一种用于病毒检测的协作免疫网络算法

程春玲[1,2],柴 倩[1],徐小龙[1],张登银[2]
(1.南京邮电大学计算机学院,江苏南京 210003;2.江苏省无线传感网高技术研究重点实验室,江苏南京 210003)

摘 要: 提出一种用于病毒检测的协作免疫网络算法,通过不同类型免疫细胞之间的激励与协作优化免疫网络中的检测器.算法引入非我集,通过成熟检测器对非我集的适应度对成熟检测器克隆选择,加强对抗体的激励作用;通过进化代数更新变异步长来自适应的改变成熟检测器的变异方式;并根据整个免疫网络中抗原对抗体以及抗体之间的激励作用提出基于浓度分区的网络抑制策略.实验结果表明,算法通过增加记忆检测器的多样性,有效地提高了免疫网络的病毒检测能力.

关键词: 病毒检测; 人工免疫; 免疫网络; 协作进化
中图分类号: TP309.5 **文献标识码:** A **文章编号:** 0372-2112 (2013) 12-2518-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.12.032

A Cooperative Immune Network Algorithm for Virus Detection

CHENG Chun-ling[1,2], CHAI Qian[1], XU Xiao-long[1], ZHANG Deng-yin[2]
(1. *College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*;
2. *Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, Jiangsu* 210003, *China*)

Abstract: A cooperative immune network algorithm used for virus detection is proposed, in which detectors are optimized through cooperation and incentive between different kinds of immune cells. The non-self set is introduced, and mature detectors are selected and cloned according to the fitness between detectors and non-self set to enhance the incentive for antibodies. The mutation step is updated by the generation number of evolution to change the mutation way of mature detectors adaptively. Furthermore, the network suppression strategy based on concentration partition is proposed according to the incentive between antibodies and antigens in the entire immune network. Experimental results show that the proposed algorithm can increase the diversity of detectors, and improve the virus detection ability of the entire immune network effectively.

Key words: virus detection; artificial immune; immune network; cooperative evolution

1 引言

计算机网络的普及为网络病毒的发展和肆虐提供了便捷的环境,研究快速准确的病毒检测方法对及早发现和处理病毒至关重要.目前,病毒检测技术主要包括特征代码法、校验和法、病毒签名检测法、行为监测法和基于人工智能的方法等.特征代码法通过匹配病毒特征码发现已知病毒,算法简单,但不能检测未知及变种病毒;校验和法通过检查文件的校验和来检测,能发现未知病毒,但对隐蔽性病毒无效且误报率较高;病毒签名检测法搜索是否存在病毒嵌入的特殊标记,但须预先已知病毒签名的位置和内容;行为监测法监测病毒的行为模式,但抽取行为模式有一定难度,且误报率较高.近年来基于人工智能的新型病毒检测方法受到广泛关注,该方法通过人工智能技术抽取病毒特征并自动检测病毒,例如基于多贝叶斯分类器的方法[1]、基于支持向量机的方法[2]、基于神经网络的方法[3]和基于人工免疫的方法[4,5]等.基于贝叶斯分类器和基于神经网络的方法需要完备的数据集及较长的训练时间才能达到较高的检测效果;基于支持向量机的方法适用于小规模的训练样本;而基于人工免疫的方法模拟生物免疫系统,能快速准确识别"自体"和"非自体",为检测病毒提供了一个新途径.

免疫学中占主导地位的是 Burnet 的克隆选择学说和 Jerne 的免疫网络学说[5].克隆选择学说通过外来抗原选择处于静止状态的免疫细胞进行克隆变异,各个

收稿日期:2012-02-18;修回日期:2013-06-01;责任编辑:梅志强
基金项目:国家自然科学基金(No.61071093);江苏省研究生科研创新计划(No.CXZZ12_0483);江苏省科技支撑计划(No.BE2012849);江苏高校优势学科建设工程(No.YX002001)

