论 文

一种基于自体集层次聚类的否定选择算法

陈文^①, 李涛^{①*}, 刘晓洁^①, 张冰^②^① 四川大学计算机学院, 成都 610065^② 国家计算机网络与信息安全管理中心, 北京 100029

* 通信作者. E-mail: litao@scu.edu.cn

收稿日期: 2012-01-28; 接受日期: 2012-03-18

国家自然科学基金 (批准号: 60873246)、教育部重大项目培育基金 (批准号: 708075) 和教育部博士点基金 (批准号: 20070610032) 资助项目

摘要 否定选择算法是用于产生人工免疫检测器的重要算法, 然而传统的否定选择过程需要将随机生成的候选检测器与全部自体数据进行匹配以排除识别了自体的无效检测器, 该匹配过程导致检测器的生成效率过低, 极大地限制了免疫算法的应用. 为此, 文中提出了一种基于自体集层次聚类的否定选择算法 CB-RNSA. 算法首先对自体数据进行层次聚类预处理, 然后用聚类中心取代自体数据点与候选检测器进行匹配, 以减少距离计算代价. 在生成检测器的过程中, 候选检测器被限定在非自体空间的低覆盖率区域内, 以降低检测器冗余. 对检测器的非自体空间覆盖率进行了概率分析, 给出了中止生成检测器的条件, 该条件较传统的预设检测器数量的中止条件更为合理. 理论分析表明 CB-RNSA 的时间复杂度与自体集规模无关, 从而解决了经典的否定选择算法的时间复杂度随自体数量呈指数增长这一难题, 极大地提高了大自体样本空间下的检测器生成效率. 对比实验结果表明: 在相同的实验数据集与期望覆盖率下, CB-RNSA 的检测率比经典的 RNSA 与 V-detector 算法分别提高了 12.3% 与 7.4%, 误警率分别降低了 8.5% 与 4.9%, 产生检测器的时间代价分别降低了 67.6% 和 75.7%.

关键词 人工免疫系统 否定选择算法 人工免疫检测器 聚类 覆盖率

1 引言

否定选择算法 (negative selection algorithm, NSA) 是人工免疫理论中的一种重要的检测器生成算法. NSA 最早由 Forrest 等提出^[1], 算法通过模拟生物体内的 T 细胞在胸腺中的成熟过程以生成消除了免疫自反应 (self reaction) 的成熟检测器, 用于执行数据分类、模式识别和入侵检测等任务^[1,2].

早期的否定选择算法^[1], 本文称之为 NNSA (native negative selection algorithm) 以二进制串编码抗体 (检测器) 和抗原 (样本特征), 并使用 r -位连续匹配算法计算检测器与抗原的匹配程度. 文献 [2~5] 指出 NNSA 存在检测器生成效率过低的问题, 随机生成的候选检测器通过否定选择成为成熟检测器的概率 $P = (1 - P_m)^{N_s}$, 其中 P_m 为任意抗体与抗原匹配的概率, N_s 为自体元素个数, 当自体数量渐增时, P 趋近于 0; 此外, 为使检测器满足指定的检测失效率 $P_f \approx e^{-P_m|D|}$ 则需要的候选检测

引用格式: 陈文, 李涛, 刘晓洁, 等. 一种基于自体集层次聚类的否定选择算法. 中国科学: 信息科学, 2013, 43: 611-625, doi: 10.1360/112010-1145

器数量为 $N_0 = \frac{-\ln(P_f)}{P_m(1-P_m)^{N_s}}$, N_0 与自体集的大小 N_s 呈指数关系, 算法的时间复杂度为 $O(N_0 \cdot N_s)^{[1]}$, 检测器的生成代价过大是限制 NNSA 实际应用的主要问题^[3].

由于实际应用中的大量问题易于在实值空间进行定义和研究, 文献 [6, 7] 提出了实值否定选择算法 (real-valued negative selection algorithm, RNSA), 将检测器与抗原的各个属性值规范化到 $[0, 1]$ 实数范围, 在 d 维实值空间的单位超立方体 (hypercube) $[0, 1]^d$ 内采用 Minkowski 距离计算抗体与抗原的匹配程度. 文献 [8, 9] 提出了可变半径的实值否定选择算法 V-detector, 通过计算检测器与最邻近自体间的距离动态决定检测半径, V-detector 在仿真实验中取得了较好的检测效果.

文献 [10, 11] 指出, 在基于距离计算的模式识别算法中, 距离计算是其最主要的时间开销. 然而无论是原始的 NNSA 算法还是 RNSA 算法及其改进 V-detector 都未考虑减少距离计算代价以提高算法效率: 为训练未成熟的检测器, 否定选择过程必须计算随机生成的候选检测器与全部自体数据间的距离, 导致算法的时间开销巨大. Stibor 在文献 [3] 中指出, 由于成熟检测器的生成效率过低, 否定选择算法的执行时间已达到难以承受的地步.

本文提出一种基于自体集层次聚类的实值否定选择算法 (real-valued negative selection algorithm based on hierarchical clustering of self set, CB-RNSA): 首先对自体集进行层次聚类处理; 其次在每聚类层指定的随机值范围内生成候选检测器, 以避免在高覆盖率区域内造成检测器冗余; 候选检测器与聚类中心而非自体集进行匹配以减少距离计算的时间代价, 提高检测器生成效率; 算法以期望的检测器非自体空间覆盖率作为中止条件.

2 NSA 的基本定义及问题

免疫是机体依靠抗体细胞识别自体 “(self)” 和 “非自体 (nonself)” 排除抗原性异物以维持机体生理平衡和稳定的过程^[12]. 人工免疫理论中, 抗体被定义为检测器 (detector). 检测器用于对非自体抗原进行检测与识别, 其质量决定了免疫检测结果的准确性和有效性^[6~9]. 然而, 随机生成的候选检测器可能识别自体空间中的抗原而发生免疫自反应 (self reaction)^[13, 14]. 受生物体内新生成的免疫细胞经历自体耐受过程而成熟的启发, Forrest 等^[1] 提出了否定选择算法以排除发生自反应的检测器. 否定选择算法中的基本概念定义如下:

定义 1 所有来自特征空间的样本特征串组成抗原集合 $U = \{g \mid g = (f_1, f_2, \dots, f_n), f_i \in [0, 1]\}$, n 为数据维数, f_i 代表第 i 个样本属性规范化后的值.

