文章编号:1001 - 9081(2008)06 - 1438 - 03

一种基于自我聚类的异常检测学习方法

李娜娜[1,2],赵 政[1],刘伯颖[3],顾军华[2]
(1. 天津大学 计算机科学与技术学院,天津 300072; 2. 河北工业大学 计算机科学与软件学院,天津 300130;
3. 河北工业大学 教务处,天津 300130)
(hellosmiling@hebut.edu.cn)

摘 要:提出一种新的基于正选择的异常检测方法,该方法通过聚类学习正常空间特征,在每个类中选择有代表性的自我样本构造检测器集,之后利用正选择算法进行异常检测。这种方法既能适用于自我样本集较多的情形,克服了 T. Stibor 提出的正选择的局限,又具备了一定的学习能力。同时,该方法还避免了负选择中随机选择样本带来的弊端。通过实验分析,该方法比 VDetector 具备更好的检测性能,是一种有效的异常检测方法。

关键词:聚类;异常检测;负选择;正选择
中图分类号: TP393.08 **文献标志码**:A

Anomaly detection method by clustering normal data

LI Na-na[1,2], ZHAO Zheng[1], LIU Bo-ying[3], GU Jun-hua[2]
(1. *School of Computer Science and Technology, Tianjin University, Tianjin* 300072, *China*;
2. *School of Computer Science and Engineering, Hebei University of Technology, Tianjin* 300130, *China*;
3. *Office of Educational Affaires, Hebei University of Technology, Tianjin* 300130, *China*)

Abstract: A new anomaly detection method was proposed based on positive selection. The method learned the characteristic of "self" space by clustering, and then selected typical samples from every cluster to construct detectors. And positive selection was used to detect anomalies. The new algorithm is not only effective in certain application with large number of "self" samples, but also avoids the shortcoming by randomly selecting sample in VDetector. Experimental results on Ring data and biomedical data show that the new method is more effective in anomaly detection.

Key words: cluster; anomaly detection; negative selection; positive selection

0 引言

异常检测是当前研究的热点领域[1-4],它建立在正常模型的基础上,通过实际行为与之比较是否偏离来判断入侵行为。它能检测已知入侵,也能检测未知入侵。换句话说,异常检测也是一种特殊的分类,它的训练数据中仅包含一类可用数据,即正常数据。

基于负选择的异常检测是当前异常检测研究的主要方法之一[5],文献[6-7]提出了一种使用实值负选择算法进行异常检测的新模型。在此基础上有学者提出了 VDetector[8-9],并对 VDetector 进行了分析和完善[10-11]。文献[12-13]对 VDetector 算法提出质疑,并提出了基于正选择的异常检测方法。文献[14]则指出正选择算法只适用于正常样本不多的情形,但实际情况下的正常样本数目很多,正选择不是一种有效的学习算法。鉴于此,本文提出了一种具有学习能力的正选择算法,通过聚类学习正常空间特征,选择有代表性的自我样本进行基于正选择的异常检测。

1 基于正、负选择的异常检测

1.1 负选择

文献[6-7]中使用实值负选择算法进行异常检测的模型将所有状态均定义在一个 $[0,1]^n$ 超矩形中,检测器 d 由一二维向量 (c,rns) 描述,其中,$c \in [0,1]^n$ 表示 d 的中心,$rns \in R$ 表示 d 的非我识别半径,"自我" = (c,rs),c 表示 s 的中心,rs 表示 s 的自我半径。对于一元素 e,如果它接近某一检测器 d 的中心,并且到该检测器中心的欧几里德距离 $dist(c,e) = \left(\sum_{i=1}^{n}(c_i - e_i)^2\right)^{1/2} < rns$,那么该元素被视为"非我",否则为"自我"。

VDetector 则是一种基于可变大小检测器的实值负选择算法。随着算法的进行,检测器的半径是动态改变的,如图 1(非我空间由不带圆心的圆区域覆盖,自我表示为带有圆心的圆区域)。如果生成的检测器的数目达到一定阈值,或者检测器覆盖的非我区域满足预先制定的阈值,算法将终止执行。

