软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(10):3104−3121 [doi: 10.13328/j.cnki.jos.006017]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

邻域形态空间多源免疫检测器生成与检测*

席 亮, 姚之钰, 张凤斌

(哈尔滨理工大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

通讯作者: 席亮, E-mail: xiliang@hrbust.edu.cn

摘 要: 人工免疫系统(artificial immune system,简称 AIS)是人工智能技术的重要分支之一,被广泛应用于异常检测、数据挖掘、机器学习等多个领域.检测器是其核心知识集,其生成、优化和检测操作决定了人工免疫的应用效果.目前,人工免疫的问题空间以实值形态空间为主,但实值非自体空间"黑洞"、检测器生成速率慢、检测器高重叠冗余、"维度灾难"等问题,使得人工免疫检测的效果不甚理想.鉴于此,使用邻域形态空间,并改进邻域否定选择算法(neighborhood negative selection algorithm,简称 NNSA),引入混沌理论和遗传算法,提出了一种多源邻域否定选择算法(multi-source-inspired NNSA,简称 MSNNSA),并基于此提出邻域形态空间多源免疫检测器生成与检测方法,改进邻域形态空间下检测器的构造与生成机制,使其更具靶向性,并使获得的检测器具有更好的分布性,提高其生成效率和整体的检测性能,解决以上实值形态空间下存在的问题.实验结果表明,该方法提高了检测器生成效率以及检测的整体性能和稳定性.

关键词: 邻域形态空间;异常检测;否定选择;混沌映射;遗传算法

中图法分类号: TP18

中文引用格式: 席亮,姚之钰,张凤斌.邻域形态空间多源免疫检测器生成与检测.软件学报,2021,32(10):3104−3121. http://www.jos.org.cn/1000-9825/6017.htm

英文引用格式: Xi L, Yao ZY, Zhang FB. Multi-source-inspired immune detector generation and detection in neighborhood shape-space. Ruan Jian Xue Bao/Journal of Software, 2021,32(10):3104−3121 (in Chinese). http://www.jos.org.cn/1000-9825/6017.htm

Multi-source-inspired Immune Detector Generation and Detection in Neighborhood Shape-space

XI Liang, YAO Zhi-Yu, ZHANG Feng-Bin

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

Abstract: Artificial immune system (AIS) is one of the important branches of artificial intelligence technology, and it is widely used in many fields such as anomaly detection, data mining, and machine learning. The detectors are its core knowledge set, and the application effects are determined by the generation, optimization, and detection of the detectors. At present, the problem space of AIS mainly applied real-valued shape-space. But the detectors in the real-valued shape-space have some problems that have not been solved, such as the holes in the non-self-shape-space, slow speed of generation, detector overlapping redundancy, dimension curse, which lead to the unsatisfactory detection effects. In view of this, based on the neighborhood shape-space, a new shape-space, and the improved neighborhood negative selection algorithm, a multi-source-inspired neighborhood negative selection algorithm (MSNNSA) is proposed by introducing chaotic map and genetic algorithm. And then, based on this algorithm, the multi-source-inspired immune detector generation and detection methods in neighborhood shape-space are built to make the construction and generation more targeted, so that the generated detectors have better distribution performance. Meanwhile, the method also improves the detectors' generation efficiency and the detection

* 基金项目: 国家自然科学基金(61172168); 黑龙江省自然科学基金(F2018019)

Foundation item: National Natural Science Foundation of China (61172168), Natural Science Foundation of Heilongjiang Province, China (F2018019)

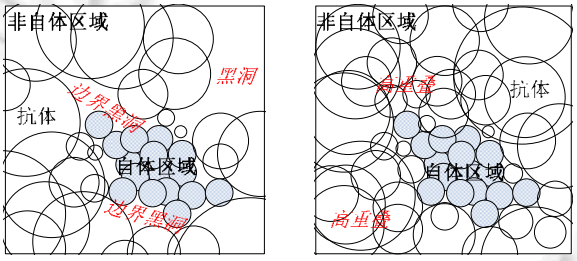
收稿时间: 2019-05-06; 修改时间: 2019-08-22, 2019-12-19; 采用时间: 2020-01-31

performances, and overcomes the shortcomings in the real-valued shape-space mentioned before. Experimental results show that the proposed method enhances generation efficiency, whole detection performances, and stability.

Key words: neighborhood shape-space; anomaly detection; negative selection algorithm; chaotic map; genetic algorithm

作为人工智能的一个重要分支,AIS 是一种模拟生物免疫功能的智能技术[1],现已被广泛应用于异常检测、数据挖掘、医学影像等各个领域[2,3].生物免疫系统与异常检测的基本作用机理十分相似,人工免疫方法经常以异常检测或基于异常的入侵检测为应用对象展开研究[4].而且,基于免疫的异常检测技术因其具有良好的预知未知异常的能力和较强的鲁棒性等特点而成为本领域的研究热点[5].如:文献[6]使用 AIS 构建适合监控环境的分布式传感器网络来防范 DoS;文献[7]改进了树状细胞算法,提出了一种免疫异常检测模型,并以此为核心建立实时入侵检测模型;文献[8]提出了基于免疫启发合作 agent 的安全防御系统,它使用增强型否定选择算法,使系统更好地覆盖自体或非自体(异常),以提高系统的整体异常检测性能.

目前,AIS 应用主要基于实值形态空间展开,匹配策略主要根据样本间的 Euclidean 距离和 Manhattan 距离来度量,其“维度灾难”问题严重影响了属性选取范围和计算效率;而且,实值检测器的高重叠和“黑洞”问题也是影响人工免疫算法应用效果的关键问题,如图 1 所示[9].因此,许多学者就这些问题展开了大量研究工作[10].文献[11]利用抗原空间密度计算抗原密集聚集的低维子空间,并在这些子空间中直接生成检测器来解决高维空间中存在的问题,并通过抑制被其他成熟检测器识别的候选检测器,采用“抗体抑制率”代替“预期覆盖率”作为终止条件来解决检测器冗余的问题.文献[12]基于克隆选择提出一种实值否定选择改进算法 CB-RNSA(clonal-selection-based RNSA),首先生成大半径候选检测器,覆盖远离自体空间以减少检测器数量;然后生成小半径候选检测器,并使它们逐渐覆盖自体和非自体之间的边界区域以便减少黑洞,从而有效提高检测器的检测率,并降低误报率.然而,实值形态空间的“维度灾难”及样本重叠造成的“黑洞”和规模过大等问题并没有从根本上得到解决,因而使得人工免疫算法在属性选取、计算效率、优化和检测等方面还需进一步展开研究[9].



(a) 黑洞 (b) 高重叠

Fig.1 Problems in real-valued shape-space

图 1 实值形态空间存在的问题

鉴于此,针对上述问题,本文使用由我们研究团队首先提出的形态空间表示法——邻域形态空间表示法[13],提出邻域形态空间多源免疫检测器生成与检测方法:通过改进邻域否定选择算法(neighborhood negative selection algorithm,简称 NNSA)解决“维度灾难”问题,并采用随机、混沌映射和 DNA 遗传变异多源机制构造候选检测器进行耐受训练,实时更新检测器集,使其具有对非自体区域的完备覆盖;并基于邻域的集合特性设计亲和力计算方法,解决实值形态空间的样本高重叠和“黑洞”问题;最终设计基于此的人工免疫检测模型,以提高检测器集的整体检测性能为目标,提升检测器的构造、生成(收敛)与检测效率[9].

本文第 1 节介绍人工免疫系统和相关的免疫异常检测基本模型,并分析混沌映射与遗传算法的特点.第 2 节详述本文提出的方法,并进行算法收敛性和时间复杂度分析.第 3 节是实验结果与分析.最后进行结论性总结和展望.

1 相关工作

1.1 人工免疫与异常检测模型

AIS模拟生物免疫系统学习外界物质的自然防御机理,提供了一种强大的信息处理和问题求解范式[14].AIS的问题域(形态空间)目前主要包括两种:二进制和实值表示法,包含了许多核心模型和算法,如图2所示.

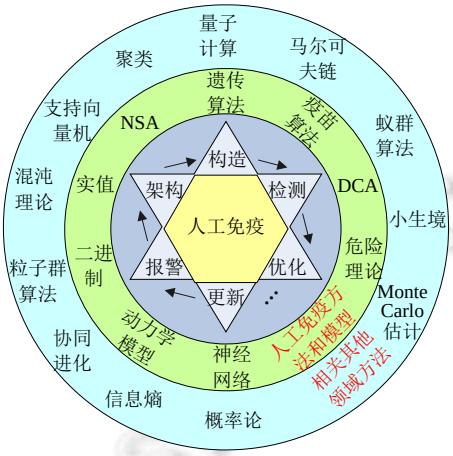


Fig.2 Algorithms and models in artificial immune system

图2 人工免疫相关算法及模型

异常检测是一种基于行为的检测,通过建立正常行为的模式轮廓,将当前活动与之相比较:若违反规则 3,则被视为异常[15].异常检测的方法很多,其中,人工免疫和异常检测的作用机制极其相似,基于免疫机制的异常检测是本领域非常重要的研究模型[3].免疫异常检测是模拟生物免疫系统各种相关细胞和抗体耐受与构造过程的否定选择和遗传变异等机制,实时进化更新以保持对抗原的识别,其核心知识集是检测器集,由候选检测器或抗体等机制经过耐受训练获得[9].检测器的规模和完备性问题是异常检测的核心问题[16],基本模型如图3所示.