定义 2 自体集 $\text{Self} \subseteq U$ 代表正常样本特征串, 自体半径 $r_s \in \mathbb{R}^+$ 代表自体的变化范围; 非自体集 $\text{Nonself} \subseteq U$ 代表异常样本特征串, 且 $\text{Self} \cup \text{Nonself} = U$, $\text{Self} \cap \text{Nonself} = \emptyset$.

定义 3 检测器 $d\langle c, r \rangle$ 用于识别非自体抗原, 其中 $c \in \text{Nonself}$ 代表检测器位置向量, $r \in \mathbb{R}^+$ 代表检测半径, 与 d 的距离小于 r 的抗原被识别为非自体元素.

定理 1 若检测器 d 与任意自体 s 的距离满足: $\text{dis}(d, s) > r + r_s$, 则该检测器在其识别范围内避免了免疫自反应.

证明 设存在 $d \in D$, $d = \langle c, r \rangle$, $\forall s \in \text{Self}$, 且 d 与 s 间的距离 $\text{dis}(d, s) > r + r_s$. 在 s 的自体边缘存在点 p 处于检测器位置向量 c 与自体中心 s 的连线上, 则 $\text{dis}(d, p) = \text{dis}(d, s) - r_s$. 根据假设可得 $\text{dis}(d, p) > r + r_s - r_s = r$, 由定义 3 可知 p 将被检测器 d 识别为自体元素. 由于 s 的自体范围内任意元素与 d 的距离均大于 $\text{dis}(d, p)$, 因此检测器 d 在其识别范围内避免了免疫自反应. 证毕.

不失一般性, 否定选择算法定义如下:

表 1 CB-RNSA 算法

Table 1 CB-RNSA

CB-RNSA (S, P_{exp}, D)	S : self set	P_{exp} : expected coverage	D : detector set	C : cluster center set
Step1	Call procedure Hierarchical-Cluster(S, C) to generate self clusters. The cluster centers of the i th level are stored in C_i .			
Step2	Initialize the detector set D as null set: $D \leftarrow \emptyset, i = 1$.			
Step3	Call procedure Detector-Create (C_i, P_{exp}) to generate detector set D_i in the i th level.			
Step4	Incorporate T into set D : $D = D \cup D_i$, increase i , goto step 3.			

目标: $|D| = \text{num}$,

条件: $\forall d \in D, s \in \text{Self} : \text{dis}(d, s) > r + r_s$,

其中: $d = \langle c, r \rangle, c = (x_1, x_2, \dots, x_n) \in X \subseteq \mathbb{R}^n, X = \{(x_1, x_2, \dots, x_n) | 0 \leq x_i \leq 1\}$.

算法的目标是生成指定规模 num 的检测器集 D , 并满足任意检测器与自体元素间的距离均大于检测阈值. 以上定义中, $|\cdot|$ 代表集合中的元素个数, X 代表 n 维特征空间, c 为检测器位置向量, r 为检测半径, r_s 为自体半径.

从以上定义可知, 任意候选检测器 d 均需与全部自体元素进行距离值 $\text{dis}(d, s)$ 的计算, 如引言部分所述: 当自体元素个数 N_s 递增时, d 通过否定选择 $\text{dis}(d, s) > r + r_s$ 的概率 P 趋近于 0, 算法需要的候选检测器数量 N_0 激增, 从而导致否定选择过程的效率低下; 另一方面, 如文献 [3,5] 所述, 由于检测器间存在重叠的检测区域, 因此难以准确估计算法需要的检测器数量 num . 上述问题极大地限制了 NSA 的实际应用.

3 CB-RNSA 算法实现策略

算法的基本思想是对自体集进行层次聚类; 在每一聚类层, 将随机值范围内生成的候选检测器与该层的聚类中心进行距离计算与相似度比较, 排除覆盖聚类范围的候选检测器; 算法以期望的检测器非自体空间覆盖率作为中止检测器生成过程的条件. CB-RNSA 算法的否定选择过程定义如下:

目标: $P_{\text{cov}} \geq P_{\text{exp}}$,

条件: $\forall d \in D, c \in C : \text{dis}(d, c) > r + r_c$,

其中: $d = \langle c, r \rangle, c = (x_1, x_2, \dots, x_n) \in X, X = \{(x_1, x_2, \dots, x_n) | l_i \leq x_i \leq u_i\}, l = (l_1, l_2, \dots, l_n), u = (u_1, u_2, \dots, u_n)$.

算法在避免检测器覆盖自体聚类范围的前提下, 生成检测器集 D 并满足 $P_{\text{cov}} \geq P_{\text{exp}}$, 其中 P_{cov} 为检测器非自体空间覆盖率, P_{exp} 为期望覆盖率, C 为聚类中心集合, r_c 为聚类半径, X 为 n 维特征空间, l, u 为检测器位置向量的随机值范围. CB-RNSA 算法的描述如表 1 所示.

其中 Hierarchical-Cluster(S, C) 执行自体集层次聚类, 聚类结果存放在集合 C 中. Detector-Create (C_i, P_{exp}) 用于产生第 i 层聚类对应的检测器. Hierarchical-Cluster 与 Detector-Create 的详细描述见 3.1 小节与 3.3 小节.

3.1 自体集层次聚类

自体集层次聚类旨在将相似的自体数据归入同一聚类中, 用聚类中心代表聚类成员与候选检测器

表 2 层次聚类算法

Table 2 Hierarchical-Cluster

Hierarchical-Cluster(S, C)	S : self set	C : cluster center set
Step1	Set s as the single cluster of the first level, cluster level $l=1$, cluster radius $r=\sqrt{n}$.	
Step2	For each cluster $_i$ in level l , find new cluster center c_k by (2), update matrix M , $M_{kj} = \text{dis}(c_k, c_j)$. If there exists no such center, go to step 4.	
Step3	Based on the clustering rule of theorem 2, group the self elements which are closest to c_k with distances less than r into the k th cluster, increase k , go to step 2.	
Step4	Set $r = r/2$, if r is less than self radius r_s then exit, otherwise reset M , increase l , go to step 2.	

