基于矩阵形式的否定选择算法研究

张雄美　易昭湘　宋建社　李俊山
(西安高技术研究所　西安　710025)

摘　要：现有的状态空间表示形式和匹配规则已经成为否定选择算法研究的瓶颈。为此，该文将状态空间从向量扩展到矩阵，提出了一种基于矩阵形式的否定选择算法。引入矩阵表示自我和非我空间，定义了元素匹配距离，在此基础上建立了双向匹配规则；同时根据状态空间特征建立了基于覆盖检验的检测器生成算法。实验结果表明该算法性能明显优于实值否定选择算法，有效解决了检测率和误报率联动的问题，且能产生更为高效的检测器。

关键词：覆盖检验；否定选择算法；双向匹配
中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1009-5896(2010)11-2701-06
DOI: 10.3724/SP.J.1146.2009.01489

Research on Negative Selection Algorithm Based on Matrix Representation

Zhang Xiong-mei　Yi Zhao-xiang　Song Jian-she　Li Jun-shan
(*Xi'an Research Institute of Hi-Tech Hongqing Town, Xi'an* 710025, *China*)

Abstract: Due to the bottleneck of the current representation of the state space and match rule in the negative selection algorithm, a negative selection algorithm based on the matrix representation is presented, which extends the state space from the vector to the matrix. The elemental match distance is defined by introducing the matrix to denote self and nonself space, the bi-directional match rule is established. Moreover, a detector generating algorithm based on coverage rate testing is developed according to the characteristics of state space. The experimental results show that the proposed algorithm achieves better performance than the real-valued negative selection algorithm, and solves effectively the problem of the linkage of the detection rate and false rate. Furthermore, it is verified to generate more effective detectors.

Key words: Coverage rate testing; Negative selection algorithm; Bi-directional match

1　引言

否定选择算法是根据生物免疫系统中否定选择原理所提出的人工免疫算法[1]，由于该算法适合于小样本检测，不需要先验知识，且具有很强的鲁棒性和并行性等优点，因而在模式识别、病毒检测、网络入侵检测、异常检测等工程领域得到了广泛应用[2−5]。在应用过程中，否定选择算法得到不断的改进，主要表现在：(1)状态空间表示形式由最初的二进制字符串拓展到实数值向量[5]，更能体现出样本的数据特征，接近于最初的问题空间；(2)匹配规则由连续 r 比特[1]和 Hamming 距离[6]演化到 Euclid 距离[7]，由固定长度演变为自适应长度[5]，更适合于对问题空间的划分；(3)形成了大量的检测器生成算法[5,7−9]，包括线性检测器产生算法、贪婪检测器产生算法、基于遗传算法的检测器生成算法等，更易于产生高效的检测器覆盖自我和非我空间。总的说来，否定选择算法得到了有效的拓展，算法的性能和解决问题的能力明显得到提升。但是，否定选择算法还存在许多亟需解决的问题，其中普遍存在的是算法在获得较高的检测率的同时也会产生较高的误报率，这种联动一直影响着算法的可靠性和可用性[4,5]。文献[10,2]指出，问题的根源并不在于算法，而是在于问题的表示空间和匹配规则。也就是说，现有算法的空间表示形式和在该空间表示形式上建立的匹配规则已成为否定选择算法的瓶颈，限制了算法的拓展。

为此，本文将状态空间从向量扩展到矩阵，提出了基于矩阵形式的否定选择算法，验证了算法的性能及参数对算法的影响，并与实值否定选择算法进行了比较和分析。

2　状态空间表示形式

状态空间描述的是问题空间，文献[1]用二进制字符串表示状态空间[1]，只能表示有限的状态，与最

2009-11-20 收到，2010-04-28 改回
国家自然科学基金(60272022)资助课题
通信作者：张雄美　zxw.ok@163.com

初的问题空间存在较大区别。为此，文献[6]提出了实值否定选择算法，用实数值形式来刻画状态空间，更能体现出样本的数据特征，同时提升了算法的速度。由于单个实数值向量只能表示部分特征信息，因此本文将多个实数值向量组合成一个矩阵来描述问题空间，可以更好地包含样本集元素内在的特征，更有利于对自我和非我的划分。

