基于改进负选择算法的异常检测

汪慧敏
(中国空空导弹研究院,河南 洛阳 471009)

摘 要:为实现用较少数目的检测器覆盖较大的非自体空间,提出一种基于渐增式矩形检测器的负选择算法。该方法采用 D_∞ 距离匹配原则,检测器在每一维的方向上呈指数形式逐渐增长,直至与自体空间相匹配,从而使得产生的每个检测器在空间的每一维都延伸至最大,能够产生足够优秀的检测器集覆盖非自体空间。通过对检测器的合并,消除重叠等简化处理,实现了检测器的数目和大小的双重优化。对不同几何形状的数据集合进行了仿真。实验结果表明,该算法在对非自体空间的覆盖和检测率的提高方面有显著的效果。

关键词:异常检测;人工免疫系统;负选择算法;超矩形检测器
中图分类号:TP202 **文献标识码**:A **文章编号**:1673-629X(2009)08-0041-04

Anomaly Detection Using Modified Negative Selection Algorithm

WANG Hui-min
(China Airborne Missile Academy, Luoyang 471009, China)

Abstract: In order to occupy more coverage of the non-self space using fewer detectors, a new negative selection algorithm based on incremental rectangle detectors is proposed. In this scheme, the D_∞ distance is employed to measure the self and nonself space coverage. The sizes of the detectors are increased exponentially in each dimension until they overlap with the self samples. The generation strategy of detectors ensures that every detector is extended to its maximum size in each dimension. The number and size of these detectors are optimized by eliminating the redundancy among the existing detectors. Consequently, the detection efficiency of individual detector is increased. Some datasets of different geometry shapes are applied to examine the proposed NSA. Experimental results show that the proposed algorithm has remarkable advantages in both coverage of anomaly state space and detection rate.

Key words: anomaly detection; artificial immune system; negative selection algorithm; hyper-rectangle detector

0 引 言

基于传统的解析模型和先验知识的方法能够实现对设备进行异常检测,但是对于复杂系统,精确模型是很难建立的,先验知识又难以获取[1]。对于大多数设备来说,正常数据容易获取,故障数据是极为缺乏的,而借鉴免疫系统的负选择机理[2]提出的负选择算法在异常检测时不需要先验知识,是基于大量的正常数据进行的[3],具有很强的鲁棒性、并行性和分布式检测等特点,非常适合于故障样本缺乏或未知故障的异常检测问题。

目前,基于负选择的异常检测算法中亟待解决的问题是检测器的数目和检测器对非自体空间的覆盖[4]。为了实现用较少数目的检测器覆盖更大的非自体空间,提出了一种新的基于渐增式矩形检测器的负选择算法进行异常检测。该方法检测器的产生是渐增式的,使得产生的每个检测器都足够大,通过对检测器的简化处理,实现了对检测器的数目和大小的双重优化。通过对多组数据的仿真,验证了该方法的有效性。

收稿日期:2008-12-02;修回日期:2009-03-01
作者简介:汪慧敏(1985-),女,江苏人,硕士研究生,研究方向为系统算法及飞控软件设计;导师:高晓智,教授,主要研究方向为神经网络控制、模糊逻辑控制等。

1 基于负选择算法的异常检测

1.1 负选择算法

Forrest 于 1994 年提出了负选择算法[5](negative selection algorithm, NSA),算法的基本思想就是通过正常数据集合生成检测器集合,然后使用检测器来检测系统或设备的状态。

根据检测器的表示方式,可以把现有的负选择算法分为两类:编码负选择算法与实值负选择算法(real-valued NSA, RNSA)。编码负选择算法是把实值数据映射成二进制串或其它数字字符串,每个检测器都用这些数字串表示检测器和自体样本;RNSA 用高维的结构来表示数据,加快了检测器的产生过程。