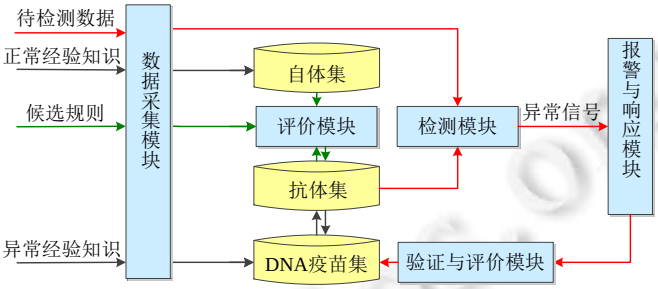


Fig.3 Basic model of immune-inspired intrusion detection

图3 免疫异常检测基本模型

1.2 *n*维混沌映射

混沌机制由于其所具有的随机遍历性等特征,已逐渐被应用在仿生智能算法的优化机制中[17],并在不同的应用背景下取得了良好效果[18].文献[19]将混沌引入到基于 Hash 的级联驱动系统中以提高其可控性.文献[20]利用混沌映射实现多服务器网络框架下的认证密钥协议并取得了很好的实验效果.张楠等人将混沌机制引入入侵检测,采用 Logistic 映射以解决个体冗余的问题[21].文献[22]基于混沌提出一种新的图像加密算法,提升其破解难度.在与人工免疫算法的联合应用中,文献[23]利用遗传算法并引入混沌算子来保持解决方案的多样性,以解决急诊室过度拥挤长期逗留的问题,利用混沌的随机遍历性,有效避免了遗传算法的过早收敛.文献[24]利

用混沌算子提高了樽海鞘群算法(salp swarm algorithm,简称 SSA)的收敛速度.

不同映射的混沌序列发生器产生不同的混沌序列,其概率分布是不同的,会影响到算法的效率.当前,Logistic 映射和自映射是两种主要的方式.Logistic 映射进行有限次折叠,得到的空间结构相对简单,存在着较多的安全问题;自映射产生的混沌序列具有更好的随机性和初值敏感性,算法收敛速度快,空间分布更好[9].Lyapunov 指数常用于衡量系统的混沌性质[25],而且由于基于自映射的 Lyapunov 指数高于基于 Logistic 的 Lyapunov 指数,因此自映射比 Logistic 映射具有更好的问题空间覆盖性能,本文将采用该方法产生混沌序列.而且,文献[26]将混沌映射应用于实值形态空间下的否定选择算法,并且证明了自映射在 n 维空间具有良好的遍历性和收敛性,从而能够提高检测器的分布能力.

1.3 遗传算法(genetic algorithm,简称GA)

GA 是一种人工免疫自适应进化方法,具有较强的鲁棒性,经过不断的迭代,可以较大范围地覆盖问题域,其基本流程如图 4 所示[9].但它不能保证搜寻到全局最优,容易陷入局部最优.目前,GA 在各个领域都有所应用:文献[27]利用 GA 求解公交网络设计问题;文献[28]引入 GA 来解决特征选择问题,通过 GA 提高了算法的执行效率;文献[29]指出,利用 GA 优化解决了串联-并联系统的冗余分配问题;文献[30]利用 GA 实现图聚类匿名以保护隐私,利用 GA 的全局化搜索优化能力保障图聚类质量;文献[31]使用 GA 进行文本的感情分析,使适应度逐渐提高,最终找到最大化准确的词典.GA 与其他传统的搜索算法相异之处在于:GA 从随机生成的一组初始解开始进行解的搜索,通过交叉、变异、选择等操作实现后续的不断迭代,根据适应度函数衡量,最终收敛于最优个体[32].

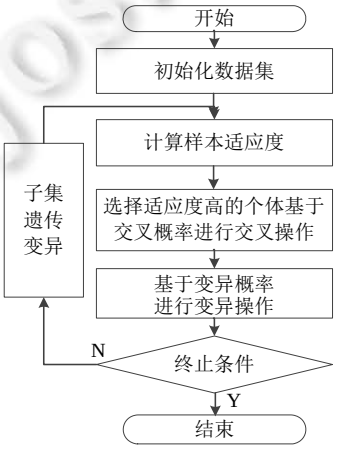


Fig.4 Flow chart of the genetic algorithm

图 4 遗传算法流程

2 多源邻域检测器生成与检测

2.1 邻域形态空间及否定选择算法

邻域形态空间的数据基础是集合论,基于离散拓扑理论进行建模,可有效避免实值形态空间的“维度灾难”等问题,其数学模型可定义为集合$[0,1]^n$的离散拓扑空间的一个子集,在集合的第 j 维$[0,1]^j$上,将其平均分为 m_j 个子集,这 m_j 个子集彼此互不相交[9],即:

$$(x_0,x_1)\cup(x_1,x_2)\cup\ldots\cup(x_{m_{j-1}},x_{m_j})^j,$$

其中,x_0=0, $x_{m_j}=1$. 扩展到 n 维空间,形态空间是所有邻域块集合的集族:

$$\bigcup(x_{i-1},x_i)^n,\ 1\leqslant i\leqslant m_j,\ j=1,2,\ldots,n.$$

邻域自体或检测器可表示为

$$\boldsymbol{s}=[s_1,s_2,\ldots,s_n]^T \tag{1}$$

其中,$s_i(i=1,2,\ldots,n)$为该样本在第 i 维上的划分编码值,$s_i=(s_{i(j-1)},s_{ij}),s_{i(j-1)}$和 s_{ij} 分别为上下界,且满足 $j\in[0,m_i]$, $s_{i0}=0$, $s_{im_i}=1, s_{i0}<s_{i(j-1)}<s_{ij}<s_{im_i}$, m_i 为第 i 维划分的邻域数[9].

免疫检测方法主要基于 NSA,其大多模拟生物免疫中的相关细胞耐受及成熟过程来区分自体与非自体元素,其核心是亲和力计算与匹配规则设计[9].在邻域形态空间下,NSA 的亲和力计算如下.

(1) 邻域样本间计算

设 $\boldsymbol{X}[\boldsymbol{s}_x,\boldsymbol{M},otherAttributes]^T$ 和 $\boldsymbol{Y}[\boldsymbol{s}_y,\boldsymbol{M},otherAttributes]^T$ 为待匹配邻域样本,$\boldsymbol{s}$ 为原始邻域样本:$\boldsymbol{s}_x[x_1,x_2,\ldots,x_n]^T$ 和 $\boldsymbol{s}_y[y_1,y_2,\ldots,y_n]^T$,$\boldsymbol{M}=[m,\sigma_1,\sigma_2,\ldots,\sigma_n]^T$ 为匹配阈值向量,m 为总匹配度,$otherAttributes$ 为活跃度、生命期、匹配阈值等其他必要属性,λ_i 为该类别权重,其邻域匹配度 m 为

$$m=\sum_{i=1}^{n}\lambda_i\sigma_i \tag{2}$$

其中,

$$\sigma_i=\sum_{i=1}^{n}\mu_i(x_i\oplus y_i) \tag{3}$$

计算方法是:属性匹配以改进实值 Manhattan 距离(Minkowski 距离特例)为重合度匹配进行,类别匹配以邻域加权二进制 Hamming 匹配统计进行,其中,μ_i 为属性权重[9].

(2) 邻域与待检测样本(实值表示)间计算

设 $\boldsymbol{X}[\boldsymbol{s}_x,\boldsymbol{M},otherAttributes]^T$ 和 $\boldsymbol{Y}[\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n]^T$ 为待匹配样本,m 和 σ_i 同上,其中,

$$x_i\oplus y_i=\begin{cases}0, & y_i\notin(x_{i(j-1)},x_{ij})\\ 1, & y_i\in(x_{i(j-1)},x_{ij})\end{cases} \tag{4}$$

该算法相对于传统的实值否定选择算法有以下优势.

(1) 邻域亲和力计算方法令每维属性独立进行匹配,然后综合判定,削弱数据属性间的耦合关系,其时间复杂度不会随维数增加而变大,可解决“维数灾难”问题,从根本上突破了维度限制,解放了属性维度限制,不用进行降维即可高效计算;

(2) 基于此的检测方法突出单维属性的独立作用,可保证更精确的检测判断,加速问题求解速度;

(3) 邻域形态空间以集合论为数学基础,基于离散拓扑理论进行建模,将实值样本归为集合形式表示,在极大地缩小样本规模的同时,也有效解决了实值样本的高重叠问题,并基于邻域样本与实值样本的计算方法,可高效地进行异常检测,从而使人工免疫更好地应对更加复杂的问题域[9].

2.2 多源邻域免疫检测器生成算法

本文算法以基于多源邻域否定选择算法(multi-sources NNSA,简称 MSNNSA)为核心来实现候选检测器的构造、耐受和择优,从而获得成熟的检测器集.其中,$otherAttributes$ 主要包括生命期 $weight$(初始值设为 0,并结合基于匹配计数的度量值进行动态调整)和构造方式 $construction$(包括随机、混沌映射、遗传变异这 3 种方式,采用轮盘赌方式进行选择),算法直到非自体空间的覆盖率达到期望值或者检测器数量达到最大值时结束,然后用生成的检测器进行异常检测[9],其基本流程如图 5 所示.

算法需对数据集进行标准化处理:

$$x_i'=\frac{x_i-\min(i)}{\max(i)-\min(i)} \tag{12}$$

其中,$\max(i)$和 $\min(i)$分别表示该属性的最大值与最小值.

具体而言,算法的伪代码实现如下.