虽然,通过实验证明负选择算法有比较好的检测性能,但是无论是实值负选择算法还是 VDetector 都存在以下缺点。

1)正常样本是从训练数据中随机选择的,如果选取数量小的话,正常样本很难反映正常模式的空间特征;如果选取太多的话,又会导致计算量的增加。

2)会产生很多无用的检测器,效率较低。如图 2(a)中的点及图 2(b)中的等半径的圈表示自我区域,图 2(a)中的圈及图 2(b)中的不等半径的圈表示检测器集。可见,实值负选择和 VDetector 产生的检测器间有很多重复,生成了很多无用

收稿日期:2007-12-10;**修回日期**:2008-02-25。
基金项目:天津市自然科学基金资助项目(05YFJMJC05700);河北省自然科学基金资助项目(F2006000109)。
作者简介:李娜娜(1980-),女,河北保定人,讲师,博士研究生,主要研究方向:智能优化算法; 赵政(1949-),男,天津人,教授,博士,主要研究方向:数据库、网络; 刘伯颖(1979-),男,河北保定人,助教,硕士研究生,主要研究方向:智能信息处理; 顾军华(1966-),男,天津人,教授,博士,主要研究方向:智能信息处理。

的检测器。

3)不能够检测出异常的种类。负选择算法只能检测出数据是正常还是异常,而无法判断出属于哪种异常。

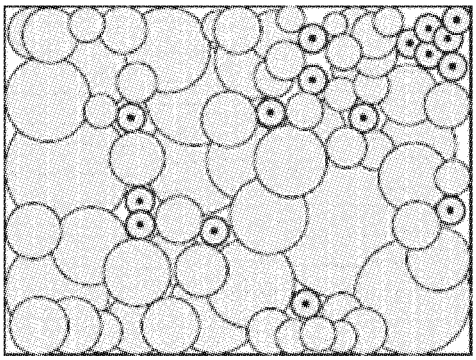
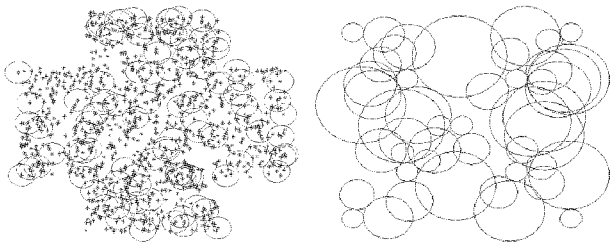


图 1　可变检测器的实值负选择算法在二维空间上的应用



(a) 实值负选择算法产生的检测器　(b) VDetector算法产生的检测器

图 2　生成的检测器集

1.2　基于实值正选择的异常检测

文献[12 – 13]中的实值正选择算法也将所有状态均定义在一个 $[0,1]^n$ 超矩形中,自我元素被定义为自我检测器,r_s 为自我识别半径。对于一元素,如果它接近某一检测器 s 的中心,并且到该检测器中心的欧几里德距离小于 r_s,那么该元素被视为“自我”,否则为“非我”。这意味着,该方法不需要通过学习产生非我检测器,只需要利用自我样本即可进行检测。但是,自我样本数量较多时,该方法的运算时间呈指数增长。

随后,文献[14]指出该算法仅适用于自我样本较少的情形。然而在很多实际应用中,自我样本的数量很多,根本没有办法直接使用正选择方法进行异常检测,此外该正选择方法不具备任何学习能力,严格地说,它不是一种学习算法。

2　基于自我聚类的异常检测

虽然,在实际应用中自我样本的数量很多,不适于使用实值正选择异常检测算法,但是这些数据中包含很多重复的或相似的信息,我们可以通过学习算法抽取核心信息,精简自我样本集合。鉴于以上原因,本文提出一种基于正选择的异常检测方法,即自我聚类法。现有的基于聚类的异常检测主要是对既包括正常又包括异常的数据进行聚类,偏离较远的即视为异常,很少有人通过对自我聚类找自我的模式进行检测。设自我集合为 S,它的每一个元素都是一个 n 维向量。

