基于切割的检测器生成与匹配算法

蔡 涛,鞠时光,仲 巍,牛德姣
(江苏大学计算机学院,江苏镇江 212013)

摘 要: 检测器生成和匹配算法直接影响到人工免疫系统的检测效率和非法抗原的检测率.为了改进现有算法存在的生成检测器与识别非法抗原的时间和空间开销较大、对非法抗原检测率较低等问题,本文提出基于切割的检测器生成与匹配算法.针对现有检测器表示方法存在的缺陷,用正超立方体表示检测器,为减少匹配算法的时间和空间开销提供了基础;依据空间包含关系设计基于空间包含的匹配算法,减少了选择检测器和检查抗原的时间和空间开销,使得分析检测器所覆盖的非法抗原较方便;依据自体在论域空间的分布,引入切割空间的方法生成检测器,消除所生成检测器间的冗余信息,减少了检测漏洞,使得所生成的检测器具有较高的非法抗原检测率和检测效率.文中建立了算法的原型系统,构造不同类型的数据集,测试识别非法抗原所需的检测器数量,以及当系统中保存不同数量的检测器时所具有的非法抗原检测率,与现有算法进行比较,验证了基于切割的检测器生成与匹配算法能有效的提高否定选择算法的性能.

关键词: 人工免疫算法; 检测器生成算法; 匹配算法; 否定选择算法; 信息安全
中图分类号: TP301 **文献标识码**: A **文章编号**: 0372-2112 (2009) 4A-131-04

A Cutting Based Detector Generating and Matching Algorithm

CAI Tao, JU Shi-guang, ZHONG Wei, NIU De-jiao
(*Department of Computer, Jiangsu University, Zhenjiang, Jiangsu* 100084, *China*)

Abstract: The detector generating and matching algorithm is important for efficiency and accuracy of the artificial immune system. In order to reduce the time and space consumption of detector generating, and improve efficiency and validity of detector, we present a cutting based detector generating and matching algorithm. Using super cube to present detector and providing the basis of reducing matching algorithm overhead. Using the spatial inclusion relationship to express the matching relationship between detector and antigen, then reducing time and space consumption of matching and in favour of analyzing accuracy. Cutting space to generate detector by location of self in domain, elemiating redundance between detectors, reducing hole of inspecting, and ensuring high efficiency and accuracy of inspecting antigen. Realizing prototype to test and compare with current algorithms. The results prove cutting based detector generating and matching algorithm can improve the performance of the artificial immune system.

Key words: artificial immune algorithm; detector generating algorithm; matching algorithm; negative selection algorithm; information security

1 引言

当前计算机安全系统的时间和空间开销较大,影响了被保护系统的性能,因此研究时间和空间开销较小的安全机制是一个重要的问题.人工免疫算法模拟生物免疫系统抵抗病毒和细菌等病原体的机制,具有耐受性、分布性、鲁棒性、适应性、多样性、免疫反馈和自组织性等特性[1],能有效地减少计算机安全系统的时间和空间开销. Forrest 于 1994 年提出的否定选择算法[2,3]是经典的人工免疫算法,NSA 通过模拟 T 细胞成熟机制选择成熟检测器,再使用成熟检测器区别自体和非自体,达到检测异常信息的目的.初始检测器的生成算法和检查检测器是否与自体及抗原匹配的匹配算法是决定否定选择算法性能的关键算法,也是人工免疫算法研究中的热点问题.

当前否定选择算法中主要使用二进制字符串[1~5]或空间向量[6,7]表示自体、检测器和抗原.使用二进制字符串表示形式时,主要的初始检测器生成算法有穷举法、线性法和贪心法. Forrest 于 1994 年提出了穷举法[1],使用枚举所有可能检测器的方法生成初始检测器,只能依据是否于自体匹配筛选成熟检测器有效性的机制,算法的时间和空间开销大,且无法保证初始检测器对非自

收稿日期:2008-10-22;修回日期:2008-12-23
基金项目:国家自然科学基金(No. 60773049);江苏省自然科学基金(No. BK2007086)