定义 1 自我集为 $\boldsymbol{S}=\{\boldsymbol{S}^1,\boldsymbol{S}^2,\cdots,\boldsymbol{S}^p\}$，$|\boldsymbol{S}|=p$，其中，自我集元素 $\boldsymbol{S}^i$ 为系统正常执行时 n 个连续状态所构成的集合，表示为一个 $m\times n$ 的矩阵。

$$\boldsymbol{S}^i=\begin{bmatrix} S_{11}^i & S_{12}^i & \cdots & S_{1n}^i \\ S_{21}^i & D_{22}^i & \cdots & S_{2n}^i \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1}^i & S_{m2}^i & \cdots & S_{mn}^i \end{bmatrix}$$

定义 2 检测集为 $\boldsymbol{D}=\{\boldsymbol{D}^1,\boldsymbol{D}^2,\cdots,\boldsymbol{D}^q\}$，$|\boldsymbol{D}|=q$，其中，检测器 $\boldsymbol{D}^i$ 是一个 $m\times n$ 的矩阵。

$$\boldsymbol{D}^i=\begin{bmatrix} D_{11}^i & D_{12}^i & \cdots & D_{1n}^i \\ D_{21}^i & D_{22}^i & \cdots & D_{2n}^i \\ \vdots & \vdots & \ddots & \vdots \\ D_{m1}^i & D_{m2}^i & \cdots & D_{mn}^i \end{bmatrix}$$

3 距离和匹配规则

在否定选择算法中，Forrest[1]等人建立了连续 r 匹配规则，即二进制字符串的 r 个连续位相同；在实值空间，则是通过计算自我集元素和检测集元素的 Euclid 距离，再根据距离是否小于域值来判断是否匹配[8]。对于实数形式的数据，需用距离来反映数据本质上的相似和差异[5]，因此，本文定义了元素匹配距离用以描述自我集元素和检测集元素的匹配程度。

定义 3 自我集元素 $\boldsymbol{S}^i$ 第 k 个行向量 $\boldsymbol{S}_k^i$ 与检测集元素 $\boldsymbol{D}^i$ 的第 k 个行向量 $\boldsymbol{D}_k^i$ 的距离 $d_k(\boldsymbol{S}^i,\boldsymbol{D}^i)$ 定义为

$$d_k(\boldsymbol{S}^i,\boldsymbol{D}^i)=\frac{\left\|\boldsymbol{S}_k^i-\boldsymbol{D}_k^i\right\|_2}{\left(\left|\boldsymbol{S}_k^i\right|^2+\left|\boldsymbol{D}_k^i\right|^2\right)^{1/2}} \tag{1}$$

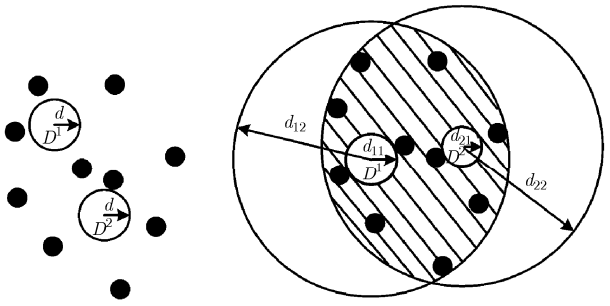
在式(1)基础上可以定义元素匹配距离来建立 $\boldsymbol{S}^i$ 和 $\boldsymbol{D}^i$ 的量度。

定义 4 自我集元素 $\boldsymbol{S}^i$ 与检测集元素 $\boldsymbol{D}^i$ 的元素匹配距离定义为行向量距离所构成向量的 2 范数，表示为

$$d(\boldsymbol{S}^i,\boldsymbol{D}^i)=\left\|d_k(\boldsymbol{S}^i,\boldsymbol{D}^i)\right\|_2 \tag{2}$$

式(2)将 $\boldsymbol{S}^i$ 和 $\boldsymbol{D}^i$ 存在的关系映射为可比较的值，为建立匹配规则提供了统一的量度。

在否定选择算法中，要进行检测器生成匹配和数据检测匹配，现有算法中的两个过程都采用同一种匹配规则。实际上，算法中生成的检测器 $\boldsymbol{D}^i$ 通过否定选择后，其与自我集之间的距离介于一个最小值 $d_{\mathrm{Min}}(\boldsymbol{D}^i,\boldsymbol{S})$ 和最大值 $d_{\mathrm{Max}}(\boldsymbol{D}^i,\boldsymbol{S})$ 之间，因而可以根据最小值和最大值之间构成的环形区域将自我和非我空间进行划分。基于此，本文建立了双向匹配规则：先用自我集半径生成检测器，再根据每个检测器和自我集之间的距离范围来检测非我。图 1(a)为现有的匹配规则，图 1(b)为先半径后范围的匹配规则，对比可见，现有的匹配规则中，检测器生成和检测监控数据都采用同一半径，限制了对自我和非我空间的分辨能力；而双向匹配规则在检测非我时利用检测器内在的距离特征，将自我集空间限制于环形区域之内，增大了对非我空间的覆盖，因而较少数量的检测器就可以较好地划分自我和非我空间。