该方法的流程如下:

1) 使用 K-means 算法对自我数据集 S 进行聚类;

2)在每一类中选取代表样本;

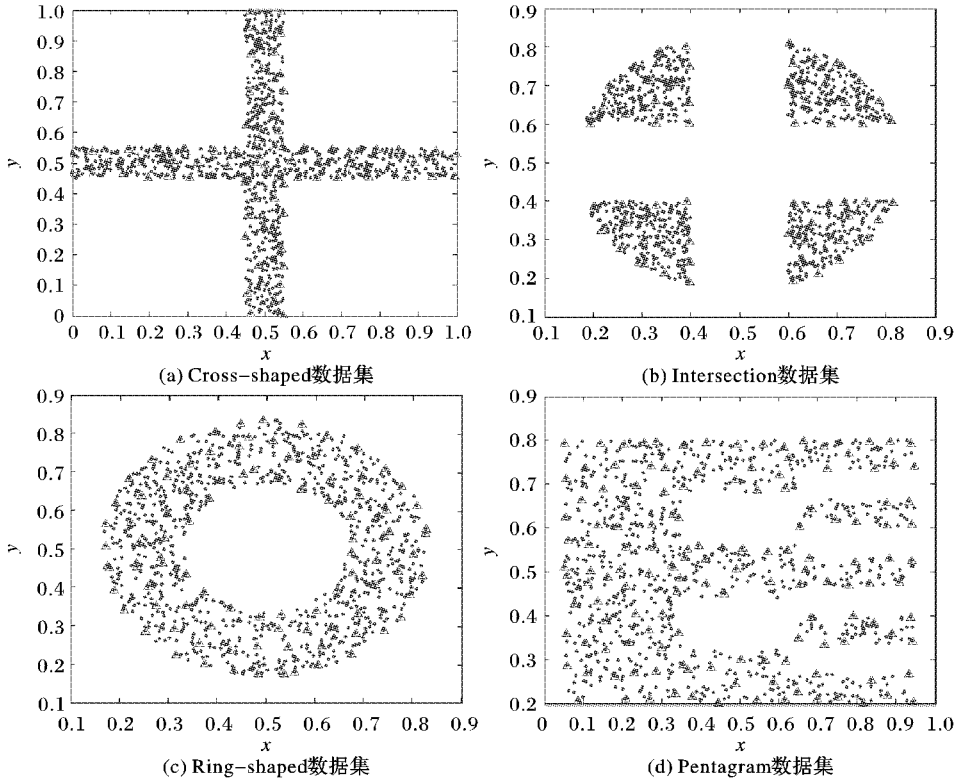
3)根据自我样本构造自我检测器集;

4)通过正选择算法进行异常检测。

2.1　K-means 聚类

使用 K-means 方法将自我集合按照特征相似度进行聚类,选取每一类中的代表性样本,使得该代表性样本集能够反映原数据集的特征。采用的方法是从聚类中心每隔距离 d 取 r 个样本,并通过相似度控制“扎堆”现象的发生,使得选取的样本能够均匀的分布在每一类中,避免了负选择中随机选择样本的弊端。以文献[8]中使用的测试数据为例,经过聚类后选择得到的样本集合如图 3 所示。

从图 3 可见,根据聚类结果选取的样本均匀地分布在正常空间内,能够反映正常空间的特征,有很好的代表性,避免了随机选择正常样本带来的弊端。



(a) Cross-shaped数据集　(b) Intersection数据集

(c) Ring-shaped数据集　(d) Pentagram数据集

注: 圆点表示正常训练集; 三角表示通过聚类选择的样本集。

图 3　测试数据集

2.2 自我检测器

聚类后选取所得的样本是 S 的子集，表征已知的正常数据。本文使用超球形结构对这些代表性自我样本构造检测器，称之为自我检测器。

本文用超球形表示自体的结构模型，球心表示的是样本集中的自体个体，半径是对应的灵敏度，灵敏度越大，球半径越大，即每个个体覆盖的空间越大。灵敏度半径用来描述一个正常个体允许与训练集中元素偏差的距离。

