·开发研究与工程应用· 文章编号：1000-3428(2020)06-0303-05 文献标志码：A 中图分类号：TP18，TP309

基于检测器集层次聚类的否定选择算法

王韫烨，孔 珊

（郑州师范学院 信息科学与技术学院，郑州 450044）

摘 要： 传统的否定选择过程需要将全部检测器与测试数据进行匹配以排除异常数据，该匹配过程需要花费大量时间，导致检测效率过低。为此，提出一种基于检测器集层次聚类的否定选择算法。对生成的检测器进行层次聚类，减少需要计算距离的检测器数量，不再将与检测器不匹配的数据标记为正常数据，而是基于该数据与自体集和检测器集距离的计算结果将其标记为正常数据或异常数据。实验结果表明，与 V-detector 算法和免疫实值否定选择算法相比，该算法的检测效率显著提高，误检率明显降低。

关键词： 异常检测；检测器集；否定选择算法；层次聚类；检测效率

开放科学（资源服务）标志码（OSID）：

中文引用格式： 王韫烨，孔珊. 基于检测器集层次聚类的否定选择算法[J]. 计算机工程，2020，46(6)：303-307.

英文引用格式： WANG Yunye，KONG Shan. Negative selection algorithm based on hierarchical clustering of detector set[J]. Computer Engineering，2020，46(6)：303-307.

Negative Selection Algorithm Based on Hierarchical Clustering of Detector Set

WANG Yunye，KONG Shan

（College of Information Science and Technology，Zhengzhou Normal University，Zhengzhou 450044，China）

【Abstract】 The traditional negative selection process takes a long time to match all detectors with test data to eliminate abnormal data，resulting in low detection efficiency. Therefore，this paper proposes a negative selection algorithm based on hierarchical clustering of the detector set. The number of detectors that need to calculate the distance is reduced by hierarchical clustering of the generated detectors. The data that does not match the detector is no longer directly marked as normal data，but is marked based on the calculation results of the distance between the data and the self-set and the detector set. Experimental results show that compared with the V-detector algorithm and the real-valued negative selection algorithm of immunity，the proposed algorithm significantly improves the detection efficiency and reduces the false detection rate.

【Key words】 anomaly detection；detector set；negative selection algorithm；hierarchical clustering；detection efficiency

DOI： 10.19678/j.issn.1000-3428.0055114

0 概述

异常检测是在网络和大数据安全分析中广泛使用的关键技术。由于异常检测系统与人体免疫系统有着高度的相似性，基于免疫机制的否定选择算法在异常检测中得到了深入应用并取得良好的效果[1]，成为解决异常检测问题的主要算法之一。检测器的生成是否定选择算法中的重点部分，对检测效率和准确度具有重要影响[2]。因此，设计高效的检测器生成算法是否定选择算法研究的关键和热点[3]。

否定选择算法一般用字符串（含二进制字符串）和实值表示[4]。实值否定选择算法将自体与检测器的各项属性值表示为 n 维[0，1]实数范围（$[0,1]^n$）内的超立方体，更适合对现实问题进行描述，因而得到广泛应用[5]。由于半径固定的实值否定选择算法存在许多黑洞区域无法被检测器覆盖和检测，文献[6]提出 V-detector 算法，该算法是一种半径可变的实值否定选择算法，可以显著提高算法的检测率，为实值否定选择算法中最具代表性的算法，众多后续研究与应用都基于该算法展开[7-16]。

文献[7]提出改进的否定选择算法，该算法增加了检测器的覆盖半径。文献[8]通过二次否定过程提高了检测器生成性能。文献[9]基于自体分布由远及近分层次产生检测器，优化了否定选择算法的性能。

基金项目： 国家自然科学基金(61572447)；河南省科技攻关计划项目(162102310238)。

作者简介： 王韫烨(1980—)，女，讲师、硕士，主研方向为网络安全、数据异常检测；孔 珊，硕士。

收稿日期： 2019-06-04 **修回日期：** 2019-07-25 **E-mail：** zz_paper@126.com

文献[10]在小型样本空间中分离自体与非自体空间,提高了检测器的检测效率。文献[11]通过分析抗原空间的密度,提升了实值否定选择算法的性能。文献[12]采用基于子空间密度搜索的改进否定选择算法提高了空间覆盖率。文献[13]对否定选择算法性能参数进行分析,指出各参数对检测性能的影响。文献[14]将主动学习和否定选择算法进行集成后应用于垃圾邮件分类,增加了垃圾邮件分类准确性。文献[15]将改进的否定选择算法用于故障检测,提升了检测率。文献[16]用否定选择算法生成测试数据,有效提高了测试数据路径覆盖率。