进行匹配以减少距离计算次数. 在自上而下的层次聚类过程中, 聚类半径逐层减半, 下层聚类范围将更加收敛从而暴露出更多的非自体空间, 这些空间将被下层新产生的检测器所覆盖.

对于 n 维特征空间, 置第 1 层的聚类半径 $r_1 = \sqrt{n}$ 以包含单位特征空间中的全部自体数据; 为了获得第 $i+1$ 层聚类, 对第 i 层各聚类中的自体数据按照半径 $r_{i+1} = r_i/2$ 再次执行聚类, 且满足 (1) 式: 聚类成员到聚类中心的距离之和最小. 执行层次聚类过程直到聚类半径小于自体半径 r_s .

$$\min \sum_{i=1}^m \sum_{s \in \text{cluster}_i} \text{dis}(c_i, s). \quad (1)$$

聚类中心的选取标准如 (2) 式所示, 以尚未划分聚类的自体数据中到其余自体的距离之和最小为标准进行选取, 直到不存在可用的聚类中心为止.

$$c_k = \arg \min_i \sum_{\text{clustered}(s_j)=\text{false}} \text{dis}(s_i, s_j). \quad (2)$$

为提高聚类效率, 层次聚类过程采用三角不等式原理排除冗余的距离计算, 若确定自体 s 距聚类中心 c_i 的距离最近, 则不再计算 s 与其余聚类中心间的距离, s 被直接划入第 i 个聚类.

定理 2 对于自体点 s , 聚类中心 c_i 和 c_j , 若 $2\text{dis}(s, c_i) \leq \text{dis}(c_i, c_j)$, 那么 $\text{dis}(s, c_i) \leq \text{dis}(s, c_j)$.

证明 由三角不等式原理^[15]可知 $\text{dis}(c_i, c_j) \leq \text{dis}(s, c_i) + \text{dis}(s, c_j)$, 因此 $\text{dis}(c_i, c_j) - \text{dis}(s, c_i) \leq \text{dis}(s, c_j)$, 由假设 $2\text{dis}(s, c_i) \leq \text{dis}(c_i, c_j)$, 则 $2\text{dis}(s, c_i) - \text{dis}(s, c_i) \leq \text{dis}(c_i, c_j) - \text{dis}(s, c_i) \leq \text{dis}(s, c_j)$, 因此 $\text{dis}(s, c_i) \leq \text{dis}(s, c_j)$ 成立. 证毕.

根据定理 2, 当已知自体点 s 与聚类中心 c_i 的距离值的 2 倍小于 c_i 与另一聚类中心 c_j 间的距离时, 可以避免 $\text{dis}(s, c_j)$ 的计算. 计算出 $\text{dis}(s, c_i)$ 后, 扫描聚类中心间的距离矩阵 M 的第 i 行, 若 $2\text{dis}(s, c_i) \leq \min M_{ij}$, 可知 s 与 c_i 距离最近, s 被直接划入第 i 个聚类, 从而避免了 s 与其余聚类中心间的距离计算代价. 层次聚类算法如表 2 所示.

3.2 检测器生成策略

3.2.1 非自体空间覆盖率分析

传统的否定选择算法以检测器数量达到预设值作为中止检测器生成过程的条件^[4]. 然而, 准确估计覆盖非自体空间所需的检测器数量较为困难; 另一方面由于检测器间存在重叠的覆盖区域, 即使产生所需数量的检测器, 也无法保证达到期望的非自体空间覆盖率. 因此预设检测器数量的中止条件缺

乏准确性和有效性^[16~18]. 本文对检测器非自体空间覆盖率进行了概率分析, 直接以期望的检测器非自体空间覆盖率作为中止生成检测器的条件.

检测器非自体空间覆盖率 p 等于被检测器覆盖的非自体空间体积 V_{covered} 占全部的非自体空间体积 V_{nonself} 的比例^[8], 如 (3) 式所示:

$$P = \frac{V_{\text{covered}}}{V_{\text{nonself}}} = \frac{\int_{\text{covered}} dx}{\int_{\text{nonself}} dy}. \quad (3)$$

由于直接计算 (3) 式不可行, 本文通过固定规模的连续非自体采样对 p 进行估计: 对于检测器集合 D , 任意非自体抗原 X 被检测器覆盖的概率服从二项分布^[9]: $p\{X=1, X \text{ 被检测器覆盖}\}=p$, $p\{X=0, X \text{ 未被检测器覆盖}\}=1-p$.

定理 3 存在正整数 n , 当连续采样的非自体样本数量大于 n 时, 检测器的非自体空间覆盖率在置信度 $1-\alpha$ 下的置信下限 $P_{\min}$ 等于 $\frac{1}{2a}(-b-\sqrt{b^2-4ac})$, 其中 $a=n+Z_{\alpha/2}^2$, $b=-(2n\bar{x}+Z_{\alpha/2}^2)$, $c=n\bar{x}^2$, $\bar{x}$ 为样本均值, $Z_{\alpha/2}$ 为标准正态分布分位点.

证明 根据中心极限定理, 当连续的非自体样本数量 n 较大时, 下式中 $\hat{P}$ 近似服从标准正态分布:

$$\hat{P} = \frac{\sum_{i=1}^n x_i - np}{\sqrt{np(1-p)}} = \frac{n\bar{x} - np}{\sqrt{np(1-p)}}, \quad (4)$$

其中, $\bar{x}$ 为样本均值, n 为非自体样本数量, 于是

$$P = \left\{ -Z_{\alpha/2} \leq \frac{n\bar{x} - np}{\sqrt{np(1-p)}} \leq Z_{\alpha/2} \right\} \approx 1 - \alpha, \quad (5)$$

其中, $Z_{\alpha/2}$ 为标准正态分布的分位点, $1-\alpha$ 为置信度. 求解 (5) 式, 得到非自体空间覆盖率 p 的置信下限:

$$P_{\min} = \frac{1}{2a}(-b - \sqrt{b^2 - 4ac}), \quad (6)$$

其中, $a=n+Z_{\alpha/2}^2$, $b=-(2n\bar{x}+Z_{\alpha/2}^2)$, $c=n\bar{x}^2$.