(a) 现有的匹配规则 (b) 双向匹配规则

图 1 不同匹配规则下的自我和非我空间的划分

4 检测器生成算法

在实值空间中，由于空间较为复杂，难以建立合适的度量方法，因此目前还没有形成一个确定性的检测器生成机制[5]。最简单的检测器生成算法是随机生成检测器，当合法检测器达到一定数量就退出，由于没有考虑检测器对非我空间的覆盖，检测效果时好时坏。文献[8]用 Monte Carlo 方法评估自我和非我空间，计算所需的检测器数量，并用模拟退火方法优化检测器在非我空间的分布，缩短了计算时间，且具有更大的非我空间覆盖。文献[11]提出了基于 V-detector 的检测器生成算法，通过估计覆盖率来决定需要生成的检测器数量，同时生成半径变化的检测器以获得更好的覆盖。本文中的自我和非我用矩阵表示，状态空间比实值向量空间更为复杂，难以通过计算超空间体积等方法来计算检测器对空间的覆盖。为此，本文借鉴文献[12]中的假设检验思想，通过评估检测集对非我空间的覆盖来引

导检测器的生成，建立了如表 1 所示的基于覆盖检验的检测器生成算法，其主要思想是：在取样的总次数 N_R、所要求的覆盖比例 p 和显著性水平 α 确定的情况下，用 N_T 记录在 N_R 取样过程中，样本被检测器集合覆盖发生的次数，如果

$$\frac{N_T}{\sqrt{N_R p(1-p)}} - \sqrt{\frac{N_R p}{1-p}} \geq z_\alpha \tag{3}$$

则接受检测集对非我空间的覆盖比例达到 p 的假设，反之，则扩充检测器集合。

表 1 基于覆盖检验的检测器生成算法

Input	S(self set)，r_S (radius)，p(coverage rate)，z_α (α is significance level)， $N_{\max}$ (the maximum of detectors)
Output	D(detector set)
Step 1	$D \leftarrow \varnothing$
Step 2	$k \leftarrow 0$
Step 3	$j \leftarrow 0$
Step 4	Repeat
Step 5	$r \leftarrow$ infinite
Step 6	$d_{\text{Min}} \leftarrow \infty$ and $d_{\text{Max}} \leftarrow 0$
Step 7	overlay $\leftarrow$ false
Step 8	$x \leftarrow$ random sample from $[a_{ij}]_{m\times n}$
Step 9	$k \leftarrow k+1$
Step 10	for every D^i in $D=\{D^1, D^2, \cdots, D^q\}$
Step 11	$d_D \leftarrow$ matrix distance between x and D^i
Step 12	if $d_D \leq r(D_i)$ ($r(D_i)$ is the detector radius of D_i)
Step 13	overlay $\leftarrow$ true
Step 14	if overlay=true
Step 15	$j \leftarrow j+1$
Step 16	if $\frac{j}{\sqrt{kp(1-p)}} - \sqrt{\frac{kp}{1-p}} \geq z_\alpha$ then return D
Step 17	else
Step 18	for every S^i in $S=\{S^1, S^2, \cdots, S^p\}$
Step 19	$d_S \leftarrow$ matrix distance between x and S^i
Step 20	if $d_{\text{Min}} > d_S$ then $d_{\text{Min}} \leftarrow d_S$
Step 21	if $d_{\text{Max}} < d_S$ then $d_{\text{Max}} \leftarrow d_S$
Step 22	if $d_S - r_S \leq r$ then $r \leftarrow d - r_S$
Step 23	if $r \geq 0$ then $D \leftarrow D \cup \{< x, r, d_{\text{Min}}, d_{\text{Max}} >\}$
Step 24	until $\lvert D\rvert = N_{\max}$
Step 25	return D