上述否定选择算法均取得了较好的效果,显著提高了检测器的生成效率。但采用否定选择算法对数据进行异常检测时,需要对所有的检测器进行距离计算,这降低了检测效率,增加了检测时间,不利于快速检测。

本文提出一种基于检测器集层次聚类的否定选择算法,对检测器集进行从上到下的层次聚类处理,只计算待检测数据与聚类中心检测器之间的距离,以减少计算时间和能耗,对未被检测器匹配的数据进行分类,并分别从检测率、误检率及检测时间等方面将本文否定选择算法与 V-detector 算法和免疫实值否定选择算法进行对比分析。

1 算法关键技术

1.1 否定选择算法的检测过程

否定选择算法的检测过程由两个阶段构成:一是训练阶段,即检测器的生成阶段;二是检测阶段,即将待测数据通过检测器按照是否正常进行分类的阶段[17]。

在检测阶段,待检测数据需要与检测器进行匹配,因此,需要匹配的检测器数量越少,对数据做出异常检测的判断就越快。

V-detector 算法及其改进算法检测阶段的过程为[6-11]:对任一需要检测的数据,计算其与所有检测器的距离,若待检测数据被任一检测器覆盖,则该待检测数据被标记为异常,否则该待检测数据被标记为正常。对于 n 维空间中待检测数据 $t=(c_t,r_t)$ 而言,将其与检测器 $d=(c_d,r_d)$ 之间距离记为 $\mathrm{dis}=(t,d)$,其中 c_t 为待检测数据的中点,r_t 为待检测数据的半径,c_d 为检测器的中点,r_d 为检测器的半径。为判断待检测数据是否异常,需要计算各待检测数据 $t_i(t_i\in t)$ 与所有的检测器 $d_i(d_i\in d)$ 之间的距离,其计算公式如下:

$$\mathrm{dis}(c_t,c_d)=\sqrt{(c_{t_1}-c_{d_1})^2+(c_{t_2}-c_{d_2})^2+\cdots+(c_{t_n}-c_{d_n})^2}=\sqrt{\sum_{i=1}^{n}(c_{t_i}-c_{d_i})^2} \tag{1}$$

若 $\mathrm{dis}(c_t,c_d)<r_d$,则该数据被标记为异常,否则该数据被标记为正常。因为该算法是在二维实值空间建立仿真实验,因此 $n=2$。

由否定选择算法的检测过程可以看出,该过程计算会耗费较多的计算时间,这不利于该算法的广泛应用。

1.2 改进的否定选择算法

本文在上述否定选择算法的基础上进行改进,改进后算法的主要思想为:对检测器集 D 进行从上到下的层次聚类处理,只计算待检测数据 t 与聚类中心检测器 C 之间的距离,不再计算待检测数据 t 与所有聚类成员检测器集 D 之间的距离,减少了计算时间和能耗。

1.3 检测器集的层次聚类

本文设计的检测器层次聚类过程为:

1)对否定选择过程生成的成熟检测器集 D 进行层次聚类处理,每层生成的聚类中心构成新的检测器集 C。

2)对于每个检测集 T 中的数据 $t=(c_t,r_t)$,分别计算其与检测器集 C 中每个检测器 C_i 的距离,从而判断该待检测数据是否正常并进行分类。其中,c_t 为待检测数据的坐标,c_c 为检测器 C_i 的坐标,r_{ci} 为检测器 C_i 的半径。具体的判断方法为:若 $\mathrm{dis}(c_t,c_c)<r_{ci}$,则说明待检测(分类)数据被检测器集 C 中的某一检测器覆盖,该检测数据就被标记为异常数据。

在二维实值空间 $[0,1]^2$ 进行仿真验证如下:

设定首层聚类半径 $r_1=\sqrt{2}$;第 2 层聚类半径 $r_2=r_1/2$;第 $n+1$ 层聚类半径 $r_{n+1}=r_n/2$;最小化聚类数据与聚类中心的距离和为 $\min\sum_{n=1}^{m}\sum_{d\in D}\mathrm{Dis}(c_n,d)$,其中,$d$ 为待聚类的检测器,m 为总聚类层数,c_n 为每层聚类中心。当聚类半径小于检测器集 D 中最小检测器的半径时,此聚类过程结束。