根据定理 3, 当连续采样的非自体样本数量达到预设规模 n 且非自体空间覆盖率的估计值下限 $P_{\min}$ 大于期望覆盖率 P_{exp} 时, 停止随机采样过程; 否则以取得的非自体样本点作为位置向量, 生成新的检测器, 并开始下一轮采样过程.

3.2.2 候选检测器随机值范围

随机值范围 (random value ranges, RVR) 设置了候选检测器的位置向量在各维度的取值区间. RVR 定义为 d 维超立方体: $\text{RVR} = \{\text{hypercube} \mid \text{hypercube} = ([\text{low}_1, \text{low}_2, \dots, \text{low}_d], [\text{high}_1, \text{high}_2, \dots, \text{high}_d])\}$, high_i 和 low_i 分别代表检测器位置向量第 i 维的取值上下限.

层次聚类的 l 层第 i 个聚类对应的 RVR 为: $([c_{i1}-r_l, c_{i2}-r_l, \dots, c_{id}-r_l], [c_{i1}+r_l, c_{i2}+r_l, \dots, c_{id}+r_l])$, 其中 c_i 为聚类中心, r_l 为聚类半径. 由于在第 $l+1$ 层的检测器生成过程中, 第 l 层随机值范围以外的非自体空间已被前 l 层生成的检测器覆盖, 因此新生成的候选检测器被限定在第 l 层的随机值范围以内, 使检测器集中在覆盖率较低的区域, 以提高检测器非自体空间覆盖率并减少检测器冗余.

图 1 以 setosa 数据作为自体集. sepal L 和 sepal W 作为抗原的第 1 维和第 2 维的属性值, 图中方框代表随机值范围, 虚线圆代表聚类范围, 实线圆代表生成的检测器

Figure 1 The setosa data composite self set. Sepal L and sepal W are the first and second data dimension, respectively, the rectangles are random value ranges, the dashed cycles are clusters and the solid cycles are generated detectors. In (a) the RVR for level “0” is $[0, 1]^2$; in (b) the shadow is the covered non-self space in the first cluster level. (a) The first cluster result and generated detectors; (b) the second cluster result and generated detectors

表 3 检测器生成算法

Table 3 Detector-Create

Detector-Create (C, P_{exp})	C : cluster center set	P_{exp} : the expected coverage
Step1	Sample element X from the random value ranges.	
Step2	Calculate the distances between X and the cluster centers, if X belongs to any self clusters, then drop X and go to step 1, otherwise, increase the non-self sample number t .	
Step3	If X is covered by any detector in D , then increase the cover number m .	
Step4	Generate new detector $d(x, r)$; put d into the temporary set T_d .	
Step5	If the non-self sample number equals predefined number n , P_{min} is estimated by Theorem 3. If P_{min} is bigger than P_{exp} , then stop the generation process; otherwise, incorporate T_d into D , reset t , m and T_d , go to step 1.	

Iris data set 数据集是 University of California Irvine(UCI) 发布的经典的机器学习数据集, 被国内外研究者广泛应用于模式识别、数据挖掘等领域. 以 Iris data set 中类别为 “setosa” 的数据记录作为自体抗原, 图 1 展示了第 1 层和第 2 层聚类过程中的随机值范围及检测器的生成情况. 从图 1(b) 中可以看出, 第 1 层聚类的随机值范围外的非自体空间已被检测器充分覆盖, 因此第 2 层的候选检测器被限定在第 1 层聚类的随机值范围内生成, 以避免检测器冗余覆盖非自体空间.

3.3 检测器生成算法

检测器的生成过程包括非自体取样与检测器非自体空间覆盖率估计两个主要步骤: 在随机值范围内采样非自体点, 排除落入聚类范围的样本, 保留余下的样本以生成新的检测器; 当连续采样的非自体样本数量达到预设值 n 时, 按定理 3 对检测器非自体空间覆盖率进行估计, 以决定是否终止检测器的生成过程. 检测器生成算法如表 3 所示.

表 4 检测生成过程时间复杂度

Table 4 The time complexity of detector generation procedure

Algorithm	Time Complexity
NNSA	$O(-\frac{\ln P_f \cdot N_s}{P_m \cdot (1-P_m)^{N_s}})^{[1]}$
RNSA	$O(\frac{ D \cdot N_s}{(1-P)^{N_s}})^{[6]}$
V-detector	$O(\frac{ D \cdot N_s}{(1-P)^{N_s}})^{[9]}$
CB-RNSA	$O(N_0 \cdot C + (1-p) \cdot N_0 \cdot (D + C))$

相对于文献 [6 ~ 9] 在 $[0,1]^d$ 空间内随机生成全部的候选检测器, 步骤 1 中的候选检测器 (样本点) 被限定在低覆盖率的随机值范围内, 降低了检测器冗余.

步骤 2 用聚类中心与样本点 X 进行距离计算以排除落入聚类范围的样本, 时间复杂度为 $O(|C|)$, 而传统的否定选择算法中该步骤的时间复杂度为 $O(|S|)$. 由于聚类中心的数量 $|C|$ 远小于自体元素的数量 $|S|$, 因此算法对样本点的选择效率得到了提高.

步骤 3 累积样本点被检测器覆盖的次数 m , m 被用于步骤 5 中检测器覆盖率下限的计算过程. 根据定理 1, 步骤 4 中设置检测器的半径 r 等于 X 与最邻近聚类中心的距离与聚类半径的差值, 以避免检测器发生自反应.

步骤 5 通过定理 3 计算检测器的覆盖率下限 $P_{\min}$, 若 $P_{\min}$ 大于期望覆盖率 $P_{\exp}$, 则中止检测器生成过程, 否则将步骤 4 中得到的新检测器并入 D , 并开始新一轮采样.