目前 RNSA 中检测器的形状有三种：超矩形[6]、超球形和超椭球形。考虑到 RNSA 是在归一化的超矩形空间 $[0,1]^n$ 中进行操作，n 表示维数空间大小，并且超矩形具有很好的覆盖空间的能力，因此文中提出了一种新的渐增式超矩形检测器的生成算法，该算法虽简单却很有效。

1.2 异常检测

在异常检测的过程中，要分析检测方法的有效性，为此首先定义几个变量[7]：*TP*— 异常样本中被检测为异常的个数；*FN*— 异常样本中被检测为正常的个数；*FP*— 正常样本中被检测为异常的个数；*TN*— 正常样本中被检测为正常的个数。那么，检测率为 DR = $\frac{TP}{TP+FN}$，误报率为 FAR = $\frac{FP}{TN+FP}$，异常检测方法的有效性将使用这两个值来衡量。

2 基于渐增式矩形检测器的异常检测

超矩形检测器 d 的表示形式：d = (low, high)，其中，low = $(low_1, low_2, \cdots, low_n)$，high = $(high_1, high_2, \cdots, high_n)$，low 和 high 分别表示了超矩形的最低的转角点和最高的转角点。由于高维空间中的超矩形的产生和计算较为复杂，所以文中的讨论均是在二维空间中，即 $n = 2$，检测器的形状为矩形。

2.1 渐增式矩形检测器的产生算法

输入参数：自体集 S，最大检测器数 T_{max}，自体半径 Rs，检测器最小值 Rr。

输出参数：检测器集合 D。

Step1：$D \leftarrow \emptyset$；

Step2：While(num(D)! = T_{max})；

Step3：$x_r \leftarrow$ 从$[0, 1]n$ 中随机产生矢量；

Step4：For 检测器集 D 中的每个检测器 d = (low, high)；If $x_r(i) \in [low_i, high_i]$，其中 $i = 1, \cdots, n$，则 x_r 落入已生成的检测器内，取消该矢量，重新产生新的 x_r；

Step5：For 自体集合 S 中的每个元素 $s = (c_s, Rs)$，判断 x_r 是否与自体元素相匹配，若是，则重新产生新的候选矢量 x_r；

Step6：初始化检测器，low = high = x_r，衰减系数 Th，步长初值 Lh，增长代数 Max$t$，当前代数置为 t = 0；

Step7：While (t ! = Max$t$)；

Step8：For $j = 1, 2, \cdots, n$ 做 ① 至 ③ 步：

①low(num, j) = low(num, j) (Lh * exp(− t/Th)

high(num, j) = high (num, j) (Lh * exp(− t/Th)；

②If low(num, j) < 0，则 low(num, j) = 0

If high(num, j) > 1，则 high(num, j) = 1；

③For 自体集合 S 中的每个元素 s = (cs, Rs)，判断新产生的检测器是否与自体集相匹配，如果匹配是由 low(num, j) 的变化引起的，则：

low(num, j) = low(num, j) (Lh * exp(− t/Th)

如果匹配是由 high(num, j) 的变化引起的，则：

high(num, j) = high(num, j) (Lh * exp(− t/Th)

Step9：$D \leftarrow D \cup \{<$ low, high $>\}$；

Step10：End；

Step11：对 $j = 1, 2, \cdots, n$ $\exists j$，使 high(num, j) − low(num, j) < Rr，则取消该检测器；

Step12：End。

2.2 渐增式矩形检测器的算法分析

(1)自体半径 Rs 的选取：自体半径大小的选取是使自体集尽可能地覆盖待检测数据中正常的样本。

(2)检测器最小值 Rr 的选取：Rr 是每一维的最小长度，表示所产生的矩形检测器 d ≥ Rr × Rr。文中的矩形检测器的最小值取为 Rr = 2 * Rs。

(3) D_∞ 距离匹配原则：D_∞ 距离可以用来检测特征空间中 □ 形超立方体结构，其定义为，若有 n 维空间上的两点 $x = (x_1, x_2, \cdots, x_n)$，$y = (y_1, y_2, \cdots, y_n)$，则 x 和 y 两点之间的 D_∞ 距离为 $d(x, y) = \max_{i=1,2,\cdots,n} | x_i - y_i |$。