输入:邻域划分步长 $step$;

成熟检测器生成最大数量 N_d;

成熟检测器数量 n;

亲和力匹配阈值 ρ;

非自体空间覆盖率 p;
轮盘赌选择概率 $p_r[3]$;
非自体空间期望覆盖率 P_{cov};
自体集合 $\boldsymbol{Self}$;
候选检测器 $\boldsymbol{d}_c$;
候选检测器集合 $\boldsymbol{D}_c$;
候选检测器个数 N_c;
变异概率 P_v;
交叉概率 P_c;
记录适应度值大小的数组 $\boldsymbol{F}[\cdot]$;

输出:成熟检测器集合 $\boldsymbol{D}$.

1. **Begin**
2. $\boldsymbol{D}=\varnothing;\boldsymbol{F}[\cdot]=\varnothing;n=0;$ /*按照公式(5)对测试数据集进行标准化等预处理,如图 5 中步骤 1*/
3. *Neighborhood*($\boldsymbol{Self}$,*step*); /*根据 *step* 值划分邻域形态空间,将实值表示的多个自体样本转换为邻域形式的单个自体样本,并精确样本中每维属性边界,确定样本在每维属性下的步长,如图 5 中步骤 2*/
4. **while** ($p<P_{cov}||n<N_d$) { /*如果覆盖率没有达到 p 或生成检测器数量小于 N_d,继续循环;否则算法结束,如图 5 中步骤 6*/
5. *flag*=*Rand*(·); /*根据 3 种构造方式的成熟检测器的累计概率设置权重,采用加权轮盘赌选择的方式选择使用随机、混沌映射或遗传变异的方式来构造候选检测器,如图 5 中步骤 3*/
6. **if** (*flag*≤$p_r[0]$),$\boldsymbol{D}_c$=*RandGenerate*(·); /*随机方式利用系统随机函数构造候选检测器*/
7. **elseif** (*flag*>$p_r[0]$ && *flag*≤($p_r[0]+p_r[1]$)),$\boldsymbol{D}_c$=*ChaosGenerate*(·); /*混沌映射方式构造候选检测器*/
8. **else** { /*遗传变异方式按照图 4 的流程构造候选检测器*/
9. Calculate *fitness*→$\boldsymbol{F}[\cdot]$; /*适应度值计算*/
10. Selection and *Crossover*(P_c)→$\boldsymbol{D}_c$; /*按照 P_c 进行交叉操作*/
11. *Mutation*($\boldsymbol{D}_c$,P_v); /*按照 P_v 对 $\boldsymbol{D}_c$ 进行变异操作*/
12. }
13. **for** (i=0;$i<N_c$;i++){
14. t=*Affinity*($\boldsymbol{Self}$,$\boldsymbol{d}_{ci}$); /*通过 NNSA 对候选检测器进行亲和力计算与匹配,如图 5 中步骤 4*/
15. **if** ($t<\rho$){$\boldsymbol{d}_{ci}$→$\boldsymbol{D}$;$n=n+1$;} /*亲和力小于ρ,将该候选样本认定为成熟检测器,如图 5 中步骤 5*/
16. **else** Discard $\boldsymbol{d}_{ci}$; /*丢弃*/
17. }
18. *MonteCarlo*(p); /*使用 Monte Carlo 方法计算检测器的覆盖率*/
19. }
20. **End**

本文算法通过邻域划分将属性值相近的样本映射到同一空间,构造出具有集合特性的邻域样本,可解决样本高重叠问题,从而极大地缩小了样本规模.同时,采用公式(2)~公式(4)的邻域亲和力匹配规则进行耐受训练判定,也可在降低计算复杂度的同时,有效地避免“维度灾难”问题.而且在构造候选检测器时,利用随机、遗传变异、混沌映射这 3 种来源构造候选检测器,也可最大化基于检测器的问题空间.因此,本文算法从多个角度显著提升了检测器生成与检测的计算效率.

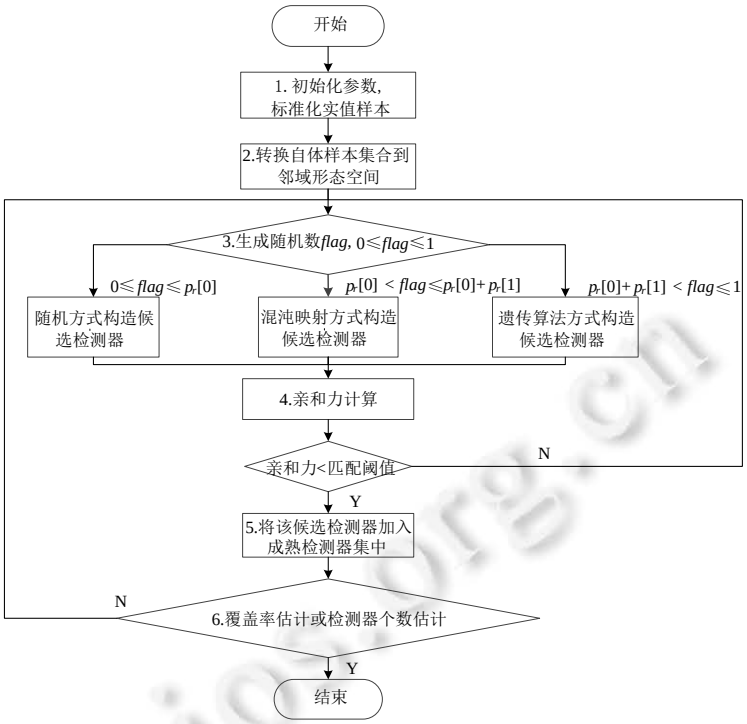


Fig.5 Flow chart of multi-source-inspired immune detector generation algorithm

图 5 多源邻域免疫检测器生成算法流程

(1) 随机方式的优点是实现简单、速度快,但分布性不可控、不均匀分布的候选样本对算法的寻优速度影响较大,使算法容易陷入局部最优区域,检测器生成的效率低,算法收敛速度慢,且不可避免地存在“黑洞”问题;

(2) 遗传变异方式提取有用 DNA,克服随机因素,使检测器的构造和生成具有靶向性,能够加速算法收敛,但算法只能在当前搜索到的子空间内维持较好的分布性能,而且通过优秀 DNA 构造的检测器检测容易造成“马太效应”,使新兴个体难以进入搜索视野范围,导致种群缺乏多样性,不能保证找到全局最优解;

(3) 混沌映射方式由于其本身具有的遍历性,可以使数据源均匀分布在问题空间,当算法陷入因遗传变异方式造成的局部最优时,通过混沌映射构造的解可以跳出局部最优,从而提高检测器的分布能力.然而,在高维形态空间下,由于属性较多、搜索空间较大,很难较快地搜索到最优解空间,使算法的搜索效率降低,收敛速度下降;

(4) 在构造候选检测器来源的选择上,算法通过加权轮盘赌的方式来进行,每种构造候选检测器方法在轮盘赌中的权重随着基于该方式检测器的检测性能而发生改变:生成的检测器检测性能越好的构造来源在轮盘中的累计概率越大,权重越大,则这种来源在构造候选检测器时被选用的概率也越大[9];

(5) 设算法初始解为$\boldsymbol{X}=(x_1,x_2,...,x_n)$,同时,设当前最优解为$\boldsymbol{X}_{Best}=\boldsymbol{X}$.设算法总迭代次数为$m$,其中,采用随机、混沌映射和遗传变异方式的迭代次数分别为 m_1、m_2 和 m_3,则 $m=m_1+m_2+m_3$;设当前迭代次数为 k,算法基于加权轮盘赌的方式使算法在上述 3 种方式中不断切换来优化空间.同时,伴随检测实时调整权重:性能越好的方式,迭代次数就越多,从而使算法自适应调整种群进化状态;若生成的解 $\boldsymbol{X}_c$ 优于当前最优解,则 $\boldsymbol{X}_{Best}=\boldsymbol{X}_c$,直至 $k=m$,则停止搜索[9];

(6) 因此,3 种候选检测器构造方式优势互补:遗传变异基于 DNA 疫苗的有用知识,可靶向性解决随机和

混沌映射方式搜索效率低、收敛速度慢的问题;混沌映射的均匀覆盖特性,可缓解遗传变异的“马太效应”;随机方式的速度优势,可从总体上加快算法的结束时间;3 种方式不断地自适应调整,最终较为全面地覆盖问题解空间,并使算法收敛于全局最优[9].

2.3 邻域检测器检测算法

本文算法基于 NNSA,参数设定、规范化处理与生成算法相同,如图 6 所示.其关键是检测的亲和力计算匹配,采用邻域的检测器与实值的待检测样本间的计算,基于公式(2)~公式(4)进行[9].

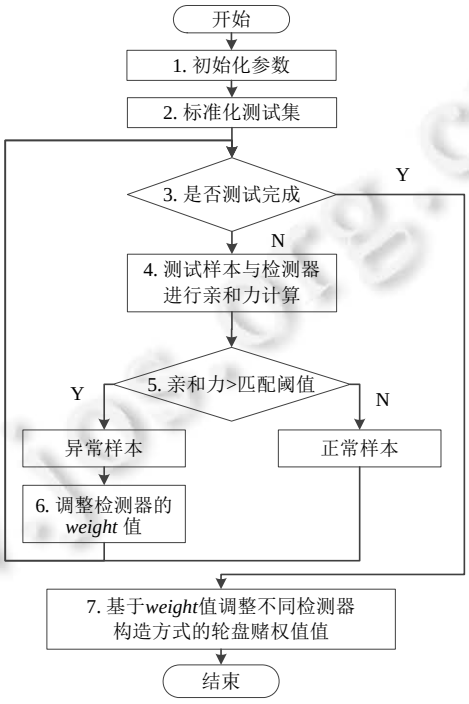


Fig.6 Flow chart of MSNNSA detection algorithm

图 6 MSNNSA 检测算法基本流程

具体而言,检测算法的伪代码实现如下.