有了自我检测器集，便可以通过正选择的方法对待测个体进行检测。将待测个体与自我检测器集进行比较，如果该个体与检测器集中某个自我检测器的偏差小于灵敏度，则该个体是正常的。如果与任何一个检测器的距离都大于灵敏度，则该个体是异常的。实际上，为了实现对一个数据的异常程度的多级判断，灵敏度半径往往被设计成一组数值，以与自体个体偏差的距离来判断一个数据的异常程度。

2.3 复杂度分析

基于自我聚类检测法的复杂度构成主要包括两部分：一是源于 K-means 聚类，这部分的复杂度是 $O(tkn)$，其中，t 表示循环次数，k 表示聚类的个数，n 表示样本个数；另一部分源于样本的选取，该部分复杂度大小与参数 r 成正比，要选取的代表性样本 r 越多，复杂度越大。然而，一般情况下，r 都远小于 n，故该方法的复杂度主要来源于聚类操作，复杂性较小。

综上所述，该方法实现简单，有较小的复杂度。而且，通过聚类得到的自我检测器集有效地避免了上面提及的 VDetector 的缺点。由于正常样本是通过聚类从训练数据中选择的，它们既能够反映正常模式的空间特征，又不会产生很多无用的检测器，提高了效率。

3 实验分析

本文使用文献[8]中的 Ring 数据集和 biomedical 数据集作为测试数据。为了测试本方法的性能，计算检测率 DR 和误报率 FA 两大指标值，其定义如下：

$$DR = \frac{TP}{TP + FN} \tag{1}$$

$$FA = \frac{FP}{FP + TN} \tag{2}$$

其中，TP(True Positives) 被定义为异常的异常元素，TN(True Negative) 定义为正常的正常元素，FP(False Positives) 定义为异常的正常元素，FN(False Negatives) 定义为正常的异常元素。

图4为在 Ring 数据集上的检测结果，训练数据集为在自我空间随机生成的1000个点，通过聚类选择80个样本点，重复运行此算法30次得到的检测率和误检率变化趋势。从图中可以看出，半径越小，检测率越高，误检率也越高；半径越大，检测率和误检率都随之降低。与 VDetector 在此数据集上的检测结果比较，可发现该方法在误检率为0的同时，检测率仍然高达85%，是一种有效的异常检测算法。

同时，将本文提出的算法作用于 biomedical 数据集，得到如下的结果。biomedical 数据集来自于对209个人的四种血液测量值，其中，75个人携带一种疾病，另外134个是正常人。我们将134个正常人的数据作为自我训练集，75个病人的数据作为测试集。

从图5中也可以看出，自我半径 r_s 是影响检测率的一个重要因素。另外，该算法有较高的检测率和较低的误检率。而 VDetector 算法最高检测率仅约为94%，误检率却高达90%。并且，该方法在 $r_s = 0.1$ 时，误检率便减至为0，检测率仍为60%以上，优于 VDetector 算法。