(4)检测器增长方式及参数选取：文中检测器的增长方式为指数衰减形式。为使产生的检测器充分大，初始步长 Lh 不能取很大，文中取 Lh < 2 * Rs；另外，增长代数要保证检测器增到最大，即每一代增长的步长之和 sum 至少应满足 sum = $\sum_{t=0}^{Maxt} Lh * \exp(-t/Th) \geqslant 1$

(5)检测器的增长过程：检测器的增长过程即 2.1 小节中的 Step6 至 Step9。

2.3 检测器的简化

由于上述方法产生的检测器之间可能交叠很大，需对其进行简化。简化处理主要如下：

(1)消叠：消除检测器交叠较大的部分。

(2)舍弃：如果两个检测器为包含或几乎包含关系，则舍弃被包含的那个检测器，判断两个检测器是否包含时所允许的阈值定义为 Ra。

(3)拆分：对于两个相互交叉的检测器，将其中的一个检测器去除与另一检测器相交叉的部分，从而将其拆分成两个小的检测器。

(4)合并：如果两个检测器在 (n − 1)维上重合或相近，则将其合并成一个检测器，在判断两个检测器在

某一维是否相近时所允许的阈值定义为 Rc。

2.4 基于渐增式矩形检测器的异常检测

矩形检测器对时间序列异常信号数据的检测问题可描述如下:对于 $\forall$ 待检测矢量 $\boldsymbol{x}$,如果 $\exists$ d = (low, high) $\in D$, If $\boldsymbol{x}(i) \in [\text{low}_i, \text{high}_i]$, $i = 1, 2, \cdots, n$,则认为 $\boldsymbol{x}$ 落入检测器 d 内,检测器被激活,该段数据发生异常,标志为 1;否则,$x \in S$,该段数据正常,标志为 0。

文中产生的矩形检测器具有以下优点:

(1)通过使用更少数量的检测器来节省产生检测器和检查新的样本的时间,也节省了存储检测器所需要的空间。

(2)可以接受一些更小的检测器,所以对孔洞[8]的覆盖更容易。

(3)通过对矩形检测器简化处理,使得检测器的数目和大小达到最优,提高了检测器的利用率。

(4)产生的矩形检测器均位于 $[0,1]^n$ 空间里,而超球形检测器会有一些边缘检测器(检测器的空间范围超出了检测空间 $[0,1]^n$)。

3 仿真结果

为了验证该算法的实用性,特对各种几何特征的数据集合进行仿真。仿真数据一共有 8 对(训练数据集合和待检测数据集合),每个数据集合均有 1000 个点,每个点均位于空间 $[0,1]^2$ 中。这 8 对数据一共包括 4 种几何形状(训练数据集合所有点组成的图形):十字形(Cross)、交叉形(Intersection)、环形(Ring)、三角形(Triangle)。每种形状还有它们的互补图形,命名为 * * * _neg。

由于数据均处于 $[0,1]^2$,所以不需要对数据进行归一化处理。建立自体集后,用文中提出的负选择算法产生检测器集合,最后对待检测数据集合进行异常检测。

在仿真过程中各参数值分别为:自体半径 Rs = 不定,窗口长度 $n = 2$,滑动步长 step = 1,初始检测器个数 $m = 20$,检测器进化的步长初值 Lh = 0.02,检测器进化的衰减系数 Th = 100,检测器增长的代数 Max$t$ = 100,检测器每一维的最小长度 Rr = 2 * Rs,合并阈值 Rc = 0.05,阈值 Ra = 0.01。

自体集和检测器集对检测空间的覆盖分别如图 1 至图 4 所示,其中圆形覆盖的区域为自体集,矩形为产生的检测器。每组数据均对 10 次仿真取平均值,所得结果如表 1 所示。