输入:亲和力匹配阈值ρ;

待检测样本集合 ***Detect***;

成熟检测器集合 ***D***;

输出:轮盘赌选择概率 p_r[3];

检测结果.

1. **Begin** /*设定算法亲和力匹配阈值ρ等参数,如图 6 中步骤 1*/
2. *Normalize*(***Detect***) /*按照公式(5)对测试数据集进行标准化处理,如图 6 中步骤 2*/
3. **for** (*i*=0;*i*<***Detect***.size;*i*++){ /*依次检测每个待检测样本,如图 6 中步骤 3*/
4. *t*=*Affinity*(***D***,***Detect***$_i$); /*通过公式(2)~公式(4)进行待检测样本与成熟检测器之间的亲和力计算,如图 6 中步骤 4*/
5. **if** (*t*<ρ){***Detect***$_i$→*normal*;} /*亲和力小于ρ,则该检测样本正常,如图 6 中步骤 5*/
6. **else** {***Detect***$_i$→*abnormal*;***D***$_i$.*weight*++;} /*亲和力大于ρ,则该检测样本异常,增加相应检测器 *weight* 属性值,如图 6 中步骤 6*/

7. }
8. **for** (*i*=0;*i*<***D***.*num*;*i*++){ /*根据每种构造方式生成的检测器 *weight* 属性值更新其在加权轮盘赌选择机制中的累计概率,如图 6 中步骤 7*/
9. **if** (***D***$_i$.*construction*==*random*){*weightofrandom*+=***D***$_i$.*weight*;}
10. **elseif** (***D***$_i$.*construction*==*chaotic*){*weightofchaotic*+=***D***$_i$.*weight*;}
11. **else** {*weightofgenetic*+=***D***$_i$.*weight*;}
12. }
13. *sum*=*weightofrandom*+*weightofchaotic*+*weightofgenetic*;
14. p_r[0]=*weightofrandom*/*sum*;
15. p_r[1]=*weightofchaotic*/*sum*;
16. p_r[2]=*weightofgentic*/*sum*;
17. **End**

2.4 算法分析

(1) 时间复杂度分析

设 N_s 为自体数量,待检测样本规模为 N,则检测器生成所需要的总时间复杂度为 $O((k+3m)N_c\times N_s)$,检测所需总时间复杂度为 $O(N_d\times N)$.

证明:在检测器生成阶段,初始化参数和选定生成源的时间复杂度均为 O(1).随机和混沌映射方式的时间代价为 $O(N_c)$.采用遗传变异构造候选检测器时,计算种群 k 个个体适应度和选择 m 个优势个体的时间复杂度为 $O(k+m)$,交叉和变异操作的时间复杂度均为 $O(((k+3m)N_c\times N_s)m)$.因此,遗传变异构造候选检测器的时间复杂度为 $O(k+3m)N_c$.从而有,检测器生成所需要的总时间复杂度为 $O((k+3m)N_c\times N_s)$,这种时间代价在检测器生成阶段是可接受的[9].

在检测阶段,需待检测样本与检测器进行亲和力计算,所需总时间复杂度为 $O(N_d\times N)$,这种时间代价在检测阶段也是可接受的. □

(2) 检测器生成的收敛性分析

建立吸收态 Markov 过程模型,证明其是可收敛到全局最优解的.

定义 1. 设 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 对任意 n 个不同的 $t_1,t_2,\ldots,t_n\in T$ 且 $t_1<t_2<\ldots<t_{n-1}$,有:

$$P(\boldsymbol{X}(t_n)\leqslant\boldsymbol{X}_n|\boldsymbol{X}(t_{n-1})=\boldsymbol{X}_{n-1},\ldots,\boldsymbol{X}(t_1)\leqslant\boldsymbol{X}_1)=P(\boldsymbol{X}(t_n)\leqslant\boldsymbol{X}_n|\boldsymbol{X}(t_{n-1})=\boldsymbol{X}_{n-1}),$$

则称 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 为 Markov 过程[9].

定义 2. 任意给定某一区域 Markov 过程 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 和该区域的最优状态空间 $\boldsymbol{Y}^*\subset\boldsymbol{Y}$,若满足:

$$P(\boldsymbol{X}(t+1)\notin\boldsymbol{Y}^*|\boldsymbol{X}(t)\in\boldsymbol{Y}^*)=0,$$

则称 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 为一个吸收态 Markov 过程[9].

引理 1. 设生成检测器至最优状态的过程为 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$,则 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 具有 Markov 性.同时,$\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 也是一个吸收态 Markov 过程[9].

证明:由多源邻域检测器生成过程可知,$\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 为离散时间的随机过程.因为检测器在当前迭代中的状态 $\boldsymbol{X}(t)$只由 $\boldsymbol{X}(t-1)$决定,$\boldsymbol{X}(0)$在初始化时可随机选取,因此,

$$P(\boldsymbol{X}(t)|\boldsymbol{X}(t-1),\boldsymbol{X}(t-2),\ldots,\boldsymbol{X}(0))=P(\boldsymbol{X}(t)|\boldsymbol{X}(t-1)).$$

即 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 具有 Markov 性.

由算法流程可知,解空间是有限的状态空间;由检测器多源生成过程可知:当 $\boldsymbol{X}(t)\in\boldsymbol{Y}^*$为最优解空间时,$\boldsymbol{X}(t+1)\in\boldsymbol{Y}^*$.因此,

$$P(\boldsymbol{X}(t+1)\notin\boldsymbol{Y}^*|\boldsymbol{X}(t)\in\boldsymbol{Y}^*)=0.$$

即 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 是一个吸收态 Markov 过程.证毕. □

定义 3. 给定某一区域的吸收态 Markov 过程 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}(\forall \boldsymbol{X}(t)\in \boldsymbol{Y})$ 和最优状态空间 $\boldsymbol{Y}^*\subset \boldsymbol{Y}$.记 $\mu(t)=P(\boldsymbol{X}(t)\in \boldsymbol{Y}^*)$ 表示时刻 t 在某区域达到最优状态的概率.若 $\lim\limits_{t\to\infty}\mu(t)=1$, 则称 $\{\boldsymbol{X}(t)\}_{t=0}^{\infty}$ 收敛[9].

定理 1. 多源邻域检测器生成过程是以概率 1 收敛的.

证明:通过 Markov 过程模型来描述本文算法中状态的转移过程,将初始群体的全部近似解认为是状态 $\boldsymbol{S}^1$:$\boldsymbol{S}^1$ 等于 $\boldsymbol{Y}^n$ 中的某个点.中间种群规模为 $\boldsymbol{S}$ 的群体的全部近似解认为是 $\boldsymbol{S}^2$,$\boldsymbol{S}^2=\boldsymbol{Y}^N$.当没有必要区分 $\boldsymbol{S}^1$ 和 $\boldsymbol{S}^2$ 时,用 $\boldsymbol{S}$ 代替.用 $\boldsymbol{s}_i\in\boldsymbol{S}$ 表示 $\boldsymbol{s}_i$ 是 $\boldsymbol{S}$ 中的一个状态, $\boldsymbol{X}_t^i$ 表示在第 t 代种群 $\boldsymbol{X}_t$ 处于状态 $\boldsymbol{s}_i$,设:

$$\boldsymbol{s}_i=(x^1,x^2,\ldots,x^n)\in\boldsymbol{S}.$$

记 $f(\boldsymbol{s}_i)=(f(x^1),f(x^2),\ldots,f(x^n))$,若 $f(\boldsymbol{s}_i)=f(\boldsymbol{s}_j)$或者 $f(\boldsymbol{s}_i)-f(\boldsymbol{s}_j)$的第 1 个非零分量为正,则记为 $f(\boldsymbol{s}_i)\geqslant f(\boldsymbol{s}_j)$.此外,记 $I=\{i|\boldsymbol{s}_i\geqslant \boldsymbol{s}_j,\forall \boldsymbol{s}_j\in\boldsymbol{S}\}$.由以上定义可知,若 $i\in I$,则 $\boldsymbol{s}_i=(x^1,x^2,\ldots,x^n)$满足:

$$f(x^1)=f(x^2)=\ldots=f(x^n)=f^*.$$

因此,$\boldsymbol{s}_i\cap\boldsymbol{Y}^*\neq\varnothing$.

设随机过程$\{\boldsymbol{X}_t\}$的转移概率为 p_{ij}:

$$p_{ij}(t)=p\left\{\frac{\boldsymbol{X}_{t+1}^j}{\boldsymbol{X}_t^i}\right\}.$$

本文算法采用 Monte Carlo 方法估计覆盖率来进行下一次检测器的生成,所以在任意的 $t\geqslant 0$ 时,有 $f(\boldsymbol{X}_{t+1})>f(\boldsymbol{X}_t)$.

当 $i\in I,j\notin I$ 时,有:

$$p_{ij}=0.$$

即:如果父代中出现了最优解,那么无论再经过多少代,它都不会变差.