体的覆盖率,易造成检测漏洞. Forrest 等于 1994 年提出了线性初始检测器生成算法[8,9],改善生成初始检测器的有效性和效率,但算法的时间和空间开销仍与论域的大小成指数关系,同时检测器中会包含较多冗余的信息. Forrest 又提出贪心初始检测器生成算法,消除了部分冗余,但仍没有优化算法的时间开销[10]. 上述三种算法中,由于使用二进制字符串自体、检测器和抗原,难以直观的表示自体、检测器和抗原在论域中的分布,给判断检测器对非法抗原的覆盖率和减少冗余检测器带来了较大的困难,算法效率较低. Gonzalez 和 Dasgupta 将 N 维空间作为人工免疫算法的论域[6,7],用空间超球体表示自体、检测器和抗原,能直观的描述自体、检测器和抗原以及它们之间的关系,有利于优化初始检测器的生成,但空间超球体的大小固定,易造成检测漏洞;Ji Zhou 和 Dasgupta 提出以可变半径的空间超球体表示自体、检测器和抗原,改进了算法的检测效率和非法抗原检测率[11];但以上两个算法没有利用自体在论域空间的分布信息,优化生成初始检测器,以保证检测器对非法抗原的覆盖率,减少检测器间的冗余信息. 在国内张衡等于 2005 年提出了 r 可变阴性选择算法,提高了检测器对非法抗原的覆盖率[12]. 2007 年中国科技大学的何申和罗文坚提出使用可变长度的检测器,使用 h 叉满树,判断串之间的包含关系,消除人工免疫算法中的检查漏洞,提高了检测效率[13]. 这两种算法中均使用二进制字符串表示自体、检测器和抗原,存在难以判断检测器对非法抗原的覆盖率及查找冗余检测器困难等问题. 因此利用自体在论域空间的分布信息,研究一种自适应的检测器生成算法,是提高人工免疫算法的检测效率和对非法抗原的检测率的重要方法.

使用二进制字符串表示形式时,主要的匹配算法包括:1994 年 Forrest 提出的 r-contiguous 匹配算法[1]、2002 年 Balthrop 提出的 r-chunk 匹配算法[14]和 Farmer 在 1996 年提出的 Hamming 距离匹配算法[15]及其多种变形等. r-contiguous 匹配算法和 r-chunk 匹配算法中,统计自体与检测器及检测器与抗原之间对应连续相同子串的长度,如超过系统设置的匹配阈值,则判断它们两者匹配;Hamming 距离匹配算法统计检测器与抗原之间相同对应位的个数,如超过系统设置的匹配阈值则判断它们匹配. 这些匹配算法的时间和空间开销较大,此外存在难以分析检测器对非法抗原的覆盖情况的问题,很难保证人工免疫算法对非法抗原的检测率和检测效率. 使用空间向量表示形式时,主要的匹配算法是基于欧氏距离的匹配算法[6]及其一系列变形,当计算出的距离值不大于系统设置的匹配阈值时,判断自体与检测器或检测器与抗原匹配,这些算法的时间和空间开销同样较大. 因此改变自体、检测器和抗原的表示方式,研究时间和空间开销小且便于分析检测器对非法抗原覆盖率的匹配算法,是人工免疫系统中研究的重点问题.

我们提出新型的检测器和抗原表示方法,利用检测器与自体和抗原间的空间包含关系,设计新型的匹配算法;提出基于切割的检测器生成算法,依据自体在论域空间中的位置切割获得检测器;并实现原型系统,测试和分析算法的性能.

2 元素的定义

我们首先给出相关元素的定义如下.

定义 1 论域空间 X:N 维中的一个封闭空间,是人工免疫算法所讨论的范围,包含所有的抗原和检测器.

定义 2 抗原 g:是人工免疫算法的检测对象,我们用 N 维正超立方体表示,$g=(X_1, X_2, \cdots, X_i, \cdots, X_n, r)$,$i=1,2,\cdots,n$,其中 r 为正超立方体边长的一半,X_i 是正超立方体的球心在第 i 维上的坐标值. 抗原包含自体和非自体两类.

定义 3 自体 S:人工免疫系统中合法抗原的集合,$S \subseteq X$.

定义 4 非自体 NS:人工免疫系统中非法抗原的集合,$NS \subset X$.

根据以上定义,推论一必然成立.