由于检测器的非自体空间覆盖率随着新检测器的加入而发生改变, 因此在步骤 4, 5 中, 先保持检测器集合 D 不变, 当非自体样本数量达到预设值 n 且 $P_{\min}$ 小于 $P_{\exp}$ 时, 才将 T_d 中的检测器加入 D .

3.4 检测器生成代价分析

3.4.1 检测器生成过程的时间复杂度

定理 4 CB-RNSA 检测器生成过程的时间复杂度为 $O(N_0 \cdot |C| + (1-p) \cdot N_0 \cdot (|D| + |C|))$, 其中 N_0 为候选检测器数量, $|C|$ 为聚类数量, $|D|$ 为检测器数量, p 为检测器平均自反应率.

证明 CB-RNSA 通过循环调用 Detector-Create 在各聚类层生成检测器. Detector-Create 在步骤 2 采用聚类中心与样本点进行比较, 时间复杂度为 $O(|C_i|)$, $|C_i|$ 为第 i 层的聚类数量; 步骤 3 判断样本点是否被检测器覆盖的时间复杂度为 $O(|D|)$; 步骤 4 计算新生成的检测器半径需要计算样本点与聚类中心间的最邻近距离, 时间复杂度为 $O(|C_i|)$. 记第 i 层的候选检测器数量为 N_i , 则在步骤 2 未被排除的样本点数量为 $(1-P_i) \cdot N_i$, $P_i = \frac{|C_i| \cdot \prod_{j=1}^{d/2}}{4^d \cdot \Gamma(d/2+1)}$ 为检测器落入聚类范围的概率, 见 3.4.2 小节. 因此, 第 i 层生成检测器的总时间复杂度为 $O(N_i \cdot |C_i| + (1-P_i) \cdot N_i \cdot (|D| + |C_i|))$. 由 3.1 小节层次聚类算法可知层次聚类树的高度 $k = \lfloor \log_2 \sqrt{n}/r_s \rfloor$, 因此 CB-RNSA 循环调用 Detector-Create 生成检测器的总时间复杂度为 $O(\sum_{i=1}^k (N_i \cdot |C_i| + (1-P_i) \cdot N_i \cdot (|D| + |C_i|)))$, 候选检测器总数 $N_0 = \sum_{i=1}^k N_i \approx |D|/(1-P)$. 化简后的算法时间复杂度等于: $O(N_0 \cdot |C| + (1-p) \cdot N_0 \cdot (|D| + |C|))$, 其中 $P = |C| \cdot \prod_{i=1}^{d/2} [4^d \cdot \Gamma(d/2+1)]$ 为检测器平均自反应率, $|C|$ 为各聚类层的平均聚类数量. 证毕.

NNSA^[1]、RNSA^[6] 和 V-detector^[9] 是目前 3 种主要的检测器生成算法, 被广泛应用于基于人工免疫系统的模式识别、入侵检测和疾病诊断等领域^[1,2,4]. 表 4 列出了这 3 种否定选择算法和 CB-RNSA 的检测器生成过程的时间复杂度, 其中 P_m 为检测器识别任意抗原的概率, N_s 为自体元素个数, P_f 代

表检测失效率. 从表 4 可知, 传统的否定选择算法的时间复杂度与 N_s 呈指数关系, 当自体元素增多时, 算法的时间开销将迅速增加, 甚至已达到难以承受的地步 [3]. CB-RNSA 算法的时间复杂度与聚类数量 $|C|$, 检测器集合大小 $|D|$ 以及检测器的自反应率 p 有关, 而与自体数量 N_s 无关, 由此消除了自体集的规模增长对 CB-RNSA 算法的时间复杂度的影响. 因此, CB-RNSA 算法能够有效提高大自体样本空间下的检测器生成效率.

3.4.2 候选检测器数量

文献 [16, 17] 提出用否定选择算法产生的全部候选检测器的数量来衡量检测器生成代价. 假设达到期望的检测器非自体空间覆盖率, 至少需要 $|D|$ 个成熟检测器, 则所需候选检测器的数量 $N_0^{[1]}$:

$$N_0 = \frac{|D|}{(1 - P)^{\text{num}}}. \quad (7)$$

由 (7) 式可知, 候选检测器发生自反应的概率 p 和自体数量 num 越大, 生成 $|D|$ 个成熟检测器所需的候选检测器数量 N_0 越大, 相应的检测器生成代价也越大.

对于 RNSA^[6] 和 V-detector^[9] 算法, p 等于检测器覆盖自体范围的概率, num 等于自体数量; 对于 CB-RNSA 算法, p 等于检测器覆盖聚类范围的概率, num 等于聚类数量. 为简化讨论, 假设自体范围或聚类间不存在重叠.

对于 RNSA 和 V-detector 算法, 由于在 $[0, 1]^d$ 空间内随机生成候选检测器, 因此检测器发生自反应的概率等于代表自体的 n 个半径为 r_s 的 d 维超球体的体积占单位超立方体体积的比例:

$$P_1 = \frac{n \cdot V_{\text{self}}}{V_{\text{cube}}} = \frac{n \cdot r_s^d \cdot \prod^{d/2}}{\Gamma(d/2 + 1)}, \quad (8)$$

其中, n 为自体数量, $V_{\text{self}} = \frac{r_s^d \cdot \prod^{d/2}}{\Gamma(d/2 + 1)}$ 为单个自体体积, $V_{\text{cube}} = 1$ 为单位超立方体体积.

在 CB-RNSA 算法中, 候选检测器被限定在棱长为 $4r$ 的超立方体内, r 为聚类半径, 因此检测器发生自反应的概率等于代表聚类的 m 个半径为 r 的 d 维超球体的体积占超立方体体积的比例:

$$P_2 = \frac{m \cdot V_{\text{clu}}}{V_{\text{random_cube}}} = \frac{m \cdot r^d \cdot \prod^{d/2}}{(4 \cdot r)^d \cdot \Gamma(d/2 + 1)}, \quad (9)$$

其中, m 为聚类数量, $V_{\text{clu}} = \frac{r^d \cdot \prod^{d/2}}{\Gamma(d/2 + 1)}$ 为单个聚类超球体的体积, $V_{\text{random_cube}} = (4 \cdot r)^d$ 为超立方体的体积.