由表 1 和图 1 至图 4 可以得出以下结论:

(1)如图 1 至 4 所示,除自体空间为环补形的数据集合外,其它的所产生的检测器之间交叠都很小,甚至

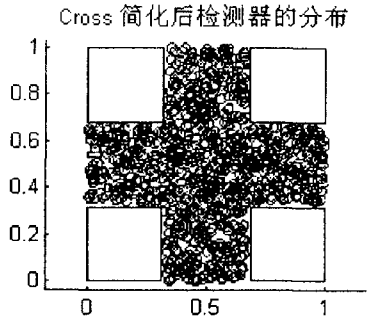
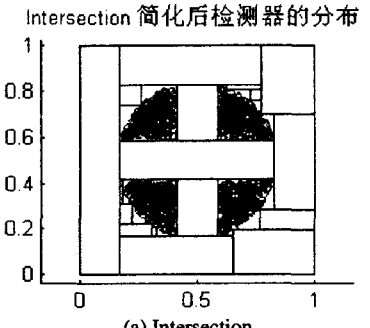
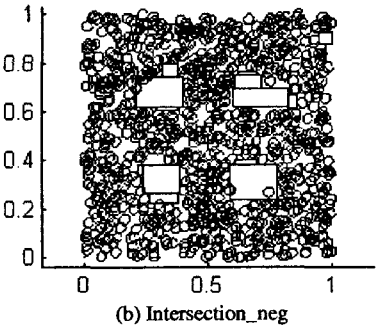


图 1　Cross 自体和检测器的分布

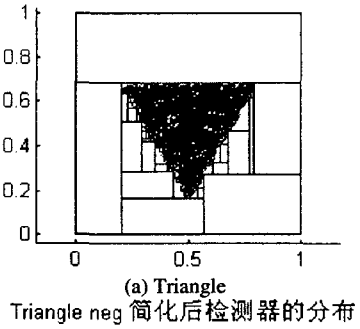


(a) Intersection

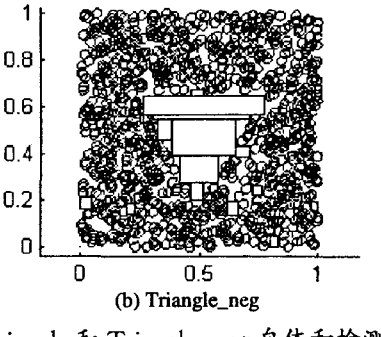


(b) Intersection_neg

图 2　Intersection 和 Intersection_neg 自体和检测器分布

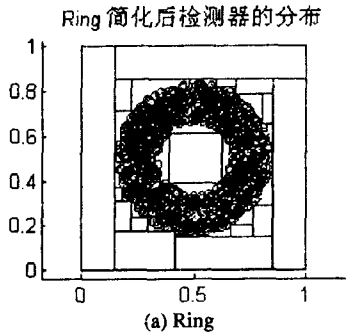


(a) Triangle

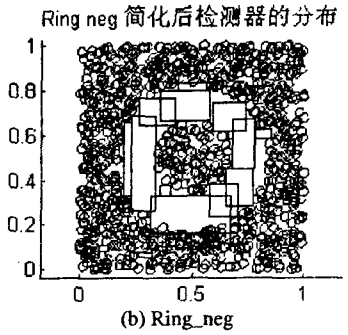


(b) Triangle_neg

图 3　Triangle 和 Triangle_neg 自体和检测器的分布



(a) Ring



(b) Ring_neg

图 4 Ring 和 Ring_neg 自体和检测器的分布

表 1 各种类型数据集合的检测结果

参数 \ 数据	Cross	Cross - neg	Intersection	Intersection - neg
Rs	0.02	0.02	0.015	0.02
Opm 大小	8.1	8.4	17.3	13.1
Opm 取值	[6, 9]	[6, 12]	[15, 19]	[10, 17]
DR (%)	88.68	91.77	91.03	67.10
FAR (%)	0	0.21	1.25	1.47
参数 \ 数据	Triangle	Triangle - neg	Ring	Ring - neg
Rs	0.015	0.02	0.02	0.02
Opm 大小	17.2	13.8	18.9	17
Opm 取值	[16, 20]	[11, 19]	[17, 21]	[13, 21]
DR (%)	95.90	87.87	85.39	83.56
FAR(%)	0	2.32	0.28	2.24