推论 1 $S \cap NS = \varnothing$ 且 $S \cup NS = X$

定义 5 检测器 d:用于检查抗原的合法性,我们用 N 维超立方体表示,超立方体的每个边均平行或垂直于坐标轴,$d=((X_1^{\min}, X_1^{\max}), \cdots, (X_i^{\min}, X_i^{\max}), \cdots, (X_n^{\min}, X_n^{\max}))$,$i=1,2,\cdots,n$,其中 $X_i^{\min}$ 和 $X_i^{\max}$ 表示超立方体在第 i 维上投影的最小和最大坐标值.

3 基于空间包含的匹配算法

匹配算法用于选择检测器和判断抗原的合法性,直接影响到人工免疫算法的性能. 现有匹配算法存在计算复杂和难以分析对非法抗原的覆盖率等问题. 在 N 维的论域空间 X 中,我们用正超立方体表示抗原,用超立方体表示检测器;图 1 中给出选择二维空间作为论域空间时,论域空间、抗原和检测器的示意.

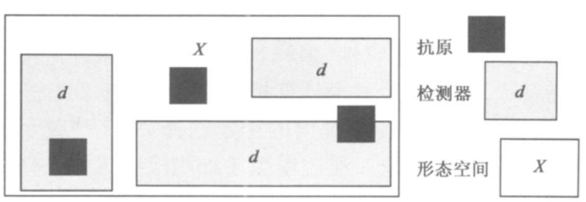


图1 论域空间、抗原和检测器的示意图

从图 1 中我们可以发现,抗原与检测器在空间位置上存在是否包含的关系,我们以表示检测器的超立方

体是否包含表示抗原的正超立方体,作为判断检测器是否与抗原匹配的标准,设计基于空间包含的匹配算法,匹配算法的定义如式(1)所示.

$$Match(d,g)=\begin{cases}\text{匹配}, f(d,g)=n\\ \text{不匹配}, \text{otherwise}\end{cases} \tag{1}$$

$f(d,g)$是表示检测器的超立方体与表示抗原的正超立方体在 N 个轴上的投影存在重叠的次数,由于表示检测器的超立方体各边均平行或垂直于坐标轴,计算函数 f 的值时,仅需依次比较正超立方体和立方体在各个轴上投影线端点的坐标,即可得出表示检测器的超立方体与表示抗原的正超立方体在 N 维空间中的包含关系,如式(2)所示.

$$f(d,g)=\sum_{i=1}^{n}\begin{cases}0, \text{otherwise}\\ 1, dX_i^{\min}-r<gX_i<dX_i^{\max}+r\end{cases} \tag{2}$$

4　基于切割的检测器生成算法

检测器生成算法首先随机的生成初始检测器,再删除与自体匹配的初始检测器构建检测器集;由于未利用自体在论域空间的分布信息,无法优化检测器对非法抗原的覆盖率,并减少检测器间的冗余信息. 在 N 维空间中表示检测器和抗原时,它们之间的空间包含关系为分析检测器能否覆盖绝大部分的非法抗原提供了基础. 我们根据检测器和抗原之间的空间包含关系,改变穷举的初始检测器生成策略,依据自体在论域空间中的位置信息,引入切割空间的方式,提出基于切割的检测器生成算法.

依据自体在 N 维空间中的位置,切割论域空间,生成检测器,再检查所获得的检测器是否于其余的自体匹配,如匹配则依据与检测器发生匹配的自体在论域空间中的位置信息,再次切割检测器,重复上述构成直到获得不与任何自体匹配的检测器. 切割检测器时首先选择空间容量较大的检测器进行切割,使得生成的检测器能匹配较多的非法抗原.