为比较算法 RNSA, V-detector 与 CB-RNSA 的检测器生成代价, 由 (8), (9) 式得到关于检测器自反应概率 p 的比例系数 ρ :

$$\rho = \frac{P_1}{P_2} = \frac{\frac{n \cdot r_s^d \cdot \prod^{d/2}}{\Gamma(d/2 + 1)}}{\frac{m \cdot r^d \cdot \prod^{d/2}}{(4 \cdot r)^d \cdot \Gamma(d/2 + 1)}} = \frac{n}{m} (4r_s)^d, \quad (10)$$

由 (10) 式得比例系数 $\rho = (\sqrt[d]{\frac{n}{m}} \cdot 4r_s)^d$. 由于自体数量 n 通常远大于聚类数量 m , 当 ρ 大于 1 时, CB-RNSA 的检测器生成代价小于传统的实值否定选择算法 RNSA 和 V-detector. 如图 2 所示在数据维数小于 20 且自体半径大于 0.01 时, ρ 大于 1.

图 2 r_s, d 对 ρ 的影响, $n/m=1000$ Figure 2 The relationship between ρ, d and $r_s, n/m=1000$

图 3 三角形自体区域

Figure 3 Triangle region of self data

图 4 五角星形自体区域

Figure 4 Pentacle region of self data

4 实验

本节通过实验验证 CB-RNSA 算法的性能. 实验数据包括经典数据集^[9]: synthetic data set (SDS), abalone, breast cancer wisconsin (BCW) 以及 spoken arabic digit (SAD), 上述四个数据集被广泛应用于检测器的性能测试以及生成效率分析等问题的研究^[6~9].

实验采用检测率 DR(detection rate) 与误警率 FA(false alarm rate) 对算法产生的检测器性能进行衡量, DR 与 FA 的定义如下:

$$DR = TP / (TP + FN), \quad (11)$$

$$FA = FP / (FP + TN), \quad (12)$$

其中, TP, TN, FP, FN 分别代表检测器正确识别的非自体的数量、正确识别的自体的数量、错误识别的自体的数量和错误识别的非自体的数量.

为测试自体数量、自体半径、自体区域形状等参数对 DR 和 FA 的影响, 本文使用 synthetic data set 对 CB-RNSA 产生检测器的性能进行验证. 实验数据分布如图 3,4 所示, 训练数据为从自体区域随机挑选出的 300 个自体点, 测试数据为从整个 $[0, 1]^2$ 区域随机选出的 1000 个数据点. 实验过程中自体半径的变化范围为 0.01~0.2, 期望覆盖率的变化范围为 85%~99.9%.

实验结果如图 5~7 所示. 从图 5 中可以看出, 在相同的自体半径下, 期望覆盖率正比于检测器的检测率和误警率; 而在相同的期望覆盖率下, 较小的自体半径训练出的检测器具有较高的检测率和误警率: $0 < r_s < 0.02, 80\% < DA, 20\% < FA$; 较大的自体半径下训练出的检测器的检测率和误警率都较

图 5 算法性能验证实验结果

Figure 5 The results of algorithm performance test. (a) DR of triangle self region; (b) FA of triangle self region; (c) DR of pentacle self region; (d) FA of pentacle self region

图 6 自体区域对检测器数量的影响

Figure 6 The impact of self region shapes on detector number

图 7 自体数量对检测性能的影响

Figure 7 The impact of self number on performance

低: $0.1 < r_s$, $DA < 60\%$, $FA < 10\%$. 因此, 对异常数据敏感的应用环境下训练检测器应采用小自体半径, 而对误报敏感的应用环境下应设置较大的自体半径.

表 5 实验参数

Table 5 Experimental parameters

Parameter	Value
Expected coverage	90% ~ 100%
Self radius	0.05 ~ 0.15
Initial cluster radius	3

图 8 检测率曲线, $r_s=0.08$
Figure 8 Curve of DR, $r_s=0.08$

图 9 误警率曲线, $r_s=0.08$
Figure 9 Curve of FA, $r_s=0.08$

从图 6 和图 7 中可以看出, 自体区域的形状和自体元素个数对检测性能也存在影响. 图 6 中的检测器数量曲线表明, 在相同的期望覆盖率条件下, 自体区域的形状越复杂 (五角星形自体区域) 所需的检测器数量越多, 这是由于复杂的自体区域中, 半径较大的检测器易覆盖自体区域而被淘汰, 因此需要大量半径较小的检测器以避免覆盖复杂的自体区域. 如图 7 所示, 随着自体个数的增加, 检测率逐渐增加, 同时误警率降低. 这是由于大量的自体元素可以有效的对检测器进行筛选, 减少自反应检测器的数量; 同时, 筛选后的检测器被集中在非自体空间, 增加了检测的准确性.

为了进一步验证 CB-RNSA 算法产生检测器的性能, 本文进行了检测器性能对比实验. 采用的比较对象为两种经典的实值否定选择算法 RNSA^[6] 和 V-detector^[9]. 实验采用的数据为 abalone. Abalone 包含 4177 条数据记录, 每条记录包含以下属性: length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, rings 以及类别标签 class (1 for male, 0 for others).

实验将类别标志为 1 的数据作为自体数据, 标志为 0 的数据为异常数据. 采用交叉验证法^[19] (cross validate method) 进行分组验证. 数据预处理阶段将数据规范化到 $[0,1]^d$ 空间中. 实验参数设置如表 5 所示, 实验过程重复 20 次.

由于传统的 RNSA^[6] 算法采用预设的检测器数量作为中止生成检测器的条件, 为了在相同的实验条件下对否定选择算法的性能进行比较, 本文对 RNSA 算法进行了修改, 使其在检测器的非自体空间覆盖率达到期望覆盖率时中止检测器的生成过程.