由上述可见,层次聚类方法中聚类半径逐渐减半,从而使得检测更精确,在减少检测时间的同时提高了检测率。

1.4 孔洞数据的判定

在二维实值空间 $[0,1]^2$ 中,孔洞区是指没有被任何检测器覆盖的检测区域[18]。对于处于孔洞区

的待检测(分类)数据,可以通过否定选择算法将其分类为正常数据。然而未被检测器覆盖的区域有可能存在异常数据并导致分类错误,造成算法误检率升高[19]。在改进后的否定选择算法中,不再将待检测数据简单标识为正常数据,其性质由检测器与自体集的位置共同定量决定,即待检测数据的异常性由该数据到最近检测器的欧氏距离和最近自体的欧氏距离共同判定,分别表示为:

$$\mathrm{Dis1} = \min \mathrm{dis}(t, c_i), c_i \in C \quad (2)$$

$$\mathrm{Dis2} = \min \mathrm{dis}(t, s_i), s_i \in S \quad (3)$$

其中,Dis1 为待分类数据 t 到检测器集合 C 中所有检测器 c_i 的最短距离,Dis2 为待分类数据 t 到自体集合 S 中所有自体 s_i 的最短距离。若 Dis1 > aDis2,则该待检测数据为正常数据,否则为异常数据。大量实验研究结果表明,a 的取值为自体半径 r_s 的 20 倍[18]。

2　算法流程

改进后否定选择算法的检验过程分为检测器生成阶段与异常检测阶段,该算法的基本步骤如下:

1)使用文献[6]中 V-detector 算法生成成熟检测器集 D。

2)对检测器集 D 进行层次聚类处理,得到检测器集 C。

3)输入数据集 T 进行异常检测(计算 T 中数据和检测器集 C 中检测器之间的距离。

4)进行异常检测判断:如果被检测数据与检测器集 C 中的任一检测器匹配,则为异常数据,直接进行第 6 步;如果被检测数据未与检测器集 C 中的任一检测器匹配,进行第 5 步。

5)如果被检测数据未与检测器集 C 中的任一检测器匹配,则认为其处于孔洞区,需进一步测试以判定其是否为正常数据并进行标识。

6)算法结束。

3　实验结果与分析

在 Windows 10 环境下,使用 JAVA 对本算法进行编程实现。实验数据集为广泛使用的人造二维数据集,对实验结果以五角星形自体集为例进行分析,并与经典的 V-detector 算法[6]和免疫实值否定选择算法[9]结果进行对比。参数取值与所对比的文献保持一致:数据取自二维实值空间$[0,1]^2$,训练数据集容量为 1 000 个,测试数据集容量为 1 000 个,目标覆盖率分别为 90% 、95% 和 99% 。分别从检测率、误检率及检测时间三方面对实验结果进行分析。

选择实验目标覆盖率为 95% ,自体半径的取值区间为[0.01,0.30](每隔 0.05 取一个值),每个自体半径进行 100 次实验并取结果的平均值。

由图 1 可以看出,3 种不同算法的检测率均随着自体半径的增大而降低,本文否定选择算法的检测率降幅比其他两种算法更小。这是因为孔洞区数据随着自体半径的增大而增加,本文否定选择算法由于对孔洞区数据的异常性进行了判定,提高了孔洞区数据检测的准确性,从而提升了算法的检测率。

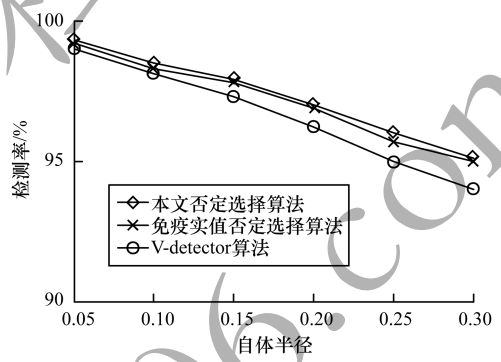

图 1　检测率随自体半径变化曲线

Fig. 1　Curve of detection rate changing with autogenous radius

由图 2 可以看出,3 种不同算法的误检率均随着自体半径的增大而降低,本文否定选择算法误检率的降幅比其他两种算法的更大。这是因为孔洞区数据随着自体半径的增大而增加,在 V-detector 算法和免疫实值否定选择算法中,孔洞区数据(部分数据可能为异常数据)全部被定义为正常数据,而本文否定选择算法对孔洞区数据异常性进行了判定,有效降低了算法的误检率。