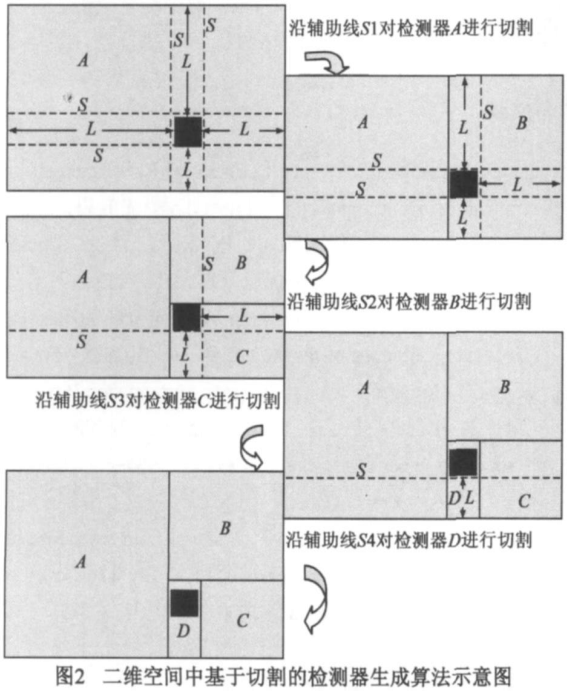


图2　二维空间中基于切割的检测器生成算法示意图

以二维空间为例,图2给出基于切割检测器生成算法的过程. 可以看出 $L1$ 长度最长,因此沿辅助线 $S1$ 对检测器 A 进行切割,生成 A 和 B 两个检测器;由于 $L2>L3>L4$,分别沿辅助线 $S2$、$S3$、$S4$ 对检测器进行切割,生成 A、B、C、D 四个检测器;再检查这些检测器是否与自体匹配,如匹配则重复上述步骤,直到所有检测器均不与自体匹配.

5　原型系统的测试

我们在Linux平台上用C语言实现了基于切割的检测器生成与匹配算法的原型系统(RVDD),为方便分析和比较算法的性能,以二维空间作为原型系统的论域空间,论域空间在每个轴上投影的取值范围是区间 $[0,16]$,表示自体的超球体和正超立方体的半径 $r=1$,使用多种类型的数据集作为原型系统的自体集,如表1所示,测试算法的性能.

表1　不同类型的自体集

自体集的类型	特性描述
随机型(RD)	自体随机分布于论域空间内.
正态型(ND)	自体正态分布于论域空间内.
多簇型(Multi-Cluster)	自体分布于论域空间中的几个部分.
圆环型(Ring)	自体按环型分布于论域空间,具有内外半径.

此外建立 real_value 算法(RV)和可变 real-value 算法(RVVD)的原型系统. 分别构建包含 10～160 个自体的自体集,测试对非法抗原的检测率为 90 %时,所需的检测器数量,比较不同算法的检测效率,测试结果如图3所示.

从图3可知,RVDD中当自体数量较少时,识别非法抗原所需要的检测器数量与自体数量成正比,自体数量的增加会使得论域空间被切割成更多的检测器,识别非法抗原所需的检测器数量也就随之上升. 当自体数量继续增加后,所需要的检测器数量缓慢下降,此时由于自体数量增加使得较多的自体聚集起来,减少了获得的检测器数量. RV 和 RVVD 中识别非法抗原所需的检测器数量一直与自体数成反比,自体数量较少时,由于检测器识别能力的限制,需要大量的检测器识别较多的非法抗原;自体数量增加后,非法抗原数量减少,所需检测器数量也呈下降趋势. 分析图3中结果可知,当自体数量较少时,RVDD中识别非法抗原所需的检测器数量远少于 RV 和 RVVD,能有效地提高人工免

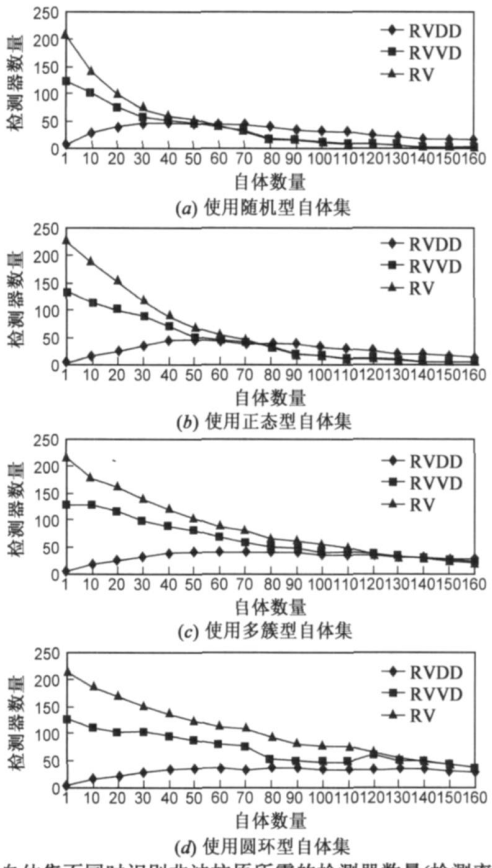


图3 自体集不同时识别非法抗原所需的检测器数量(检测率:90%)

