

实数向量型阴性选择算法的改进

王华东 刘芳¹

(胜利油田纯梁采油厂信息中心, 博兴 256504; 胜利油田采油工艺研究院信息中心¹, 东营 257000)

摘要 通过对实数向量型阴性选择算法的分析, 提出了检测器空间覆盖率的概念, 用它作为估计检测器数量的一项理论依据, 将这个估计值引入到实数向量型阴性选择算法中, 控制检测器的生成, 同时对检测器采取了新的变异操作。实验表明, 这一改进在保证算法检出率的同时, 又可降低误报率。

关键词 实数向量型阴性选择算法 检测器 空间覆盖率 变异操作

中图法分类号 TP393.08; **文献标志码** A

阴性选择算法 (Negative Selection Algorithm, NSA) 最早由 S. Forrest 在 1994 年提出, 是第一个应用到异常检测的经典人工免疫算法。它的目的是区分系统中的自体 (Self) 与非自体 (No-self), 工作机制类似人体免疫系统中的自体细胞与抗原。主要思想是根据识别的对象产生一组与自体数据不匹配的检测器, 再利用这些检测器检测自体集的变化, 根据一定的规则寻找并清除非自体数据^[1]。

为解决入侵检测系统应用中不断出现的问题, 研究人员借鉴各种思想, 在很多方面对阴性选择算法作了改进^[2-4]。

1 阴性选择算法中检测器的表示

S. Forrest 将阴性选择算法应用于异常检测时, 检测器采用二进制字符串表示, 后来又有许多研究者提出采用多维实数向量表示检测器^[5]。举例说明这两种表示法的区别: 给定一个 16 位二进制检测器 1000000100000000, 其匹配法则为 r -连续位匹配规则; 给定一个实数向量型检测器 (0.6, 0.8, 0.9), 匹配过程使用 Euclidean 距离。两种不同表示法的检测器在相同的空间里覆盖不同的区域范围, 如图 1 所示, 检测器的覆盖区域由深色表示。

从图 1(a) 中可以看出一定数量的二进制检测

器很难覆盖形状不规则的非自体, 这是二进制阴性选择算法的主要缺点。而单个二维实数向量型检测器覆盖的形状是一个圆形, 如图 1(b), 这样的形状更适合覆盖随机形状的非自体集。

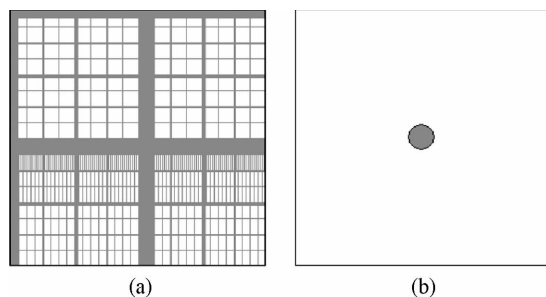


图1 单个检测器在问题空间内覆盖区域

2 实数向量型阴性选择算法 (Real-valued-vector Negative Selection Algorithm, RNSA) 研究

2.1 实数向量型阴性选择算法简述

实数向量型阴性选择算法是由 Gonzalez 等人首次提出^[5]。在该算法中, 自体和非自体空间是 $\mathbf{R}^n$ 的一个子集, 特别设定该实数向量集为 $[0, 1]^n$ 的一个子集。检测器 (抗体) 就可以被定义为一个 n 维实数向量。该 n 维实数向量可以确定某个检测器的位置, 并且规定该检测器的半径, 因此该检测器可以看作维数为 n 的一个超球体, 当 $n = 2$ 时, 该检测器是一个圆形覆盖的区域; 当 $n = 3$ 时, 该检测器是

一个球体。匹配算法采用 Euclidean 距离,计算抗原和检测器之间的亲和度。

该算法由一组 n 维的实数向量作为自体集样本,并由该自体集生成一个检测器(抗体)集合,使之能够覆盖非自体区域。该算法所要遵循的两个原则就是:

- (1) 检测器不能覆盖到自体集区域。
- (2) 让检测器尽量分散,能够最大限度的覆盖非自体区域。

这种算法思想类似于贪心阴性选择算法,但是该算法的作用域是实数空间。

2.2 空间覆盖率的概念

在给定的问题域中,必须讨论某个实数向量与该空间中随机的某个实数向量所能匹配的概率,这里用空间覆盖率的概念来表示,记为 $p_M(r_0)$ 。它实际上就是以空间中某个点为中心的超球体的体积,球体的半径为 r_0 。设问题域为 l 维实数空间,其中的向量可以表示为集合 $[0.0,1.0]^l$,如由 Euclidean 距离设定的距离阈值为 r_0 ,根据空间中超球体的计算公式,

$$p_M(r_0) = 2^l \times \underbrace{\left[\frac{\pi}{2} \times 1 \times \frac{\pi}{4} \times \frac{2}{3} \times \frac{3\pi}{16} \times \dots \times \int_0^1 \frac{y_i^{l-i} dy_i}{\sqrt{1-y_i^2}} \times \dots \times \int_0^1 \frac{y_1^{l-2} dy_1}{\sqrt{1-y_1^2}} \right]}_{(1-1)} \times \frac{r_0^l}{l} \quad (1)$$

可以计算出空间中某个随机向量对应点在阈值为 r_0 时的体积,当 $l=2$ 时, $p_M(r_0) = \pi r_0^2$,表示在二维空间中半径为 r_0 的圆形面积;当 $l=3$ 时, $p_{M2}(r_0) = \frac{4\pi}{3} r_0^3$,表示在三维空间中半径为 r_0 的球体体积。用计算出来的体积作为某个检测器的空间覆盖率,在整个空间已知的情况下就可以估计所需检测器的数量。

根据另一种空间中超球体的计算公式,

$$p_M(r_0) = \frac{2\pi^{\frac{l}{2}}}{\Gamma\left(\frac{l}{2}\right)} \times \frac{r_0^l}{l} \quad (2)$$

$x \in N$:

- (1) $\Gamma(x) = (x-1)!$;
- (2) $\Gamma(x+1) = x\Gamma(x)$;

$$(3) \Gamma(x)\Gamma\left(x + \frac{1}{2}\right) = \frac{\sqrt{\pi}}{2^{2x-1}}\Gamma(2x)。$$

用空间超球体的体积来表示空间覆盖率,根据公式(1),当 $l=13, r_0=0.05$ 时可以得到:

$$p_M(r_0) < \frac{2^{13}\pi}{13}r_0^{13} \approx 69r_0^{13}。$$

通过公式(2),可以得到 $p_M(r_0) \approx 11.838r_0^{13}$ 。可见后一种计算方法比前一种精确,但后者在整个空间一定的情况下所需的检测器数量多。可以根据需要选择其中一种计算空间覆盖率的方法。

2.3 算法的缺陷分析

当 $r_0=0.05$,根据式(1), $p_M(r_0) = 8.442 \times 10^{-16}$,根据式(2), $p_M(r_0) = 1.445 \times 10^{-16}$ 。可见这样计算出来的某个实数检测器向量的空间覆盖范围是很小的,这样会导致候选检测器在耐受期中将不会与自体集合中的向量所匹配,也就是说,在耐受期,候选检测器都将成熟,这就会造成选用的检测器数量过大,进而影响检测器搜索的效率并提高了误报率;而且已有实验表明,在检测器的数量增加到一定程度后,即使再增加检测器的数量也不会使检出率提高^[6]。因此,选用多少检测器、如何选用检测器对实数向量型阴性选择算法来说十分重要。

3 实数向量型阴性选择算法的改进

检测器对于阴性选择算法来说是一个很重要的参数,适当数量的高质量检测器能够保证较高的检测速度同时又能降低误报率,因而必须研究影响检测器数量的因素。

3.1 成熟检测器数量估计

研究表明, l 维实数空间中的某一点至少需要 $3+2^{l-2}$ 个检测因子才能保证被检测到,这是由 2 维和 3 维空间所决定。在 2 维平面空间中,至少有 3 个圆形可以靠近一点;在 3 维立体空间中,至少需要 $5(3+2)$ 个球体可以靠近一点。由此得出,靠近一个半径为 r_0 圆形区域的检测器的数量应该等于该圆形区域的覆盖率与检测器数量的乘积。设所需检测器的数量为 n ,则可以得出以下结论:

$$n \geq \frac{3 + 2^{l-2}}{p_M(r_0)}, \quad (l > 2) \quad (3)$$

通过式(3)就可以估算出在问题空间中所需要的检测器数量,在算法当中引入这个估算值作为一个参数控制检测器的产生,以降低产生检测器数量的盲目性。

同时为降低检测器的数量,还要考虑检测器的质量问题,在检测器生成时,要考虑对它们的遗传变异操作,以减少它们之间的相似性。

3.2 实数向量型阴性选择算法的具体改进方案

3.2.1 初始值确定

根据空间覆盖率、检测器半径、候选检测器的数量,确定算法的迭代次数和估计成熟检测器数量,以及是否需要将某检测器进行变异。

3.2.2 遗传变异操作

候选检测器需要根据其他检测器或自体集向量来进行变异生成,这里采用的是每个向量中的各个分量依据不同的步长逐次进化法。在该算法中,可能一个检测器需要进行多次不同的移动,以远离其他检测器,每一次移动需要在该检测器向量的每个分量上分别加上或者减去某个不同的值。例如,第一个分量的步长设为 0.001,第二个为 0.002,第三个为 0.000 1,给定三个三维的实数向量检测器,分别为(0.8,0.7,0.9),(0.6,0.65,0.8)和(0.5,0.9,0.7),假设第一个向量发生变异远离第二、三个检测器。首先确定向量(0.8,0.7,0.9)中的第一个分量的变化趋势,根据三个向量的第一个分量之间的关系,(0.8-0.6)+(0.8-0.5)=0.5,该值为正,发生位移的向量的第一个分量就为 0.8+0.001=0.801;确定第二个分量,(0.7-0.65)+(0.7-0.9)=-0.15,所以第二个分量的值为 0.7-0.002=0.698。确定第三个分量(0.9-0.8)+(0.9-0.7)=0.3,为正值,第三个分量就为 0.9+0.000 1=0.900 1,经过一次变异,该向量变为(0.801,0.698,0.900 1)。

在实际应用中,要根据具体的环境来选择适合的遗传变异策略。

3.2.3 算法描述

该算法的伪代码如下:

Improved-Real-valued-vector-Negative-Selection-Arithmetic (r_0 , num , 1, micro-step)

r_0 :距离阈值(检测器半径)

num :候选检测器的数量

l :检测器向量的维数

micro-step[0..1]:每一个检测器分量变异时所要移动的步长

(1). 随机生成 num 个候选检测器;

(2). 通过空间覆盖率函数 $Pm(r_0)$ 计算预期检测器的数量 n ;

(3). 比较 n 和 num ,确定迭带次数。

For 每一次迭代

For(每一个候选检测器 d)

. 根据阴性选择算法对检测器进行变异

IF(检测器 d 与一个自体集中的向量匹配 s)

将检测器 d 远离自体集

For(d 中的每一个分量, $i=0$ to 1)

IF(($\sum_{self} d_i - s_i > 0$))

$d_i = d_i + \text{micro-step}[i]$

Else

$d_i = d_i - \text{micro-step}[i]$

End For

Else

将检测器 d 远离其他检测器

For(d 中的每一个分量, $i=0$ to 1)

IF(($\sum_{detectors} d_i - s_i > 0$))

$d_i = d_i + \text{micro-step}[i]$

Else

$d_i = d_i - \text{micro-step}[i]$

End For

End For

End For

3.3 实验分析

实验采用的自体训练集、自体检测数据集和异常检测数据集由延迟微分方程 Mackey-Glass 产生^[7],如等式(4)。

$$\frac{dx}{dt} = \frac{ax(t-\tau)}{1+x^c(t-\tau)} - bx(t) \quad (4)$$

为便于比较,对原阴性选择算法(RNSA)和改进的阴性选择算法(IRNSA)均采用同样的训练集和检测集,每组实验重复 5 次。初始参数如下:

自体训练集大小:1 000;

自体检测集大小:1 500;

异常数据集大小:500;

成熟检测器个数:2 200;
检测器维数: $l=6$;
检测器半径: $r=0.35$ 。
实验结果如表 1 和表 2 所示。

表 1 RNSA 仿真实验结果

实验一	识别异常数据 集个体数	识别自体集 个体数	识别率/%	误报率/%
1	488	22	97.6	1.47
2	487	23	97.5	1.53
3	487	39	97.5	2.60
4	493	18	98.6	1.20
5	490	24	98.0	1.60
平均值	489	25.2	97.8	1.68