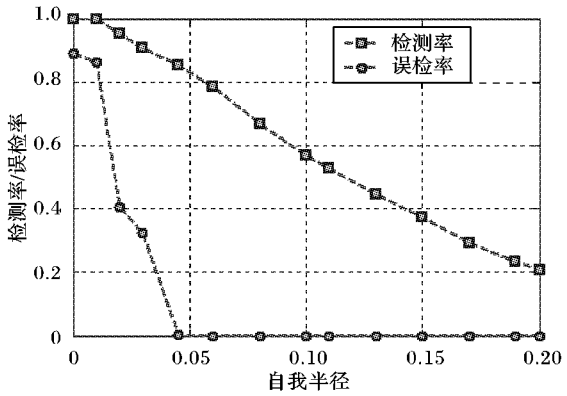


图4 Ring 数据集检测率和误报率

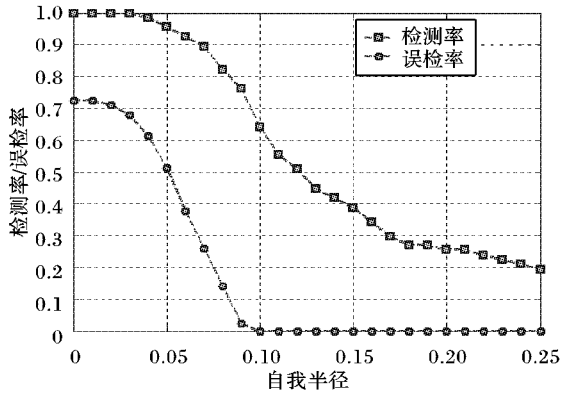


图5 biomed 数据集检测率和误报率

4 结语

基于负选择和正选择的异常检测是目前免疫机制在异常检测应用研究中的一个热点问题。本文提出了一种新的基于正选择的异常检测方法，它既能适用于自我样本集较多的情形，克服了实值正选择检测算法的局限；又具备了一定的学习能力；同时，该方法还避免了负选择中随机选择样本带来的弊端。通过实验分析，它比 VDetector 具备更好的检测性能，是一种有效的异常检测方法。但是，通过实验分析自我半径 r_s 是影响检测率的一个重要因素，该方法没有提出较好的途径解决 r_s 的选择问题，这是以后研究的一个重点。

参考文献：

[1] 张凤斌. 基于免疫遗传算法的入侵检测技术研究[D]. 哈尔滨：哈尔滨工程大学，2005.

[2] 肖娜，龚薇. 基于可达邻域的异常检测算法[J]. 计算机工程，2007，33(17)：74－76.

[3] SHON T，MOON J. A hybrid machine learning approach to network anomaly detection [J]. Information Sciences，2007，177(18)：3799－3821.

[4] 蒋盛益，姜灵敏. 一种高效异常检测方法[J]. 计算机工程，2007，33(7)：166－168.

[5] PATCHA A，PARK J-M. An overview of anomaly detection techniques：Existing solutions and latest technological trends [J]. Computer Networks，2007，51(12)：3448－3470.

[6] GONZALEZ F，DASGUPTA D，KOZMA R. Combining negative selection and classification techniques for anomaly detection [C]// Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002). Washington：IEEE Press，2002，1：705－710.

(下转第1474页)

2.3 关于测试序列生成法则的讨论

定义8 待测实体t和规范s,如果任何一个s的合法输入/输出序列,对于t而言也是合法的,则称t和s是基于输入输出一致的(本文中简称一致)。

定义9 对于待测实体t、规范s和测试套(测试用例的集合)T,

T是完备(complete)的:t和s是一致的,当且仅当t passes T;

T是合理(sound)的:如果t和s是一致的,那么t passes T;

T是完全(exhaustive)的:如果t passes T,那么t和s一致[2]66。

定理 使用前述测试序列生成法则生成的测试用例具有如下性质:

1)一个根据生成法则得到的测试用例,对于规范而言是合理的;

2)包含了所有能够通过生成法则获得的测试用例的集合是完全的。

证明

1)不含任何动作的测试用例总是合理的,下面对测试用例t的结构进行归纳:

(1)由选择1得到的测试用例总是合理的,

(2)对于选择2,如果使用下一个递归得到的测试用例是合理的,那么由选择2得到的测试用例也是合理的。

(3)对于选择3:

a)如果待测实体没有任何输出,也即输出动作是δ-stop,测试终止,赋以判决pass,它是合理的;

b)如果待测实体给出一个非法的输出(不为规范所说明),那么赋以一个判决fail,它是合理的;

c)如果待测实体给出的是一个合法输出,那么它的合理性与下一个递归一致。

2)假设有一个s的一个合法的输入/输出动作序列,那么这个动作序列必然能为生成法则所生成(选择2中对每个选择分别生成测试序列及选择3中考察所有输出动作保证了这点),因而在测试执行中,会检查待测实体是否可以通过这个测试序列。 证毕。

3 结语

本文主要讨论了如何使用RSL来基于输入/输出动作对协议进行形式化描述,它以主干进程为核心,以输入、输出动作为基本元素,能够较好地描述协议的功能规范。由于RSL的广谱特点,用它所书写的主干进程可以是抽象级别很低、面向实现的,这点将为协议的实现带来便利。