疫算法检测抗原的效率;自体数量增加后,RVDD、RV 和 RVVD 中识别非法抗原所需的检测器数基本相同. 当自体数量变化时,RVDD 中识别非法抗原所需的检测器数量变化很小,有效地改善了 RV 和 RVVD 中识别非法抗原所需的检测器数量变化较大的问题,保证了识别非法抗原所需的时间和空间开销基本稳定.

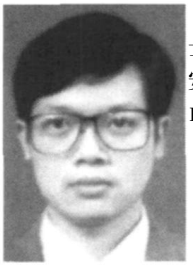
由图 3 可知,在 RVDD 中自体数量较少时,RD-Hash 和 NC-Hash 型自体集中自体的分布密度较小,需要的检测器较多,如图 3(*a*) 和 3(*b*) 所示. Multi-Cluster 和 Ring 型自体集中,自体数较少时,自体的分布密度较大,识别非法抗原所需检测器数量较少,如图 3(*c*) 和图 3(*d*) 所示. 相比 RV 和 RVVD,RVDD 能显著减少自体集类型为 Multi-Cluster 和 Ring 型时识别非法抗原所需的检测器数量,明显降低时间和空间开销;能显著减少自体集类型为 RD-Hash 和 NC-Hash 型且自体数量较少时所需的检测器数量,明显减少时间和空间开销.

6 结束语

本文使用所有边与轴平行或垂直的超立方体表示检测器,为减少匹配算法的时间和空间开销打下了基础. 设计基于空间包含的匹配算法,以表示抗原的正超立方体与表示检测器的超立方体之间的空间包含关系,作为判断检测器是否与抗原匹配的标准,由于仅需统计正超立方体和超立方体在各轴上的投影存在重叠的次数即可判断两者是否匹配,算法的时间和空间开销小. 再给出基于切割的检测器生成算法,根据自体在论域空间的分布信息,切割空间生成检测器,消除了检测器间的冗余信息,减少了检测漏洞. 最后实现了算法的原型系统,构建不同的测试环境,测试了算法的非法抗原检测率和检测效率,并与现有算法进行分析和比较,验证了基于切割的检测器生成与匹配算法能有效地提高人工免疫算法检测抗原的效率,并能以更小的时间和空间开销获得更高的非法抗原检测率.

由于限制检测器数量,算法仍然存在误差和检测漏洞. 下一步准备分析影响算法的时间和空间开销、以及非法抗原检测率的因素,引入博弈论设计检测效率和非法抗原检测率的优化机制.

作者简介:



蔡 涛 男,1976 年 3 月出生于上海市,博士研究生、讲师. 主要研究领域为人工免疫系统、安全存储系统.
E-mail:caitao@ujs.edu.cn

鞠时光 男,1955 年出生,博士、教授. 主要研究领域为信息安全、存储系统.
E-mail:jushig@ujs.edu.cn

参考文献:

[1] Forrest S, Perelson S A, Allen L, Cherukuri R. Self-nonself discrimination in a computer[A]. Proceedings of IEEE Symposium on Research in Security and Privacy[C]. Los Alamitos, CA: IEEE Computer Society Press, 1994. 202 - 212.

[2] Forrest S, Hofmeyr S A, Somayaji A. Computer immunology [J]. Communications of the ACM. 1997, 40(10): 88 - 96.

[3] Forrest S, Hofmeyr S A. Immunology as Information Processing [M]. Segel and Coheneds. Design Principles for the Immune System and Other Distributed Autonomous Systems. USA: Oxford University Press, 2000.

[4] Esponda, F, Forrest S, Helman P. A formal framework for positive an negative detection scheme [J]. IEEE Transaction on Systems, Man, and Cybernetics, 2004, 34(1): 357 - 373.