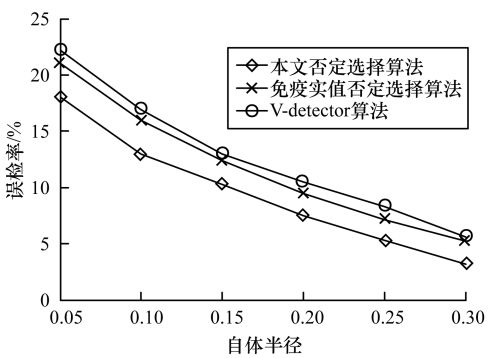

图 2　误检率随自体半径变化曲线

Fig. 2　Curve of false detection rate changing with autogenous radius

由图 1 和图 2 可知,检测性能与自体半径密切相关,随着自体半径的增大,算法的检测率和误检率均降低。因此,需根据实际问题选择适合的自体半径来调整算法的检测性能。

由图 3 和图 4 可以看出,当自体半径为 0.2 时,3 种算法的检测率和误检率均随着目标覆盖率的增大而增加;和其他两种算法相比,本文否定选择算法的检测率更高且误检率更低;当目标覆盖率为 99% 时,3 种算法的检测性能较接近,这是因为当目标覆盖率为 99% 时,所需成熟检测器的数量显著增加,对非自体区域的覆盖增大,处于孔洞区域的数据减少,此时本文否定选择算法的优势不明显。

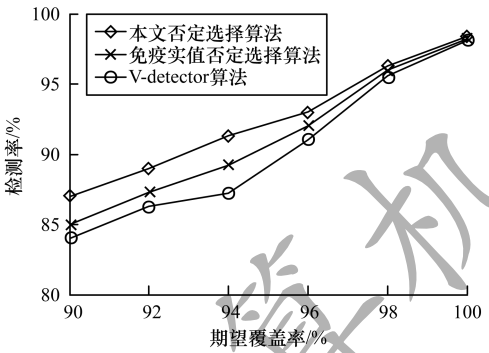

图 3 检测率随期望覆盖率变化曲线

Fig. 3 Curve of detection rate changing with expected coverage rate

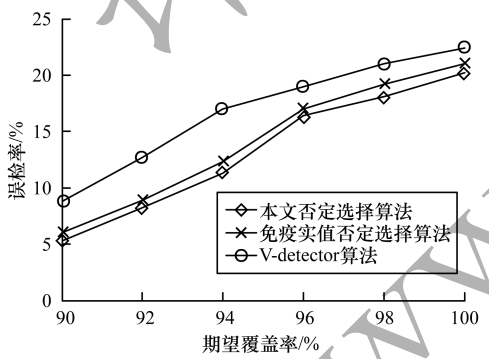

图 4 误检率随期望覆盖率变化曲线

Fig. 4 Curve of false detection rate changing with expected coverage rate

由表 1 可以看出,V-Detector 算法在采用不同形状检测器时检测时间均最长,免疫实值否定选择算法的次之,本文、否定选择算法最短;当检测器形状为环形时,本文否定选择算法的检测时间最短。这是因为 V-Detector 算法由于需要计算各待检测数据与检测器集 D 中所有检测器之间的距离以判断数据是否异常,因而所需的检测时间最长。

表 1 不同算法的检测时间结果对比

Table 1 Comparison of detection time results of different algorithms s

算法种类	条形	十字形	五角星形	环形
V-Detector 算法	7.34	7.41	7.22	6.72
典型改进否定选择算法	6.74	6.38	6.44	5.81
本文改进否定选择算法	5.34	5.58	5.79	4.21

免疫实值否定选择算法采用了检测器分级生成方式,在同样的检测率下,需要的检测器数量减少,因而所需的检测时间比 V-Detector 算法要短。本文否定选择算法由于用聚类中心检测器 C 代替检测器集 D 进行距离计算(C 的数量远小于 D),因而所需的检测时间最短。圆形检测器与环形自体集的形状较相似,匹配度较高,此时孔洞区待检测数据较少,因而检测时间最短。

4 结束语

为提高异常检测的效率,本文提出一种基于检测器集层次聚类的否定选择算法,通过对检测器集进行层次聚类,以聚类中心检测器代替初始检测器集进行距离计算,减少计算时间并对未被检测器覆盖的孔洞区域属性进行进一步判断。实验结果表明,本文否定选择算法较 V-detector 算法和免疫实值否定选择算法所需时间大幅减少,检测效率提高且误检率降低。下一步将在本文算法的基础上对聚类中心集和检测器集的数量关系进行研究。

参考文献

[1] DE ABREU C C E,DUARTE M A Q,VILLARREAL F. An immunological approach based on the negative selection algorithm for real noise classification in speech signals[J]. International Journal of Electronics and Communica-tions,2017,72:125-133.