是零交叠，极大程度地提高了检测器的空间利用率。这充分说明了对检测器简化处理的有效性。

(2)由表 1、图 1 可知，矩形检测器对于自体空间为十字(补)形即非自体空间为矩形的数据集合，检测效果最好，检测率高，几乎均在 90% 以上，而且简化后检测器的数目很少，从而实现了用尽可能少的检测器覆盖了较大的非自体空间。

(3)由表 1、图 2(a) 和图 3(a)可知，矩形检测器对于自体空间为交叉形、三角形的数据类型检测效果很好，检测率较高，一般在 90% 以上。但检测器的数目简化前后变化不大。

(4)由表 1、图 2(b)和图 3(b)可知，矩形检测器对于自体空间为三角补形的检测效果比交叉补形的好。这是由其非自体空间形状决定的，三角补形的非自体空间为坡面，而交叉补形的非自体空间存在部分曲面，用矩形覆盖相对于坡面来说会产生大的孔洞，所以其检测效果差。

(5)由表 1、图 4(a)和图 4(b)可知，矩形检测器对于自体空间为环形和环补形的数据类型检测效果良好，相对来说自体为环形的检测效果要好些，因为其非自体空间边缘为仅一个曲面，而环补形的非自体空间边缘均为曲面，即要覆盖的区域为一环形，用矩形检测器对其覆盖孔洞大，而且检测器之间也存在很大的交叠，如图 4(b)所示。

4 结束语

为了实现用较少数目的检测器覆盖更大的非自体空间，提出了一种新的基于渐增式矩形检测器的负选择算法进行异常检测。该方法采用 D_∞ 距离匹配原则，检测器首先初始化为检测空间的一个随机点，然后在每一维的方向上呈指数逐渐增长，直至与自体空间相匹配，使得产生的每个矩形检测器在空间的每一维都延伸至最大，从而能够产生足够优秀的检测器集覆盖非自体空间。文中采用的渐增式矩形检测器大小可以随数据分布而灵活变化，对于非自体空间为方形或近似方形的检测效果尤为突出，仅需极少数目的检测器就能覆盖很大的非自体空间。

由于超矩形之间相对位置关系的描述很复杂，所以这种方法目前仅是在二维空间中实现，如何延伸至高维空间将是进一步的研究目标。

参考文献：

[1] 葛丽娜，钟 诚. 基于人工免疫的入侵检测系统负选择并行算法[J]. 计算机工程，2005，31(12)：138－140.

[2] 莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨：哈尔滨工业大学出版社，2002：1－4.

[3] Dasgupta D, Forrest S. Novelty Detection in Time Series Data Using Ideas from Immunology[M]. Cary, NC: ISCA, 1996: 82－87.

[4] 董永贵，孙照焱，贾惠波. 时间序列中异常值检测的负选择算法[J]. 机械工程学报，2004，40(10)：30－34.

[5] Forrest S, Perelson A, Allen L, et al. Self/non－self Discrimination in A Computer[M]. Washington: IEEE Computer Society, 1994: 202－212.

[6] Dasgupta D, González F. An Immunity－Based Technique to Characterize Intrusions in Computer Networks [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3): 281－291.

[7] 于宗艳. 基于人工免疫算法的故障诊断问题研究[D]. 大庆：大庆石油学院，2006：43－46.

[8] 熊春柳. 人工免疫在故障诊断中的应用[D]. 杭州：浙江大学，2006：15－20.