实验结果如图 8~10 所示, 曲线上的柱状直线表示方差. 从图 8 和图 9 中可以看出, 在期望覆盖率为 90% 时, CB-RNSA 的检测率比 RNSA 与 V-detector 算法分别高出 12.3% 与 7.4%, 而误警率则分别降低了 8.5% 与 4.9%. 一方面, 这是由于 CB-RNSA 引入了随机值范围, 限定候选检测器仅在非

图 10 P_{exp} 与检测器数量关系曲线, $r_s=0.08$ Figure 10 The effect of P_{exp} on detector number, $r_s=0.08$

图 11 算法产生检测器时间

Figure 11 The time cost of algorithms

表 6 数据集属性

Table 6 Data attributes

Data set	Records number	Data type	Dimension
BCW	699	integer	9
Abalone	4177	real, integer	8
SAD	8800	real	13

自体空间覆盖率较低的区域生成, 提高了检测器的非自体空间覆盖率; 另一方面, 采用聚类中心与聚类半径代替自体数据与自体半径与候选检测器进行匹配, 对“自反应”的界定更加严格, 在降低了检测器误报率的同时, 也减少了否定选择过程中的距离计算代价. 图 8 和图 9 中, CB-RNSA 产生的检测器的检测率和误警率的方差较另两种算法小, 算法的检测性能更加稳定.

图 10 中的候选检测器数量曲线表明, 为达到相同的期望覆盖率 CB-RNSA 比 RNSA 与 V-detector 需要的候选检测器数量更少, 且随着期望覆盖率的增加, 候选检测器数量增长更加缓慢: RNSA 23419-34091, V-detector 16730-32163, CB-RNSA 7297-15023.

为了测试 CB-RNSA 的算法效率并与传统的否定选择算法进行比较, 本文对 CB-RNSA, RNSA^[6] 和 V-detector^[9] 在相同的期望覆盖率下产生免疫检测器的效率进行对比研究, 实验采用的数据为三类不同维度的 UCI (University of California Irvine) 数据集: breast cancer wisconsin (BCW), abalone 和 spoken arabic digit (SAD).

数据集属性描述如表 6 所示. 在三类数据集上分别取类别标记为 “benign”, “male” 和 “digit 0” 的记录为自体数据, 实验的预处理阶段将数据记录在各维度上的取值规范化到 [0,1] 区间, 设置期望覆盖率为 90%. 类似地, 对 RNSA^[6] 的中止条件进行了修改, 使其在检测器的非自体空间覆盖率达到期望覆盖率时中止检测器的生成过程, 以便在相同的实验条件下对算法的效率进行比较.

20 次重复实验的平均结果如图 11 和表 7 所示. 从图 11 中可以看出, 为达到期望覆盖率 (90%), CB-RNSA 在数据集 BCW, abalone 和 SAD 上产生检测器的时间开销分别比 RNSA 与 V-detector 降低了 82.6%, 67.6%, 63.2% 和 86.7%, 75.7%, 64.2%. 结合图 11 中的时间曲线与表 7 中比例系数 ρ 可知, 数据集 abalone 上的 ρ 值最大, 但 CB-RNSA 在 BCW 数据集上产生检测器的时间开销比在 abalone 上更少, 这是因为如 3.4.1 小节的理论分析所述, CB-RNSA 的时间复杂度还跟自体聚类数量

表 7 对比实验结果

Table 7 The result of comparative experiment

Data set	ρ	Algorithm	Candidate detector (k)
BCW	3.9	RNSA	7.42
		V-detector	5.39
		CB-RNSA	2.16
Abalone	13.6	RNSA	23.42
		V-detector	16.7
		CB-RNSA	7.29
SAD	1.76	RNSA	110.9
		V-detector	83.7
		CB-RNSA	56.1

有关, 自体数据分布越紧密, 自体聚类数量越少, 则候选检测器与聚类中心的比较次数越少, 反之自体分布越稀疏, 自体聚类数量越多, 算法执行否定选择的时间开销越大. 因此选择否定选择算法产生检测器时, 应充分考虑自体集的分布, 自体数量, 自体半径等数据特征.

从表 7 中可以看出, 在数据集 Abalone 和 BCW 上, 由 (10) 式计算出的系数 ρ 分别为 3.9 和 13.6, 此时在相同的期望覆盖率下, CB-RNSA 需要的候选检测器数量远少于算法 RNSA 和 V-detector. 而当采用数据集 SAD 时, 随着数据维数增加到 13, 系数 ρ 等于 1.76, 此时 CB-RNSA 需要的候选检测器数量增加且更接近于传统的否定选择算法. 另一方面, 结合表 6 与表 7, 对比 BCW、abalone 与 SAD 的数据集属性与实验结果可以发现, 当数据维度越低且样本数量越多时, 系数值 ρ 越大, 实验结果与 3.4.2 小节关于系数 ρ 的理论分析相一致.

5 小结

传统的 NSA 算法时间代价过高, 难于实际应用. 同时传统 NSA 算法在整个特征空间内随机生成检测器, 并以预设的检测器数量作为中止生成检测器的条件, 造成检测器在部分非自体空间内的冗余分布, 而对非自体空间整体的覆盖率却不足. 针对这些问题, 本文提出了一种新的基于自体集层次聚类的实值否定选择算法 CB-RNSA, 给出了检测器非自体空间覆盖率的概率分析方法, 以期望的非自体空间覆盖率作为中止生成检测器的条件. 提出了用随机值范围控制新生成的检测器分布在低覆盖率区域的方法, 降低了检测器冗余, 提高了检测器非自体空间覆盖率. 同时, 算法用聚类中心代替自体数据与候选检测器进行相似度比较, 由于聚类数量远小于自体数量, 算法降低了检测器生成过程的时间开销, 提高了成熟检测器的生成效率, 有效地解决了传统的 NSA 算法中检测器训练的时间代价随自体数量呈指数增长这一难题. 理论分析和实验结果显示 CB-RNSA 相对经典的 NSA 算法具有更好的时间效率和检测器生成质量.