[2] LI Dong,LIU Shulin,ZHANG Hongli. A boundary-fixed negative selection algorithm with online adaptive learning under small samples for anomaly detection[J]. Engineering Applications of Artificial Intelligence,2016,50(2):93-105.

[3] CHAI Zhengyi,WANG Xianrong,WANG Liang. Real-value negative selection algorithm for anomaly detection[J]. Journal of Jilin University(Engineering and Technology Edition),2012,42(1):176-181.(in Chinese)
柴争义,王献荣,王亮. 用于异常检测的实值否定选择算法[J]. 吉林大学学报(工学版),2012,42(1):176-181.

[4] JIN Zhangzan, LIAO Minghong, XIAO Gang. Survey of negative selection algorithms[J]. Journal on Communications, 2013,34(1):159-170. (in Chinese)
金章赞,廖明宏,肖刚. 否定选择算法综述[J]. 通信学报,2013,34(1):159-170.

[5] FOULADVAND S, OSAREH A, SHADGAR B. DENSA: an effective negative selection algorithm with flexible boundaries for self-space and dynamic number of detectors[J]. Engineering Applications of Artificial Intelligence,2017,62:359-372.

[6] ZHOU Ji, DASGUPTA D. V-detector: an efficient negative selection algorithm with "probably adequate" detector coverage[J]. Information Sciences, 2009, 179(10):1390-1406.

[7] CHAI Zhengyi, WU Huixin, WUYong. Optimization algorithm for immune real-value detector generation[J]. Journal of Jilin University(Engineering and Technology Edition),2012,42(5):1251-1256. (in Chinese)
柴争义,吴慧欣,吴勇. 用于异常检测的免疫实值检测器优化生成算法[J]. 吉林大学学报(工学版),2012,42(5):1251-1256.

[8] ZHENG Xufei, FANG Yonghui, LI Tao. Dual negative selection algorithm[J]. Scientia Sinica Informationis, 2013,43(4):529-544. (in Chinese)
郑旭飞,方永慧,李涛. 二次否定选择算法[J]. 中国科学:信息科学,2013,43(4):529-544.

[9] XIAO Xin,LI Tao,ZHANG Ruirui. An immune optimization based real-valued negative selection algorithm[J]. Applied Intelligence,2015,42(2):289-302.

[10] LI Dong, LIU Shulin, ZHANG Hongli. A negative selection algorithm with online adaptive learning under small samples for anomaly detection[J]. Neurocomputing,2015,149:515-525.

[11] YANG Tao, CHEN Wen, LI Tao. An antigen space density based real-value negative selection algorithm[J]. Applied Soft Computing,2017,61:860-874.

[12] LIU Zhengjun, LI Tao, YANG Jin, et al. An improved negative selection algorithm based on subspace density seeking[J]. IEEE Access,2017,5:12189-12198.

[13] CHEN Wen,LI Tao. Parameter analysis of negative selection algorithm[J]. Information Sciences,2017,420:218-234.

[14] HU Xiaojuan, LIU Lei, QIU Ningjia. A novel spam categorization algorithm based on active learning method and negative selection algorithm[J]. Acta Electronica Sinica,2018,46(1):203-209. (in Chinese)
胡小娟,刘磊,邱宁佳. 基于主动学习和否定选择的垃圾邮件分类算法[J]. 电子学报,2018,46(1):203-209.

[15] ABID A, KHAN M T, DE SILVA C W. Layered and real-valued negative selection algorithm for fault detection[J]. IEEE Systems Journal,2018,12(3):2960-2969.

[16] MOHI-ALDEEN S M, MOHAMAD R, DERIS S. Application of Negative Selection Algorithm (NSA) for test data generation of path testing[J]. Applied Soft Computing,2016,49:1118-1128.

[17] HE Jun. Research on detector generation mechanism in real negative selection algorithm[D]. Tianjing: Tiangong University,2019. (in Chinese)
何君. 实值否定选择算法中检测器生成机制研究[D]. 天津:天津工业大学,2019.

[18] GONG Maoguo, ZHANG Jian, MA Jingjing, et al. An efficient negative selection algorithm with further training for anomaly detection[J]. Knowledge-Based Systems,2012,30(2):185-191.

[19] ALIZADEH E, MESKIN N, KHORASANI K. A negative selection immune system inspired methodology for fault diagnosis of wind turbines[J]. IEEE Transactions on Cybernetics,2017,47(11):3799-3813.

编辑　宋　圆