当 $i\notin I,j\in I$ 时,因为 $f(\boldsymbol{X}_{t+1})>f(\boldsymbol{X}_t)$,所以:

$$p_{ij}>0 \tag{6}$$

设 $p_i(t)$为种群 $\boldsymbol{X}_t$ 处在状态 $\boldsymbol{s}_i$ 的概率, $p(t)=\sum\limits_{i\notin I}p_i(t)$, 由 Markov 的性质可知:

$$p(t+1)=\sum_{s_i\notin S}\sum_{j\notin I}p_i(t)p_{ij}(t)=\sum_{i\in I}\sum_{j\notin I}p_i(t)p_{ij}(t)+\sum_{i\notin I}\sum_{j\notin I}p_i(t)p_{ij}(t) \tag{7}$$

由于:

$$\sum_{i\notin I}\sum_{j\notin I}p_i(t)p_{ij}(t)+\sum_{i\notin I}\sum_{j\in I}p_i(t)p_{ij}(t)=\sum_{i\notin I}p_i(t)=p_t,$$

所以:

$$\sum_{i\notin I}\sum_{j\notin I}p_i(t)p_{ij}(t)=p_t-\sum_{i\in I}\sum_{j\in I}p_i(t)p_{ij}(t) \tag{8}$$

把公式(8)代入公式(6)和公式(7),可得:

$$0\leqslant p_{t+1}<\sum_{i\in I}^{n}\sum_{j\notin I}^{n}p_i(t)p_{ij}(t)+p_t=p_t.$$

因此, $\lim\limits_{t\to\infty}p_t=0$. 又因为:

$$\lim_{t\to\infty}\{f_t=f^*\}=1-\lim_{t\to\infty}\sum_{i\notin I}p_i(t)=1-\lim_{t\to\infty}p_t,$$

所以可得:

$$\lim_{t\to\infty}p\{f_t=f^*\}=1.$$

可证得该算法以概率 1 收敛.证毕. □

3 实验分析

衡量异常检测的检测性能需要测定检测器的计算效率、检测率(detection rate,简称 DR)和误报率(false-

positive rate,简称 FR)等评价指标[9]:

$$DR=\frac{\text{检测到的入侵数}}{\text{数据集中入侵总数}},$$

$$FR=\frac{\text{被误报为入侵的正常数据}}{\text{数据集中正常数据}}.$$

实验分为 3 部分:第 1 部分为主要参数的测定,第 2 部分为稳定性测试,第 3 部分为与其他算法的对比实验.

实验选用本领域的权威数据集 KDD CUP 1999 中的一个 10%数据子集,KDD CUP 1999 数据集是美国国防部高级规划署(DARPA)在 MIT 林肯实验室收集的一个用于异常检测、机器学习等研究领域的权威数据集,该数据集包含了 39 种异常类型,每个样本有 41 维不同类型的属性[33].因此,实验首先需要进行数据预处理,将其中的离散型属性转换成连续型属性.例如,其中的协议属性,实验设定的变换规则为 *TCP*→1、*UDP*→2、*ICMP*→3等[9].

3.1 参数测定

实验首先进行匹配阈值(ρ)、邻域划分步长(*step*)的取值测试,这两个参数的取值是否合适,将直接影响到候选检测器的耐受训练时间、检测器的分布效果以及检测算法的最终整体检测性能.因此,本节实验基于以上 3 个出发点进行设计,由两部分组成:第 1 部分是测试各个ρ值下,检测器的生成时间、检测率和误报率,并择优选择ρ值,将其取值范围缩小到一个较小的区间,然后进一步测试每种 *step* 对检测率和误报率的影响,确定检测效果较好的ρ和 *step* 的取值组合;第 2 部分以选取的[ρ,*step*]取值组合进行检测性能的稳定性测试,最终确定算法的参数.

(1) 参数取值组合实验

实验分为 3 部分:第 1 部分为主要参数的测定,第 2 部分为稳定性测试,第 3 部分为与其他算法的对比实验.

实验选用本领域的权威数据集 KDD CUP 1999 数据集的一个 10%数据子集,KDD CUP 1999 数据集是美国国防部高级规划署(DARPA)在 MIT 林肯实验室收集的一个用于异常检测、机器学习等研究领域的权威数据集,该数据集包含了 39 种异常类型,每个样本有 41 维不同类型的属性[33].因此,实验首先需要进行数据预处理,将其中的离散型属性转换成连续型属性,例如其中的协议属性,实验设定的变换规则为 *TCP*→1、*UDP*→2、*ICMP*→3等[9].

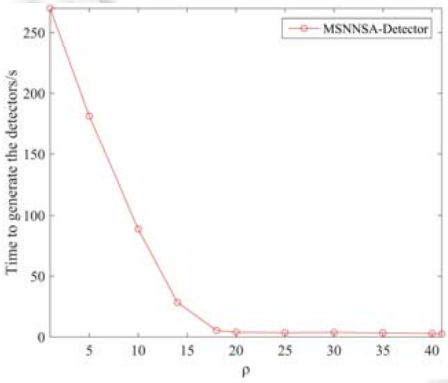


Fig.7 Time to generate detectors

图 7 检测器生成时间

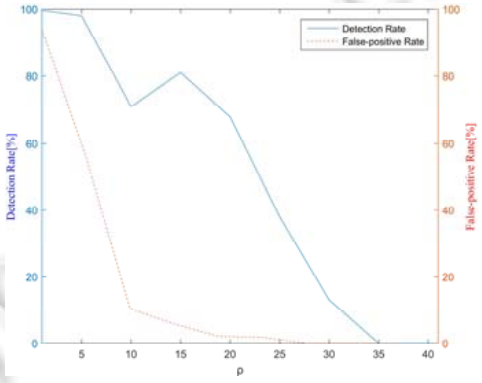


Fig.8 Detection rate and false-positive rate with different thresholds

图 8 不同ρ值下获得的检测器的检测率和误报率

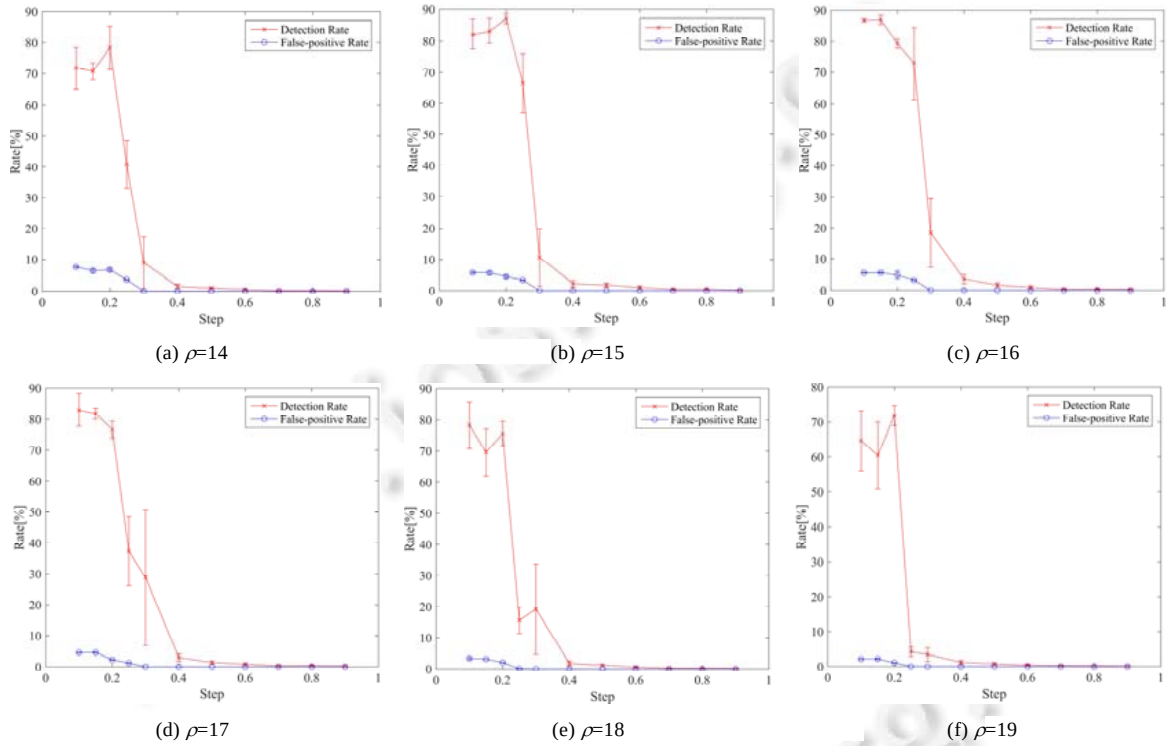
最后,从数据集中分别随机抽取正常样本的 30%、50%、70%作为训练自体集,参数设置同上.然后,随机抽取样本组成 3 组测试数据集,每组样本选取比例同上.同样,每组实验进行 3 次,结果取均值和标准差如图 9 所示.从中可以看出:随着 *step* 的不断增大,各个ρ值下的检测率都呈逐渐下降趋势,当 *step*≥0.25 时,检测效果不尽理

想.同时,随着 *step* 的不断增大,不同ρ值下的误报率逐渐降低,在 *step*>0.3 时,误报率降至 0,但与此同时,检测率也几乎降至 0.

综合以上实验结果可以看出:在当前实验环境下,当[ρ,*step*]取值为[16,0.1]、[16,0.15]和[15,0.2]时,检测的整体效果更为理想;而且,在这几组参数下的训练自体规模对于检测器性能的影响较小.

(2) 基于检测稳定性的[ρ,*step*]确定实验

由上述[ρ,*step*]取值组合进一步实验,通过观测检测稳定性确定最终参数取值.实验使用上一部分实验中的50%自体样本集合,参数设置同上.然后,随机抽取样本组成 5 组测试数据集,样本选取比例同上.在每组参数取值组合下,分别测定这 5 组数据的检测率,其中,每组数据测定 5 次,取测试结果的均值和标准差,结果如图 10 所示.