若自我集元素个数为 p，最后产生的检测器数量为 N_f，则基于覆盖检验的检测器生成算法的时间复杂度为 $O(N_f p)$。

5　基于矩阵形式的否定选择算法

基于前面的讨论，本文给出了基于矩阵形式的否定选择算法，主要步骤如下：

步骤 1　收集系统 m 个子模块的样本数据，编码成 y 个实数值向量；

步骤 2　选择参数 n，将连续 n 个向量组合成 $m\times n$ 矩阵，所得的 $y-n+1$ 个矩阵构成自我集 $\boldsymbol{S}$；

步骤 3　通过基于覆盖检验的检测器生成算法最终产生 N_f 个检测器，构成检测集 $\boldsymbol{D}$，每个检测器 $\boldsymbol{D}^i$ 的检测范围为环形区域：$[d_{\text{Min}}(\boldsymbol{D}^i, \boldsymbol{S}),\ d_{\text{Max}}(\boldsymbol{D}^i, \boldsymbol{S})]$；

步骤 4　监测系统，实时获取系统的最新样本，编码形成待检测集 $\boldsymbol{UD}$，计算 $\boldsymbol{UD}$ 中每个元素 $\boldsymbol{UD}^i$ 与 $\boldsymbol{D}$ 中每个元素 $\boldsymbol{D}^i$ 的距离，如果 $d(\boldsymbol{D}^i, \boldsymbol{UD}^i) < d_{\text{Min}}(\boldsymbol{D}^i, \boldsymbol{S})$ 或者 $d(\boldsymbol{D}^i, \boldsymbol{UD}^i) > d_{\text{Max}}(\boldsymbol{D}^i, \boldsymbol{S})$，则提示系统发生异常。

由于自我集中元素个数为 $y-n+1$，算法最终生成的检测器数量为 N_f，设待检测集元素个数为 N_{UD}，则算法的时间复杂度为 $O(N_f(N_{UD}+y))$。

6　实验验证和分析

本文以标准样本 Iris 数据为例验证算法的性能。为了便于讨论，本文指定 n =2，即将两个同类样本组合在一起作为一个样本进行分析。Setosa 组成的是第 1 类样本，Virginica 组成的是第 2 类样本，Versicolor 组成的是第 3 类样本。

用 N_{TP} 和 N_{FN} 分别表示“非我样本被判定为非我”和“非我样本被判定为自我”发生的次数，用 DR 表示检测率，那么

$$\text{DR} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{4}$$

同理，用 N_{FP} 和 N_{TN} 分别表示“自我样本被判定为非我”和“自我样本被判定为自我”发生的次数，用 FA 表示误报率，那么

$$\text{FA} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \tag{5}$$

6.1 参数分析实验

指定第 1 类样本为正常，第 2 类样本和第 3 类样本作为异常，从第 1 类中取出 29 个作为自我集样本，其余 20 个作为待检测的样本，而将第 2 类样本和第 3 类样本的共 98 个样本作为待检测的非我样本。令 α =0.05，$N_{\max}$ =200，重复实验 50 次，取 DR 和 FA 的平均值。

实验 1　自我集半径对检测率和误报率的影响。令 p 为 0.9 和 0.99，r_S 取 0.01~0.2 间隔 0.01 的 20 组数据，得到的实验结果如图 2 所示。

图 2 结果表明，算法检测率随自我集半径的增大近似呈线性递增，这说明算法生成的检测器对非我空间的覆盖率随自我集空间变大而逐渐增加，同时算法也具有很低的误报率，并随自我集半径增加逐步逼近于 0。因此，本文算法获得较高检测率的同时也可以获得较低的误报率，这是对否定选择算法的一个显著的改进。

实验 2 覆盖率对检测率和误报率的影响。令 r_S 为 0.01 和 0.11，p 取 0.01~0.99 间隔 0.05 的 20 组数据，得到的实验结果如图 3 所示。

实验结果表明，检测率随覆盖率增大而逐渐增加，当 $p<0.7$ 时，误报率呈现波动上升趋势，这是因为生成的检测器数量过少，而不同检测器对非我样本数据的敏感性存在很大差别，导致检测率上下波动。而当覆盖率 $p>0.7$ 时，产生的检测器能较好地覆盖非我空间，使得检测率近似呈指数上升。误报率也随覆盖率增大而明显减少，逐渐逼近 0。这说明可以通过增加检测器来提高覆盖率，将误报率降到更低。

