objective immune algorithm[C]. Proc of the 2010 IEEE Congress on Evolutionary Computation. Barcelona, 2010: 18-23.

[18] Jerne N K. The immune system[J]. Scientific American, 1973, 229(1): 51-60.

[19] De Castro L N, Von Zuben F J. Clonal selection algorithm with engineering applications[C]. Proc of Genetic and Evolutionary Computation Conference. Las Vegas, 2000: 36-37.

[20] 潘金京, 李宗民. 一种改进的基于Sobol序列的快速线积分卷积法[J]. 中国石油大学学报, 2011, 3(31): 162-166.
(Pan J J, Li Z M. An improved algorithm of fast line integral convolution with Sobol sequence[J]. J o f China University of Petroleum, 2011, 3(31): 162-166.)

[21] Sobol I M. The distribution of points in a cube and the approximate evaluation of integrals[J]. Computational Mathematics and Mathematical Physics, 1967, 7(4): 86-112.

[22] 陶新民, 杜宝祥, 徐勇. 基于高阶统计特征实值阴性克隆选择算法的轴承故障检测[J]. 机械工程学报, 2008, 44(7): 230-236.
(Tao X M, Du B X, Xu Y. Bearing fault detection using real-valued negative clone selection algorithm based on higher order statistics[J]. Chinese J of Mechanical Engineering, 2008, 44(7): 230-236.)

[23] Zhou Ji, Dasgupta D. Estimating the detector coverage in a negative selection algorithm[C]. Proc of the 2005 Conf on Genetic and Evolutionary Computation. Washington: ACM, 2005, (1): 281-288.

[24] 罗文坚, 曹先彬, 王煦法. 用一种免疫遗传算法求解频率分配问题[J]. 电子学报, 2003, 31(6): 915-917.
(Luo W J, Cao X B, Wang X F. Solving frequency assignment using an immune genetic algorithm[J]. Chinese J of Electronics, 2003, 31(6): 915-917.)

[25] Stibor T, Timmis J, Eckert C. A comparative study of realvalued negative selection to statistical anomaly detection techniques[C]. Proc of the 4th Int Conf on Artificial Immune Systems. Berlin: Springer, 2005: 262-275.

(上接第1129页)

[63] Sbarbaro D, Murray-Smith R. An adaptive nonparametric controller for a class of nonminimum phase non-linear system[C]. Proc of the 16th IFAC World Congress. Praga, 2005: 453-458.

[64] Rottmann A, Burgard W. Adaptive autonomous control using online value iteration with Gaussian processes[C]. Proc of IEEE Int Conf on Robotics and Automation. Kobe, 2009: 2106-2111.

[65] Ferris B, Haehnel D, Fox D. Gaussian processes for signal strength-based location estimation[C]. Proc of the Int Conf on Robotics, Science and Systems. Philadelphia, 2006: 303-310.

[66] Ko J, Klein D J, Fox D, et al. Gaussian processes and reinforcement learning for identification and control of an autonomous blimp[C]. Proc of the Int Conf on Robotics and Automation. Rome, 2007: 742-747.

[67] Ko J, Fox D, Haehnel D. GP-UKF: Unscented Kalman filters with Gaussian process prediction and observation models[C]. Proc of the Int Conf on Intelligent Robots and Systems. San Diego, 2007: 1901-1907.

[68] Deisenroth M P, Huber M F, Hanebeck U D. Analytic moment-based Gaussian process filtering[C]. Proc of the 26th Int Conf on Machine Learning. Montreal, 2009: 81-94.

[69] Ko J, Fox D. GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models[J]. Automomous Robots, 2009, 27(1): 75-90.

[70] 李鹏, 宋申民, 陈兴林. 自适应平方根无迹卡尔曼滤波算法[J]. 控制理论与应用, 2010, 27(2): 143-146.
(Li P, Song S M, Chen X L. Adaptive square-root unscented Kalman filter algorithm[J]. Control Theory & Applications, 2010, 27(2): 143-146.)

[71] 李鹏, 宋申民, 陈兴林, 等. 联合高斯回归的平方根UKF方法[J]. 系统工程与电子技术, 2010, 32(6): 1281-1285.
(Li P, Song S M, Chen X L, et al. Square root unscented Kalman filter incorporating Gaussian process regression[J]. Systems Engineering and Electronics, 2010, 32(6): 1281-1285.)

[72] Kocijan J. Dynamic GP models: An overview and recent developments[C]. Proc of the 6th Int Conf on Applied Mathematics, Simulation, Modelling. Vouliagmeni Beach, 2012: 38-43.