表 2 改进的 RNSA 仿真实验结果

实验二	识别异常数据 集个体数	识别自体集 个体数	识别率/%	误报率/%
1	496	18	99.1	1.20
2	496	20	99.1	1.33
3	486	29	97.2	1.93
4	486	19	97.2	1.27
5	477	17	95.4	1.13
平均值	488.2	25.2	97.6	1.372

在实验 1 中,检出率平均值可以达到 97.8%,实验 2 中,检出率平均值稍有降低,为 97.6%。由于简单变异策略的引入,使得少部分异常数据被当成自体集数据。但是改进的 RNSA 算法的误报率在整体水平上较原 RNSA 降低了,并且最少可以达到 1.13%。

通过实验可以看出,改进的实数向量型阴性选择算法通过控制检测器的数量和质量,在保证一定

检出率的同时,又降低了误报率,说明该算法的改进是有效的。

4 总结

本文分析了实数向量型阴性选择算法中产生检测器的缺陷,通过空间覆盖率的计算提出一种确定成熟检测器数量的方法,并结合一定的遗传变异策略,改进了实数向量型阴性选择算法。实验证明了改进算法的有效性。但是在检出率方面还有待于提高,这是我们下一步努力的目标。

参 考 文 献

1 Dasgupta D, Ji Z, Gonzalez F. Artificial immune system (AIS) research in the last five years, CEC -03, 2003. The 2003 Congress. 2003. 123—130

2 王煦法,张显俊,曹先彬,等. 一种基于免疫原理的遗传算法. 小型微型计算机系统,1999;20(2):117—120

3 曹先彬,罗文坚,王煦法. 基于免疫网络调节的改进遗传算法. 高技术通讯,2000;10: 22—27

4 张海英,管洪娜,潘永湘. 一种改进的阴性选择免疫算法. 西安理工大学学报,2005;21(3): 306—309

5 Ji Zhou, Dasgupta D. Real-valued negative selection using variable-sized detectors. In: The Proceedings of International Conference on Genetic and Evolutionary Computation (GECCO). Seattle, Washington USA, 2004, June 26—30:287—298

6 Balachandran S. Multi-shaped detector generation using real valued representation for anomaly detection. The University of Memphis, Uusa, 2005;11:42—46

7 Dasgupta D, Forrest S. Novelty detection in time series data using ideas from immunology. In: The Proceedings of the 5th International Conference on Intelligent Systems, (Received The Best Paper Award), Reno, 1996, June 19—21

(下转第 2241 页)

Type Selection Design of Support Truss System in Curtain Wall of Sightseeing Hall

ZHANG Wen-bo

(Shanghai Nuclear Engineering Research & Design Institute, Shanghai 200233, P. R. China)

[Abstract] In order to be satisfied with the aesthetic property and permeability required by architectural design, type selection analysis for support truss system with 12 meters height in curtain wall of sightseeing hall were carried out. The FEM software SAP2000 was employed for the comparing computations of steel truss and cable truss. The calculating results showed the stress ratio and displacements of the above two type trusses were satisfied with design codes. Under the given wind case, the quake case and the combined load cases, the stress ratio of the chord members in the steel truss is greater than the cable truss, the stress ratio of the web members in the steel truss is less than the cable truss, and the displacements of cable truss is greater than steel truss. Considered the values of the stress ration and displacements of the two types of truss, the cable truss is chosen to be used as the support system of curtain wall of the sightseeing hall.

[Key words] curtain wall steel truss cable truss stress ratio displacement

(上接第 2236 页)

An Improved Algorithm of Real-valued Vector of Negative Selection

WANG Hua-dong, LIU Fang¹

(Science & Technology Division, Chunliang Oil Production Co., Shengli Oilfield, Boxing 256504, P. R. China;

The Information Cerntrle of Shengli Oil Production Research Institute¹, Dongying 257000, P. R. China)

[Abstract] Based on the study of Algorithm of Real-valued Vector of Negative Selection, the concept of space coverage of detectors is introduced and used as a theoretical basis for the estimated number of detector. This estimated number to the algorithm to control the produce of detectors is applied. Then new mutation is adopt trend to detectors in the algorithm. From the results of testing on the improved step, the new algorithm can ensure preferable detection rate and low false alarm rate.

[Key words] algorithm of real-valued vector of negative selection detector space coverage mutation trend