在形式化描述的基础上,文中还研究了一种测试用例的自动生成方法。这一方法要求协议是完备的,为了满足这一要求,需要在主干进程的描述中添加δ-stop动作。使用文中给出的测试用例生成法则,得到的测试用例将是合理的和完全的。

为了使这一测试用例生成方法能更好地工作,还有两点需要作进一步研究:1)使用此方法得到的测试用例会有一定的冗余度,有必要研究并设计规则对其作进一步精简,从而简化测试过程;2)在实际测试中,需要能够更加科学设定空输出的等待时间。

志谢:本文的研究和写作工作得到了南京邮电大学计算机学院王汝传教授、中国科学技术大学计算机系赵保华教授的悉心指导,在此表示诚挚的感谢!

参考文献:

[1] The RAISE Language Group. The RAISE Specification Language [M]. New Jersey: Prentice Hall, 1992.

[2] TRETMANS J. Conformance testing with labelled transition systems: Implementation relations and test generation [J]. Computer Networks and ISDN Systems, 1996, 29(1):49-79.

[3] 龚正虎. 计算机网络协议工程[M]. 长沙:国防科技大学出版社,1993.

[4] 顾翔,邱建林,蒋峥峥. RSL在协议形式化描述中的应用研究[J].计算机应用,2007,27(9):150-152.

[5] 吴建平,尹霞. 基于形式化方法的协议测试理论[J]. 清华大学学报:自然科学版,2001, 41(4/5):203-208.

[6] BOSIK B S, UYAR M U. Finite state machine based formal methods in protocol conformance testing: From theory to implementation [J]. Computer Networks ISDN System, 1991, 22(1):7-33.

[7] 顾翔,赵保华,屈玉贵. 通信顺序进程的扩充及其在协议形式化技术中的应用[J]. 通信学报,2004,25(2):43-50.

(上接第1440页)

[7] GONZALEZ F A, DASGUPTA D, NINO L F. A randomized real-valued negative selection algorithm [C]// Proceedings of the 2nd International Conference on Artificial Immune Systems (ICARIS), LNCS 2787. Berlin: Springer-Verlag, 2003: 261-272.

[8] ZHOU JI, DASGUPTA D. Real-valued negative selection algorithm with variable-sized detectors [C]// Proceedings of Genetic and Evolutionary Computation Conference (GECCO 2004), LNCS 3102. Berlin: Springer-Verlag, 2004: 287-298.

[9] ZHOU JI, DASGUPTA D. Augmented negative selection algorithm with variable-size detectors [C]// IEEE Congress on Evolutionary Computation (CEC 2004). Washington: IEEE Press, 2004, 1: 1081-1088.

[10] ZHOU JI. A boundary-aware negative selection algorithm [C/OL]// Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing (ASC 2005). Benidorm, Spain: ACTA Press, 2005 [2007-10-26]. http://www.zhouji.net/prof/asc2005.pdf.

[11] ZHOU JI, DASGUPTA D. Estimating the detector coverage in a negative selection algorithm [C]// Proceedings of the 2005 Conference on Genetic and Evolutionary Computation Conference (GECCO 2005). New York: ACM Press, 2005, 1: 281-288.

[12] STIBOR T, MOHR P, TIMMIS J, *et al.* Is negative selection appropriate for anomaly detection? [C]// Proceedings of the 2005 Conference on Genetic and Evolutionary Computation Conference (GECCO 2005). New York: ACM Press, 2005, 1: 321-328.

[13] STIBOR T, TIMMIS J, ECKERT C. A comparative study of real-valued negative selection to statistical anomaly detection techniques [C]// International Conference on Artificial Immune Systems (ICARIS). Berlin: Spring-Verlag, 2005: 262-275.

[14] ZHOU JI, DASGUPTA D. Applicability issues of the real-valued negative selection algorithms [C]// Proceedings of 8th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO 2006). New York: ACM Press, 2006: 111-118.