实验 3 检测器数量和自我集半径的关系。令 $p=0.9$，r_S 取 0.01~0.2 间隔 0.01 的 20 组数据，得到的实验结果如图 4 所示。

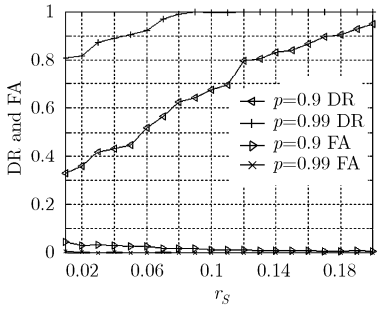


图 2 r_S 与 DR/FA 实验结果

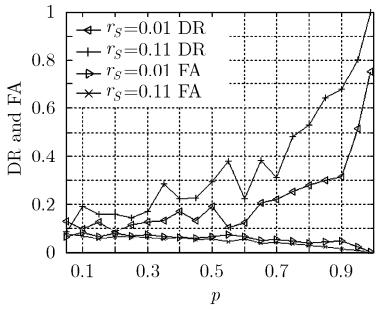


图 3 p 与 DR/FA 实验结果

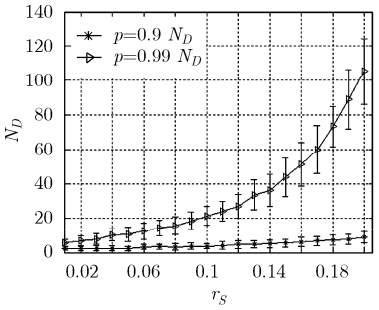


图 4 N_D 与 r_S 实验结果

实验结果表明，在通常的覆盖率下，检测器数量和自我集半径近似呈线性关系；但如果要求非常大的覆盖率，需生成的检测器数量和自我集半径近似呈指数关系。这是因为矩阵空间远比向量空间复杂，要达到更好的空间覆盖需大量地增加检测器的数量。

6.2 算法对比实验

现有的实值否定选择算法较多，其中常用的有连续位匹配实值否定选择算法、多层学习实值否定选择算法和 V-detector 实值否定选择算法[5,11,13]。其中，V-detector 实值否定选择算法比其他两种算法具有更好的检测效果[11]，因此本文选取该算法进行对比实验。取 $\alpha=0.05$，$N_{\max}=400$，$p=0.97$，$r_S=0.05$，得到的实验结果如表 2 所示。

表 2 的数据表明，在进行的 6 次对比实验中，本文算法的检测率都要高于 V-detector 实值否定选择算法，并且误报率远低于 V-detector 实值否定选择算法的。这就表明本文算法的性能明显优于 V-detector 实值否定选择算法。同时，本文统计了上述实验过程中产生的检测器数量，如表 3 所示。

表 2 算法对比实验结果

自我集	算法	检测率(%)	误报率(%)
Setosa(100%)	V-detector 实值否定选择算法	99.98	0
	基于矩阵形式的否定选择算法	99.99	0
Setosa(50%)	V-detector 实值否定选择算法	99.97	1.32
	基于矩阵形式的否定选择算法	99.98	0.01
Versicolor(100%)	V-detector 实值否定选择算法	85.95	0
	基于矩阵形式的否定选择算法	96.8	0
Versicolor(50%)	V-detector 实值否定选择算法	88.3	8.42
	基于矩阵形式的否定选择算法	88.5	0.02
Virginica(100%)	V-detector 实值否定选择算法	81.87	0
	基于矩阵形式的否定选择算法	91.4	0
Virginica(50%)	V-detector 实值否定选择算法	93.58	13.18
	基于矩阵形式的否定选择算法	93.73	0.012

表 3 检测器数量对比

自我集	算法	平均值	最大值	最小值	标准方差
Setosa(100%)	V-detector 实值否定选择算法	20	42	5	7.87
	基于矩阵形式的否定选择算法	8.5	13	5	2.01
Setosa(50%)	V-detector 实值否定选择算法	16.44	33	5	5.63
	基于矩阵形式的否定选择算法	7.3	10	5	1.76
Versicolor(100%)	V-detector 实值否定选择算法	153.24	255	72	38.8
	基于矩阵形式的否定选择算法	131.1	152	114	12.73
Versicolor(50%)	V-detector 实值否定选择算法	110.08	184	60	22.61
	基于矩阵形式的否定选择算法	116.3	128	90	13.11
Virginica(100%)	V-detector 实值否定选择算法	218.36	443	78	66.11
	基于矩阵形式的否定选择算法	130	143	106	11.25
Virginica(50%)	V-detector 实值否定选择算法	108.12	203	46	30.74
	基于矩阵形式的否定选择算法	98.4	113	84	11.12