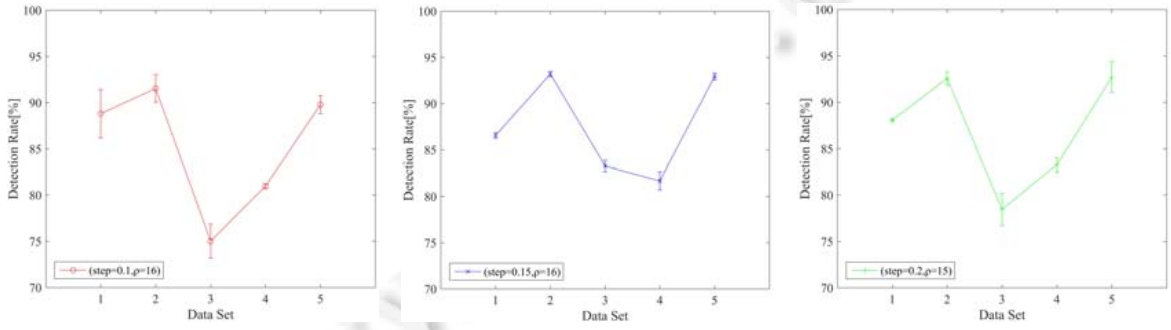


(a) ρ=14 (b) ρ=15 (c) ρ=16

(d) ρ=17 (e) ρ=18 (f) ρ=19

Fig.9 Detection rate and false-positive rate with different [ρ,*step*]

图 9 不同[ρ,*step*]取值下的检测率和误报率



(a) [ρ,*step*]=[16,0.1] (b) [ρ,*step*]=[16,0.15] (c) [ρ,*step*]=[15,0.2]

Fig.10 Stability experiments with different [ρ,*step*]

图 10 不同[ρ,*step*]算法稳定性实验

从中可以看出:每组参数在不同测试数据集下所表现出来的性能都有所波动.这是因为,算法在不同参数组合下对于不同类型的数据认知能力不同.同时,从结果的标准差也可以看出,每种参数组合下检测的整体稳定性也有所差别.

[ρ,*step*]与二进制的比特匹配和实值的半径匹配一样,都是检测器最重要的参数,直接影响了算法的最终效果.而[ρ,*step*]结合了两个方面,可有效降低二进制和实值匹配方法单匹配规则的敏感度.以上实验结果也可以佐证该优势.

综上,结合以上实验结果,当[ρ,*step*]取值为[16,0.15]时,算法表现出了最佳的性能.由此,后续实验选定该值组合进行.

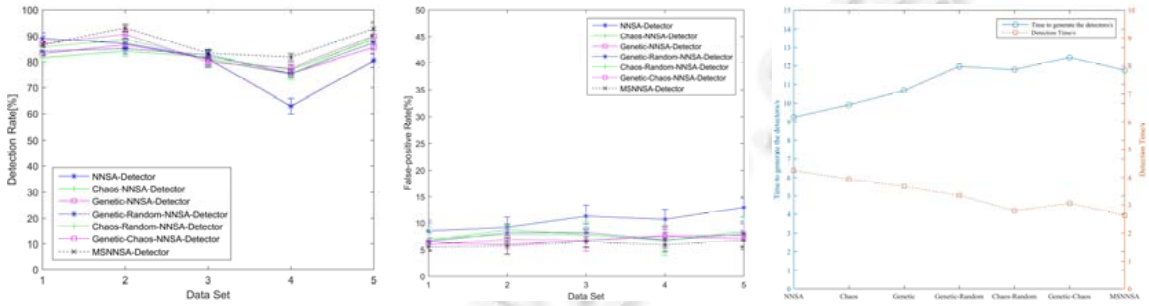
3.2 单源、双源和多源免疫检测器对比实验

为了验证本文算法的多源优越性,本部分实验对比分别使用随机、遗传算法、混沌映射这 3 种单源构造候选邻域检测器的检测器生成算法、两两组合的双源构造候选邻域检测器的检测器生成算法与 MSNNSA 在生成时间和检测性能方面的效果.本实验使用上一部分实验中的 30%自体样本集合和选取的 5 组测试数据集,参数设置同上.

检测率和误报率的实验结果如图 11(a)和图 11(b)所示.从图 11(a)可以看出:在第 1 组测试数据集中,MSNNSA 生成的检测器的检测率仅低于 NNSA(即随机单源方式)的检测器;但从检测稳定性和对不同测试数据集的检测率曲线上看,NNSA 最不稳定,遗传和混沌机制较为接近,而双源方式彼此之间差别不大,但要略优于 3 种单源方式生成的检测器.MSNNSA 优于其他 6 种单源和多源方式.从图 11(b)中可以看出:NNSA 的误报率最高,遗传和混沌机制次之且具有一定的波动性,双源方式在稳定性方面略好于单源方式,MSNNSA 的误报率最低且最稳定.

图 11(c)为 7 种方式生成 300 个检测器的生成时间和对应的检测时间对比.从生成时间结果对比中可以看出:采用随机方式所花费的时间最短,混沌映射和遗传算法次之,多源方式接近或优于双源方式.这主要是由于 MSNNSA 双/多源选择和迭代次数要高于单源方式.但从检测效果(如图 11(a)和图 11(b)所示)可以看出:这种时间代价在检测器生成对时间性能要求较低的阶段是可接受的,而且生成的检测器性能更高,因此也是值得的.而且,MSNNSA 的检测时间花费最短.这也就说明本文的检测方法兼顾二进制和实值优势,在保证整体检测性能更好的基础上,效率更高.

因此,从以上分析可以看出:MSNNSA 生成了更高质量的检测器,在检测率和误报率及其稳定性方面要优于其他 6 种单/双源方式.



(a) 检测率 (b) 误报率 (c) 生成和检测时间

Fig.11 Experimental results' comparison of MSNNSA and other 3 single-source-inspired & 3 2-source-inspired detectors generation algorithms

图 11 MSNNSA 与 3 种单源、3 种双源检测器生成算法的实验结果对比

3.3 与其他代表算法的检测性能对比实验

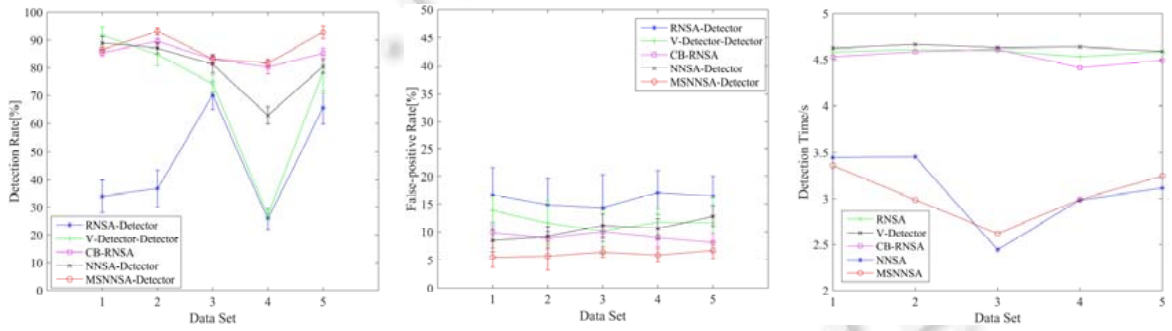
为了进一步验证本文算法的性能,本部分实验选择其他本领域相关的代表算法——实值 NSA(real-valued NSA,简称 RNSA)、V-Detector 和 NNSA 的检测器生成算法,与前文提到的 CB-RNSA[12]进行对比.CB-RNSA 引入克隆选择,采用基于自体集的聚类中心在分层有限范围内随机生成候选检测器,比其他同类型算法及改进具有更好的时间效率和检测器生成质量.实验包括两部分:(1) 采用最优参数组合的 MSNNSA 与其他算法的对比,以验证 MSNNSA 算法的优势;(2) 采用代表性参数组合的 MSNNSA 与 NNSA 和 RNSA 算法的对比,以验证 MSNNSA 算法对参数的较低敏感性.对比算法采用的参数都基于最好的设置进行.

(1) 最优参数组合的 MSNNSA 与其他算法的对比

实验使用上一部分实验中的 30%自体样本集合和选取的 5 组测试数据集,参数设置同上.

检测率和误报率的实验结果如图 12(a)和图 12(b)所示,RNSA 在 5 组测试集中的检测率均是最低的,而误报率却很高且稳定性差.NNSA 的检测率优于 V-Detector、RNSA 和 MSNNSA,但逊于 CB-RNSA.在误报率方面,NNSA 优于 V-Detector 和 RNSA,在 2-4 组测试集中略逊于 CB-RNSA,整体检测稳定性也略逊于 CB-RNSA.MSNNSA 的检测器的检测率,除在第 1 组测试集中劣于 V-detector 和 NNSA,并与 CB-RNSA 持平外,在其余各组测试集中均优于其余算法.而且,在误报率和检测的整体稳定性方面,MSNNSA 也均优于其余算法.整体对比,邻域形态空间(MSNNSA 与 NNSA)的整体均要优于实值形态空间(CB-RNSA,V-Detector 和 RNSA).而且在邻域形态空间下,MSNNSA 的检测器的检测效果优于基于 NNSA 的检测器.

图 12(c)所示为 5 种算法生成的 300 个检测器的检测时间对比情况.从中可以看出:在相同条件下,邻域形态空间(MSNNSA 与 NNSA)下检测器的检测时间要明显优于实值形态空间(CB-RNSA、V-Detector 和 RNSA)下的检测时间.