参考文献

- 1 Forrest S, Perelson A S, Allen L, et al. Self-nonsself discrimination in a computer. In: Proceeding of the IEEE Symposium on Research in Security and Privacy. Oakland: IEEE, 1994. 202-212

- 2 Li T. Dynamic detection for computer virus based on immune system. *Sci China Ser F-Inf Sci*, 2008, 51: 1475–1486
- 3 Stibor T, Timmis J, Eckert C. On the appropriateness of negative selection defined over hamming shape-space as a network intrusion detection system. In: *Proceedings of IEEE Evolutionary Computation*. Edinburgh: IEEE Computer Society Press, 2005. 995–1002
- 4 Kim J, Peter J B. An evaluation of negative selection in an artificial immune system for network intrusion detection. In: *Proceeding of GECCO*. San Francisco: Morgan Kaufmann, 2001. 199–213
- 5 Li T. *Computer Immunology*. Beijing: Publishing House of Electronics Industry, 2004. 55–56 [李涛. 计算机免疫学. 北京: 电子工业出版社, 2004. 55–56]
- 6 Gonzalez F, Dasgupta D, Nino L F. A randomized real-valued negative selection algorithm. In: *Proceeding of ICARIS*. Edinburgh, 2003. 261–272
- 7 Gonzalez F, Dasgupta D. Anomaly detection using real-valued negative selection. *Genet Progr Evolv Mach*, 2003, 4: 383–403
- 8 Zhou J, Dasgupta D. Real-valued negative selection algorithm with variable-sized detectors. In: *Proceedings of GECCO*. Seattle: Springer, 2004. 287–298
- 9 Zhou J, Dasgupta D. V-detector: an efficient negative selection algorithm with probably adequate detector coverage. *Inform sciences*, 2009, 19: 1390–1406
- 10 Edgar C, Gonzalo N. Measuring the Dimensionality of General Metric Spaces. Technical Report TR/DCC-00-1. 2000
- 11 Skala M. Measuring the difficulty of distance-based indexing. *LNCS*, 2005, 3772: 103–114
- 12 Perelson A S, Weisbuch G. Immunology for physicists. *Rev Mod Phys*, 1997, 69: 1219–1267
- 13 Han B R, Herrin B R, Cooper M D. Antigen recognition by variable lymphocyte receptors. *Science*, 2008, 321: 1834–1837
- 14 Mullighan C G, Philips L A, Su X P. Genomic analysis of the colonel origins of relapsed acute lymphoblastic leukemia. *Science*, 2008, 322: 1377–1380
- 15 Huang Y D, Di C N, Zhu S X. *The Matrix Theory and Application*. Hefei: Press of University of Science and Technology of China, 1995. 27–28 [黄有度, 狄成恩, 朱士信. 矩阵论及其应用. 合肥: 中国科学技术大学出版社, 1995. 27–28]
- 16 Stibor T, Philipp M, Jonathan T. Is negative selection appropriate for anomaly detection? In: *Proceeding of ACM GECCO*. Edinburgh: IEEE Computer Society Press, 2005. 569–576
- 17 Timmis J, Hone A, Stibor T, et al. Theoretical advances in artificial immune systems. *Theor Comput Sci*, 2008, 403: 11–32
- 18 Li T. An immunity based network security risk estimation. *Sci China Ser F-Inf Sci*, 2005, 48: 557–578
- 19 Sun C Y, Mu C X, Li X M. A weighted LS-SVM approach for the identification of a class of nonlinear inverse systems. *Sci China Ser F-Inf Sci*, 2009, 52: 770–779

A negative selection algorithm based on hierarchical clustering of self set

CHEN Wen¹, LI Tao^{1*}, LIU XiaoJie¹ & ZHANG Bing²

¹ College of Computer Science, Sichuan University, Chengdu 610065, China;

² National Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing 100029, China

*E-mail: litao@scu.edu.cn

Abstract Negative selection algorithm (NSA) is an important method of generating artificial immune detectors. However, the traditional NSAs aim at eliminating the self-recognized invalid detectors, by matching candidate detectors with the whole self set. The matching process results in extremely low generation efficiency and significantly limits the application of NSAs. In this paper, an improved NSA called CB-RNSA, which is based on

the hierarchical clustering of self set, is proposed. In CB-RNSA, the self data is first preprocessed by hierarchical clustering, and then replaced by the self cluster centers to match with candidate detectors in order to reduce the distance calculation cost. During the detector generation process, the candidate detectors are restricted to the lower coverage space to reduce the detector redundancy. In the paper, probabilistic analysis is performed on non-self coverage of detectors. Accordingly, termination condition of the detector generation procedure in CB-RNSA is given. It is more reasonable than that of traditional NSAs, which are based on predefined detector numbers. The theoretical analysis shows the time complexity of CB-RNSA is irrelevant to the self set size. Therefore, the difficult problem, in which the detector training cost is exponentially related to the size of self set in traditional NSAs, is resolved, and the efficiency of the detector generation under a big self set is also improved. The experimental results show that: under the same data set and expected coverage, the detection rate of CB-RNSA is higher than that of the classic RNSA and V-detector algorithms by 12.3% and 7.4% respectively. Moreover, the false alarm rate is lower by 8.5% and 4.9% respectively, and the time cost of CB-RNSA is lower by 67.6% and 75.7% respectively.

Keywords artificial immune system, negative selection algorithm, artificial immune detector, cluster, coverage

CHEN Wen was born in 1983. He received his Ph.D. in computer science from Sichuan University, Chengdu, in 2011. Currently, he is a Senior Engineer at the Information Security Laboratory, Sichuan University. His research interests include information security and machine learning. Dr. Chen is a member of IEEE, as well as IEICE.

LI Tao was born in 1965. He received the Ph.D. degree in computer science from University of Electronic Science and Technology of China in 1994. From 1994 to 1995, he was a visiting scholar in University of California at Berkeley for neural networks theory. Now, he is a Professor in the Department of Computer Science at Sichuan University, China. His current research interests include network security, artificial immune theory and disaster recovery application.

LIU XiaoJie was born in 1965. She received the M.S. degree in computer science from University of Electronic Science and Technology. Now, she is an Associate Professor in the Department of Computer Science at Sichuan University, China. Her current research interests include network security, artificial immune theory and disaster recovery application.

ZHANG Bing was born in 1973. He received his Ph.D. degree in Computer Science from the University of Harbin Institute of Technology in 2001. Now, he is a Senior Engineer at the National Computer Network Emergency Response Technical Team Coordination Center of China. His current research interests include network security monitor and cyberspace strategy.