免疫细胞是离散的,忽略了整体识别能力;而免疫网络理论认为免疫细胞不仅受外来抗原的刺激,细胞之间也要相互刺激和协调形成一个动态的免疫网络来完成免疫功能,更能体现免疫系统的动态性能.典型的免疫网络模型是资源受限人工免疫系统 RLAIS[6]和 aiNet 人工免疫网络[7].RLAIS 首先构造与生物免疫系统中 B 细胞功能类似的人工识别球,然后通过一些识别球及其联系构成免疫系统[6].aiNet 将免疫细胞连接为一个相互作用的网络,根据抗体与抗原之间的亲和力来选择抗体克隆变异并进行网络抑制,形成稳定的免疫系统[7].但是,aiNet 主要依赖亲和力抑制来降低检测器间的冗余度,使抑制阈值极大地影响了 aiNet 算法的免疫网络结构和大小.为了降低抑制阈值对 aiNet 的影响,文献[8]提出了矢量距免疫网络聚类算法 VD-aiNet,采用矢量距来评估抗体性能,增加了群体多样性.文献[9]引入禁忌算法,提高了 aiNet 算法的收敛性.文献[10]提出了一种多目标免疫网络优化算法,集中了抗体群在克隆变异过程中的有利信息来自适应地指导变异方向;文献[11]引入启发算法对 aiNet 进行优化,并通过刺激下一代得到问题的多样解.此外,文献[12]采用多模糊回归树的方法来优化 aiNet 以减少最终解的错误率,利用梯形函数模糊化的模糊推理策略获得最终的免疫细胞.

以上算法都对 aiNet 进行了改进,但细胞的进化主要通过克隆选择过程实现,细胞间仅表现出生存竞争关系,这种细胞间各自独立的进化方式导致整个网络的进化速度慢,不适用于快速有效的检测病毒.本文提出一种用于病毒检测的协作免疫网络算法,通过多种免疫细胞之间的激励和相互协作快速优化免疫网络,增加记忆检测器的多样性,从而有效地提高了整个免疫网络的病毒检测能力.

2　协作免疫网络算法

协作免疫网络算法由如下三个阶段构成:(1)构建免疫网络阶段:采用基于基因库的检测器生成算法来提高成熟检测器生成效率和免疫网络的构建效率.(2)协作进化阶段:设计成熟检测器的选择、分级克隆、自适应变异和基于浓度分区的网络抑制策略不断优化免疫网络中的记忆检测器,最终形成具有自稳态调节的免疫网络.(3)病毒检测阶段:利用进化后的记忆检测器检测病毒.

2.1　构建免疫网络

免疫网络是由成熟检测器构成的带权、不完全连接的无向图,定义为 $\boldsymbol{G}=\langle V,E,W\rangle$,其中节点集 V 是成熟检测器的集合;边集 E 表示检测器之间的连接;W 是 E 的权重集,用检测器之间的亲和力表示.检测器 x_i 与自我(非我)元素 x_j 的亲和力定义为:

$$f_{x_ix_j}=\begin{cases}1, & R/l\geqslant\sigma\\ 0, & \text{otherwise}\end{cases}\tag{1}$$

其中,1 表示 x_i 识别了 x_j,0 表示不识别.σ 为匹配阈值,$0\leqslant\sigma\leqslant1$,$l$ 为检测器链长,R 为 x_i 与 x_j 连续匹配的位数.

由于非我集中已经收集了异常信息,采用基于基因库的检测器生成算法可以提高成熟检测器的生成效率、降低运算开销.构建免疫网络算法描述见算法 1.

算法 1　构建免疫网络算法 Construct-ImmuneNet($\boldsymbol{M}$)

1. 利用非我集 $\boldsymbol{N}$ 初始化基因库 $\boldsymbol{J}$;
2. 成熟检测器集 $\boldsymbol{C}=\varnothing$,免疫网络 $\boldsymbol{M}=\varnothing$;
3. while(Count(C) < T)do　　// T 为生成成熟检测器的个数
   从 $\boldsymbol{J}$ 中随机选择基因片段组合得到未成熟检测器 t;
   $\forall s\in\boldsymbol{S}$,根据式(1)计算 f_{ts};　// $\boldsymbol{S}$ 为自我集
   if($\forall f_{ts}\leqslant\sigma$) then　$\boldsymbol{C}=\boldsymbol{C}\cup\{t\}$;
4. End while;
5. While($\forall x_i,x_j\in\boldsymbol{C}$)do　　//构建免疫网络
   根据式(1)计算 $f_{x_ix_j}$;
   if($f_{x_ix_j}\geqslant\delta_c$) then
   AddtoImmuneNet($x_i,x_j,\boldsymbol{M}$);
   Connect($x_i,x_j,\boldsymbol{M}$);　//连接 x_i,x_j
   $w_{x_ix_j}=f_{x_ix_j}$;　　//赋边的权值为 $f_{x_ix_j}$;
6. End while;
7. Return($\boldsymbol{M}$);