(a) 检测率 (b) 误报率 (c) 检测时间

Fig.12 Experimental results' comparison of the 5 algorithms

图 12 5 种算法实验结果对比

因此可以看出:邻域形态空间可更好地解决人工免疫系统的"黑洞"、高重叠与"维度灾难"问题,降低问题求解的计算代价,提高计算效率,获得更好的检测性能.而且,多源方式也提高了 NNSA 的检测器生成质量,使其优于实值形态空间最具代表性的一些算法.

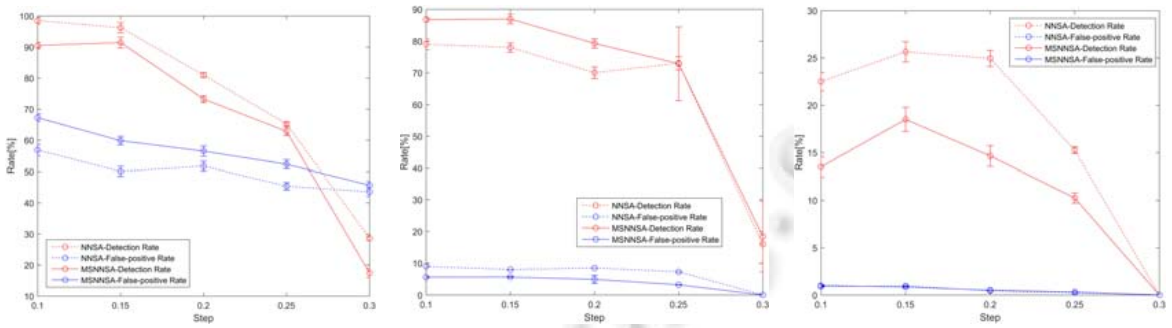
(2) 采用代表性参数组合的 MSNNSA 与 NNSA 和 RNSA 算法的对比

本部分实验选取上一部分数据集中的一组进行,对比不同[ρ,*step*]组合下 MSNNSA、NNSA 与最好参数的 RNSA 的实验结果,参数设置同上.

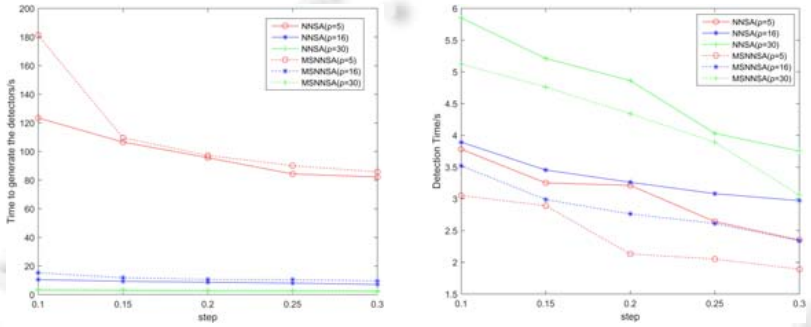
图 13 所示为在选取几组不同的 *step* 和ρ下,NNSA 与 MSNNSA 的性能对比情况:ρ为较小值时,两种算法的检测率和误报率都较高;ρ为较大值时,两种算法的检测率和误报率都较低;只有ρ为 16 左右时,两种算法的效果最为理想.在不同参数组合下,两种算法的检测率和误报率均随 *step* 值的增加而减小.结合图 11 和图 12 的实验结果综合可得:MSNNSA 在不同参数组合下的效果均优于 NNSA,且在一个相对合理的范围内($\rho\in$[14,18],*step*$\in$

[0.1,0.2])的效果都是可接受的.

图 14 为 *step*=0.15 时,在不同ρ下,MSNNSA 与 RNSA 最优实验结果的对比.从图 14(a)和图 14(b)可以看出:MSNNSA 检测率和误报率都随ρ的增加而不断降低,但在较大区间内,都好于 RNSA 的最优情况.由此可以看出,MSNNSA 对于[ρ,*step*]的敏感性要低于 RNSA 对于匹配尺度的敏感性.从图 14(c)可以看出:MSNNSA 检测时间在任何情况下都优于 RNSA 的最优情况;MSNNSA 检测器生成时间随ρ的增加而不断降低,在合理取值范围下限之后开始优于 RNSA 的最优情况.



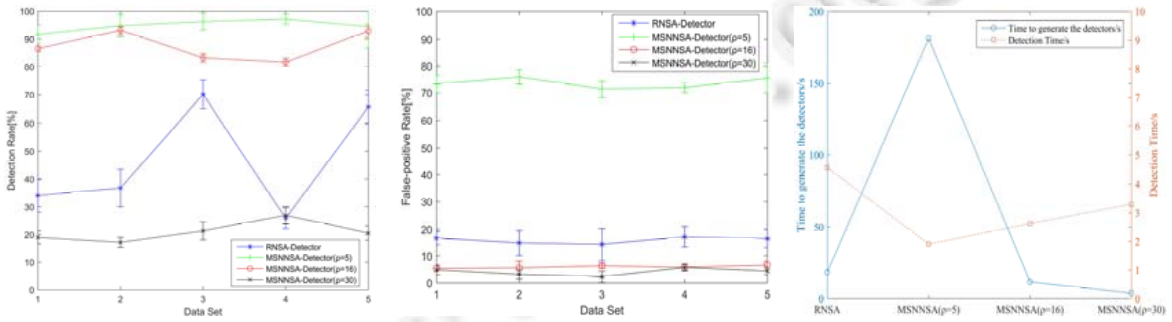
(a) ρ=5 时,不同 *step* 的检测效果 (b) ρ=16 时,不同 *step* 的检测效果 (c) ρ=30 时,不同 *step* 的检测效果



(d) 不同[ρ,*step*]组合的检测器生成时间 (e) 不同[ρ,*step*]组合的检测时间

Fig.13 Experimental results' comparison between MSNNSA and NNSA with different [ρ,*step*]

图 13 不同[ρ,*step*]组合的 MSNNSA 与 NNSA 的实验结果对比



(a) 检测率 (b) 误报率 (c) 生成和检测时间

Fig.14 Experimental results' comparison between MSNNSA with different ρ when *step*=0.15 and RNSA

图 14 *step*=0.15 时,不同ρ取值的 MSNNSA 与 RNSA 算法的实验结果对比

因此,综合以上所有实验可以看出,MSNNSA 对于参数的敏感性低于实值检测算法,在参数较为合理的范

围内,算法的执行效率和检测结果均优于实值检测算法.

综上,以上各实验,从确定关键参数取值到测试单源、多源和多源候选检测器的效果,并对比其他相关代表算法的检测性能可以看出,本文算法得到了更为理想的结果.

(1) 通过参数测定实验可以看出:MSNNSA 的[ρ,*step*]参数组合的合理范围要大于实值形态空间下的匹配半径的设定区间,其对超参数的敏感程度低于实值检测算法;

(2) MSNNSA 及检测模型基于邻域形态空间,并设计了更符合实际的匹配规则,相比于实值形态空间,MSNNSA 极大地减少了数据维度对算法的影响;

(3) 通过对比实验可以看出:MSNNSA 由于在生成候选检测器阶段利用 3 种机制对检测器进行了优化与训练,使得在各种性能指标上要优于基于单/双源机制的原始 NNSA;

(4) 通过与其他实值形态空间代表算法的对比实验中可以看出:MSNNSA 虽然在检测器生成上花费的时间较长,但仍在可接受范围内,而且生成的检测器基于更合理的亲和力计算方法,检测效率、性能和稳定性都优于其他几种算法.

这就说明邻域形态空间相比于实值形态空间具有维度无关性,且亲和力计算速率更快,并基于多源构造检测器的方式可使得邻域检测器获得更全面的知识,配以适合邻域和实值样本的检测算法,可使人工免疫更好地应用于异常检测等相关应用中.但本文算法是基于 *step* 进行邻域划分,自适应程度不够.下一阶段将继续进行改进和完善使其更适合动态环境下的邻域转换操作,从而适应具有更复杂变化特征的异常检测应用.

4 结 论

人工免疫系统有两个重要的研究对象:形态空间与检测器.本文针对人工免疫系统在实值形态空间中存在的“维度灾难”“黑洞”、高重叠等主要问题,以候选检测器构造机制为出发点,改进邻域形态空间否定选择算法,提出了邻域形态空间多源免疫检测器生成与检测方法.算法通过引入随机、混沌映射和遗传算法这 3 种机制构造候选检测器,优势互补,从而可以有效提升生成候选检测器的全局性和靶向性,提高检测器的分布性能.实验结果表明:本文算法提高了检测器的生成速率和分布性能,以此为基础的检测算法也表现出了很好的计算效率和检测水平.

相比于传统的实值形态空间和匹配策略,本文算法在算法收敛性和运行速率方面提高明显,同时也可保障系统的检测性能有效提高.基于此的异常检测/入侵检测方法可以较好地满足实际应用的实时需求.在不同异常检测应用背景下的改进和实际测试是未来研究的重点.