数据表明本文算法达到相同的检测效果所需生成的检测器数量小于 V-detector 实值否定选择算法所生成的检测器数量。而且本文算法检测器数量的标准方差更小，说明其产生的检测器质量较高，对自我和非我空间的覆盖率很高，且能较为均匀地覆盖自我和非我空间。

7 结束语

否定选择算法自 Forrest 等人[1]提出以来得到了不断的改进和完善，已成为人工免疫系统的核心算法之一。但由于算法空间表示形式和匹配规则的限制，否定选择算法还存在许多急需解决的问题。本文在现有的否定选择算法基础上，从 3 个方面对否定选择算法进行了改进：(1)将状态空间从向量扩展到矩阵；(2)将匹配规则从单一扩展到双向；(3)根据状态空间特征建立了基于覆盖检验的检测器生成算法。由此形成的基于矩阵形式的否定选择算法比现有否定选择算法更好的性能，不仅解决了检测率和误报率联动的问题，而且能获得更高质量的检测器。下一步工作是对本文算法进行具体的应用研究，并进行更为深入的理论分析。

参 考 文 献

[1] Forrest S, Perelson A S, Allen L, and Cherukuri R. Self-nonself discrimination in a computer[C]. Proceedings of the IEEE Symposium on Research in Security and Privacy, Los Alamos, CA, May 16–18, 1994: 202–212.

[2] Ji Zhou and Dasgupta D. Applicability issues of the real-valued negative selection algorithms[C]. Proceedings of the 2006 Conference on Genetic and Evolutionary Computation Conference, Seattle, Washington, USA, July 8–12, 2006: 111–118.

[3] Sarafijanovic S, Perez S, and Le Boudec J Y. Resolving FP-TP conflicting in digest-based collaborative spam detection by use of negative selection algorithm[C]. Proceedings of the Fifth Conference on Email and AntiSpam, Mountain View, California, USA, Aug. 21–22, 2008.

[4] Yi Zhao-xiang, Mu Xiao-dong, Zhang Li, and Zhao Peng. A matrix negative selection algorithm for anomaly detection[C]. Proceedings of 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, June 1–6, 2008: 978–983.

[5] Ji Zhou and Dasgupta D. Revisiting negative selection algorithms[J]. *Evolutionary Computation*, 2007, 15(2): 223–251.

[6] Gonzalez F, Dasgupta D, and Kozma R. Combining negative selection and classification techniques for anomaly detection[C]. Proceedings of 2002 IEEE Congress on Evolutionary Computation, Honolulu, May 12–17, 2002: 705–710.

[7] Gonzalez F, Dasgupta D, and Gomez J. The effect of binary matching rules in negative selection[C]. Proceedings of the Conference on Genetic and Evolutionary Computation Conference(GECCO'2003), Chicago, USA, July 12–16, 2003: 195–206.

[8] Gonzalez F, Dasgupta D, and Nino L F. A randomized real-valued negative selection algorithm[C]. Proceedings of the Second International Conference on Artificial Immune Systems, Edinburgh, UK, September 1–3, 2003: 261–272.

[9] Gao X Z, Ovaska S J, and Wang X. A GA-based negative selection algorithm[J]. *International Journal of Innovative Computing, Information and Control*, 2008, 4(4): 971–979.

[10] Balthrop J, Forrest S, and Glickman M R. Revisiting LISYS: parameters and normal behavior[C]. Proceedings of the

Conference on Genetic and Evolutionary Computation (GECCO'2002), New York City, USA, July 9–13, 2002: 1045–1050.

[11] Ji Zhou and Dasgupta D. Augmented negative selection algorithm with variable-coverage detectors[C]. Proceedings of the Congress on Evolutionary Computation, San Diego, CA, USA. July 6–9, 2004: 1081–1088.

[12] Ji Zhou and Dasgupta D. Estimating the detector coverage in a negative selection algorithm[C]. Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO'2005), Washington DC, USA, June 25–29, 2005: 88–97.

[13] Ji Zhou and Dasgupta D. V-detector: an efficient negative selection algorithm with "probably adequate" detector coverage[J]. *Information Sciences*, 2009, 179(10): 1390–1406.

张雄美： 女，1983 年生，博士生，研究方向为免疫算法、SAR 图像处理.

易昭湘： 男，1980 年生，讲师，研究方向为人工免疫、故障诊断.

宋建社： 男，1954 年生，教授，博士生导师，研究方向为信号与信息处理、智能算法.