(下转第 86 页)

[15] Y Cheng. Mean shift, mode seeking, and clustering[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790 - 799.

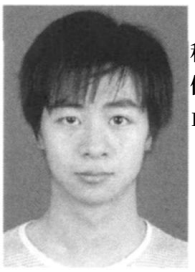
[16] M Black, D Fleet, Y Yacoob. Robustly estimating changes in image appearance [J]. Computer Vision and Image Understanding, 2000, 78(1): 8 - 31.

[17] L Liu, J Wang, X Chen, Y Guo, Q Peng. A robust and fast non-local means algorithm for image denoising[J]. Journal of Computer Science and Technology, 2008, 23(2): 270 - 279.

[18] 易翔, 王蔚然. 一种概率自适应图像去噪模型[J]. 电子学报, 2005, 33(1): 63 - 66.
Yi Xiang, Wang Wei-ran. A probability model for adaptive image denoising[J]. Acta Electronica Sinica, 2005, 33(1): 63 - 66. (in Chinese)

[19] 杨晓慧, 焦李成, 李伟. 基于第二代 bandelets 的图像去噪[J]. 电子学报, 2006, 34(11): 2063 - 2067.
Yang Xiao-hui, Jiao Li-cheng, Li Wei. Second generation bandelets based image denoising [J]. Acta Electronica Sinica, 2006, 34(11): 2063 - 2067. (in Chinese)

作者简介:



王 想 男, 1984 年出生. 南京大学计算机科学与技术系硕士研究生, 从事多媒体技术、图像视频处理等方面的研究.
E-mail: wangxiang@graphics.nju.edu.cn

郭延文 男, 1980 年出生, 博士, 讲师, 主要研究方向为计算机图形学、图像和视频处理.
E-mail: ywguo@nju.edu.cn

(上接第 134 页)

[5] Gonzalez, F, Dasgupta D, Gomez J. The effect of binary matching rules in negative selection[A]. Proceedings of Genetic and Evolutionary Computation Conference (CECCO)[C]. Chicago: Springer Berlin/ Heidelberg, 2003, Volume 2723. 195 - 206.

[6] Gonzalez, F, Dasgupta D, Nino F L. A randomized real-valued negative selection algorithm[A]. Proceedings of 2nd International Conference on Artificial Immune System (ICARIS)[C]. UK: Springer-Verlag, 2003. 261 - 272.

[7] Gonzalez, F, Dasgupta D. Anomaly detection using real-valued negative selection[J]. Journal of Genetic Programming and Evolvable Machines, 2003, 4(4): 383 - 403.

[8] D'haeseleer P. Further efficient algorithms for generating antibody strings[R]. Technical Report CS95-3, The University of New Mexico, Albuquerque, NM, 1995.

[9] Helman P, Forrest S. An efficient algorithm for generating random antibody strings[R]. Technical Report CS-94-07, The University of New Mexico, Albuquerque, NM, 1994.

[10] D'haeseleer P, Forrest S, Helman P. An immunological approach to change detection: algorithms, analysis and implications[A]. J. McHugh and G. Dinolt, editors, Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy[C]. USA: IEEE Press, 1996. 110 - 119.

[11] Ji Zhou, Dasgupta D. Real-valued negative selection using variable-sized detectors[A]. Proceedings of International Conference on Genetic and Evolutionary Computation (GECCO)[C] Seattle, WA, 2004, June26 - 30.

[12] 张衡, 吴礼发, 张毓森, 曾庆凯. 一种 r 可变阴性选择算法及其仿真分析[J]. 计算机学报, 2005, 28(10): 1614 - 1619.

[13] 何申, 罗文坚, 王煦法. 一种检测器长度可变的非选择算法[J]. 软件学报, 2007, 18(6): 1361 - 1368.

[14] Balthrop J, Forrest S, Glickman R M. Revisting LISYS: parameters and normal behavior[A]. Proceedings of the 2002 Congress on Evolutionary Computation CEC2002[C]. USA: IEEE Press, 2002. 1045 - 1050.

[15] Farmer D J, Packard H N, Perelson S A. The immune system, adaptation, and machine learning[J]. Physical D, 1986, 22(1 - 3): 187 - 204.