References:

[1] Zahra Z, Mohammad SH, Akbar R, Kamran K. A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework. Expert Systems with Applications, 2018,133:301−314. [doi: 10.1016/j.eswa.2018.06.047]

[2] Louati A, Darmoul S, Elkosantini S, ben Said L. An artificial immune network to control interrupted flow at a signalized intersection. Information Sciences, 2018,433:70−95. [doi: 10.1016/j.ins.2017.12.033]

[3] Xu Y, Yuan F, Lin Q, Deyou T, Li D. Merging event logs for process mining with a hybrid artificial immune algorithm. Ruan Jian Xue Bao/Journal of Software, 2018,29(2):396−416 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5253.htm [doi: 10.13328/j.cnki.jos.005253]

[4] Magna G, Casti P, Jayaraman SV, Salmeri M, Mencattini A, Martinelli E, Natale CD. Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system. Knowledge-based Systems, 2016,101:60−70. [doi: 10.1016/j.knosys.2016.02.019]

[5] Fernandes G, Rodrigues JJPC, Carvalho LF, Al-Muhtadi JF, Proença ML. A comprehensive survey on network anomaly detection. Telecommunication Systems, 2019,70(3):447−489. [doi: 10.1007/s11235-018-0475-8]

[6] Ring M, Schlör D, Landes D, Hotho A. Flow-based network traffic generation using generative adversarial networks. Computers & Security, 2019,82:156−172. [doi: 10.1016/j.cose.2018.12.012]

[7] Eduardo V, Altair S, Alysson B, Nuno N. BigFlow: Real-time and reliable anomaly-based intrusion detection for high-speed networks. Future Generation Computer Systems, 2019,93:473−485. [doi: 10.1016/j.future.2018.09.051]

[8] James Z, Robert G, Ilija V. Anomaly detection in wide area network meshes using two machine learning algorithms. Future Generation Computer Systems, 2019,93:418−426. [doi: 10.1016/j.future.2018.07.023]

[9] Yao ZY. Research on adaptive model of neighborhood immune detector [MS. Thesis]. Harbin: Harbin University of Science and Technology, 2020 (in Chinese with English abstract). [doi: 10.27063/d.cnki.ghlgu.2020.000878]

[10] Wen C, Santin A, Tao L. Parameter analysis of negative selection algorithm. Information Sciences, 2017,420:218−234. [doi: 10.1016/j.ins.2017.08.062]

[11] Yang T, Chen W, Li T. An antigen space density based real-value negative selection algorithm. Applied Soft Computing, 2017,61: 860−874. [doi: 10.1016/j.asoc.2017.09.005]

[12] Zhang RR, Xiao X. A clone selection based real-valued negative selection algorithm. Complexity, 2018. [doi: 10.1155/2018/2520940]

[13] Zhang FB, XI L, Wang DW, Yue X. Neighborhood shape-space and detection algorithm. Control and Decision, 2011,26(10): 1562−1566 (in Chinese with English abstract). [doi: 10.13195/j.cd.2011.10.125.zhangfb.012]

[14] Jiao LC, Du HF. Development and prospect of the artificial immune system. Acta Electronica Sinica, 2003,31(10):1540−1548 (in Chinese with English abstract). [doi: 10.3321/j.issn:0372-2112.2003.10.024]

[15] Yan Q, Xie WX. Research on and development of anomaly detection. Journal of Xidian University (Natural Science Edition), 2002, 29(1):128−132 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-2400.2002.01.030]

[16] Yang DY, Chen JY. Research on detector generation algorithm based on multiple population GA. Acta Automatica Sinica, 2009, 35(4):425−432 (in Chinese with English abstract). [doi: 10.3724/sp.j.1004.2009.00425]

[17] Sayed GI, Khoriba G, Haggag MH. A novel chaotic salp swarm algorithm for global optimization and feature selection. Applied Intelligence, 2018,48(10):3462−3481. [doi: 10.1007/s10489-018-1158-6]

[18] Xia Z, Wang X, Zhou W, Li R, Wang C, Zhang C. Color medical image lossless watermarking using chaotic system and accurate quaternion polar harmonic transforms. Signal Processing, 2019,157:108−118. [doi: 10.1016/j.sigpro.2018.11.011]

[19] Wu T, Jin JG, Wei MJ. A Hash function algorithm based on variable parameter cascading chaos. Journal of Computer Research and Development, 2016,53(3):674−681 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20148155]

[20] Irshad A, Sher M, Chaudhary SA, Naqvi H, Farash MS. An efficient and anonymous multi-server authenticated key agreement based on chaotic map without engaging registration centre. Journal of Supercomputing, 2016,72(4):1−22. [doi: 10.1007/s11227-016-1688-9]

[21] Zhang N, Li ZZ, Zhang JH. Negative selection algorithm based on chaos theory. Journal of Sichuan University (Engineering Science Edition), 2006,38(1):124−127 (in Chinese with English abstract). [doi: 10.3969/j.issn.1009-3087.2006.01.026]

[22] Asgari-Chenaghlu M, Balafar MA, Feizi-Derakhshi, MR. A novel image encryption algorithm based on polynomial combination of chaotic maps and dynamic function generation. Signal Processing, 2019,157:1−13. [doi: 10.1016/j.sigpro.2018.11.010]

[23] Yousefi M, Yousefi M, Ferreira RPM, Kim JH, Fogliatto FS. Chaotic genetic algorithm and adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. Artificial Intelligence in Medicine, 2017,84:23−33. [doi: 10.1016/j.artmed.2017.10.002]

[24] Sayed GI, Khoriba G, Haggag MH. A novel chaotic salp swarm algorithm for global optimization and feature selection. Applied Intelligence, 2018,48(10):3462−3481. [doi: 10.1007/s10489-018-1158-6]

[25] Cui L, Xie XP, Wang XW, Luo Y, Liu J. Event-triggered single-network ADP method for constrained optimal tracking control of continuous-time non-linear systems. Applied Mathematics and Computation, 2019,352:220−234. [doi: 10.1016/j.amc.2019.01.066]

[26] Zhang FB, Wang TB. Negative selection algorithm for real-valued *n*-dimensional chaotic maps. Journal of Computer Research and Development, 2013,50(7):1387−1398 (in Chinese with English abstract).

[27] Owais M, Osman MK. Complete hierarchical multi-objective genetic algorithm for transit network design problem. Expert Systems with Applications, 2018,114:143−154. [doi: 10.1016/j.eswa.2018.07.033]

[28] Das AK, Das S, Ghosh A. Ensemble feature selection using bi-objective genetic algorithm. Knowledge-based Systems, 2017,123: 116−127. [doi: 10.1016/j.knosys.2017.02.013]

[29] Tavakkoli-Moghaddam R, Safari J, Sassani F. Reliability optimization of series-parallel systems with a choice of redundancy strategies using a genetic algorithm. Reliability Engineering & System Safety, 2017,93(4):550−556. [doi: 10.1016/j.ress.2007.02.009]

[30] Jiang HW, Zeng GB, Hu KK. An image privacy anonymity privacy protection method based on genetic algorithm. Journal of Computer Research and Development, 2016,53(10):2354−2364 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20160435]

[31] Keshavarz H, Abadeh MS. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of Microblogs. Knowledge-based Systems, 2017,122:1−16. [doi: 10.1016/j.knosys.2017.01.028]

[32] Qi HL, Zhao H, Liu WW, Zhang HB. Parameters optimization and nonlinearity analysis of grating eddy current displacement sensor using neural network and genetic algorithm. Journal of Zhejiang University Science (A), 2009,10(8):1205−1212. [doi: 10.1631/jzus.A0820564]

[33] The UCI KDD archive: KDD CUP 1999 data set. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

附中文参考文献:

[3] 徐杨,袁峰,林琪,汤德佑,李东.基于混合人工免疫算法的流程挖掘事件日志融合方法.软件学报,2018,29(2):396−416. http://www.jos.org.cn/1000-9825/5253.htm [doi: 10.13328/j.cnki.jos.005253]

[9] 姚之钰.邻域免疫检测器自适应模型研究[硕士学位论文].哈尔滨:哈尔滨理工大学,2020. [doi: 10.27063/d.cnki.ghlgu.2020.000878]

[13] 张凤斌,席亮,王大伟,等.邻域形态空间与检测算法.控制与决策,2011,26(10):1562−1566. [doi: 10.13195/j.cd.2011.10.125.zhangfb.012]

[14] 焦李成,杜海峰.人工免疫系统进展与展望.电子学报,2003,31(10):1540−1548. [doi: 10.3321/j.issn:0372-2112.2003.10.024]

[15] 阎巧,谢维信.异常检测技术的研究与发展.西安电子科技大学学报(自然科学版),2002,29(1):128−132. [doi: 10.3969/j.issn.1001-2400.2002.01.030]

[16] 杨东勇,陈晋音.基于多种群遗传算法的检测器生成算法研究.自动化学报,2009,35(4):425−432. [doi: 10.3724/sp.j.1004.2009.00425]

[19] 吴涛,金建国,魏明军.一种基于变参级联混沌的 Hash 函数算法.计算机研究与发展,2016,53(3):674−681. [doi: 10.7544/issn1000-1239.2016.20148155]

[21] 张楠,李志蜀,张建华.基于混沌理论的否定选择算法.四川大学学报(工程科学版),2006,38(1):124−127. [doi: 10.3969/j.issn.1009-3087.2006.01.026]

[26] 张凤斌,王天博.实值 *n* 维混沌映射否定选择算法.计算机研究与发展,2013,50(7):1387−1398.

[30] 姜火文,曾国荪,胡克坤.一种遗传算法实现的图聚类匿名隐私保护方法.计算机研究与发展,2016,53(10):2354−2364. [doi: 10.7544/issn1000-1239.2016.20160435]

席亮(1983－),男,博士,副教授,主要研究领域为人工智能及应用,深度学习,网络与信息安全,物联网安全.

张凤斌(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能及应用,深度学习,网络与信息安全.

姚之钰(1994－),女,硕士,主要研究领域为人工智能及应用,网络与信息安全.