2.2　免疫网络的协作进化

为了克服 aiNet 算法中免疫细胞间仅为生存竞争关系的缺点,考虑到非我集已收集了异常信息,本文引入非我集,通过抗体对非我集的适应度加强协作;提出激励水平,并给出激励水平与浓度的动态方程,使得各个免疫细胞通过自我识别、相互激励和制约构成一个动态平衡的网络结构.协作进化策略包括:

(1)成熟检测器的选择策略

引入非我集后,通过计算成熟检测器对非我集的适应度增强非我集对抗体的激励作用,删除免疫网络中适应度低即受非我集激励作用小的的检测器;选择适应度高的检测器进行克隆变异,增加记忆检测器的多样性.成熟检测器对非我集的适应度定义如下:

$$\text{fitness}(x_i,\boldsymbol{N})=\sum_{x_j\in\boldsymbol{N}}f_{x_ix_j}\tag{2}$$

$f_{x_ix_j}$为 x_i 与非我集中的元素 x_j 之间的亲和力.

(2)成熟检测器的分级克隆策略

适应度表达了成熟检测器与非我集的亲和力,适应度越高则检测器的检测能力越强,因此根据成熟检测器对非我集的适应度从低到高形成等差级数关系克隆相应的数量.设成熟检测器按适应度从低到高排序

表示为$\{x_1, x_2, \cdots, x_n\}$,则每个被激励的抗体 $x_i(1 \leqslant i \leqslant n)$的克隆数目 s_i 可通过式(3)计算:

$$s_i = \text{round}\left(\frac{n \cdot s_1 - s_n}{n-1} + \frac{s_n - s_1}{n-1} i\right) \tag{3}$$

其中:round()为四舍五入的取整函数.

(3)成熟检测器的自适应变异策略

aiNet 算法的变异方式为:$C^* = C + \alpha N(0,1)$,其中,α 为抗体变异步长,$\alpha = (1/\beta)\exp(-f_i)$,$\alpha \in (0,1]$;$\beta$ 为用户预先设定的参数,$\beta \in [1, +\infty]$.α 控制了进化的平均变异情况,其值设置偏小则算法的收敛速度慢,而过大则不利于检测器向最优检测能力变异.本文设置 β 为进化代数 t 的函数 $\beta = \exp(t)$,在进化初期,β 值小则 α 值大,有利于提高检测器的多样性;而多次迭代后,减小 α 值有利于检测器朝最优方向变异,实现了参数 α 根据进化代数的自适应调整.改进后的变异方式为:

$$C^* = C + \alpha N(0,1)$$
$$\alpha = \frac{1}{\exp(t)}\exp(-f_i); \tag{4}$$

(4)基于浓度分区的网络抑制策略

aiNet 通过亲和力对抗体进行抑制,减少了抗体的多样性.为克服这一缺点,本文提出激励水平和基于浓度分区的网络抑制策略.

定义 1 激励水平 $A_i(t)$:检测器 x_i 在 t 时刻受免疫网络中免疫细胞的激励程度,用免疫细胞对该检测器的亲和力之和来度量.

定义 2 浓度 $\alpha_i(t)$:激励水平 $A_i(t)$在区间(0,1)上的映射.

激励水平和浓度关系的动态方程为:

$$\frac{\mathrm{d}A_i(t)}{\mathrm{d}t} = \left(r\frac{\sum_{x_j \in M} f_{x_i x_j}\alpha_j(t)}{\sum_{x_j \in M} f_{x_i x_j}} + q\sum_{x_k \in T} f_{x_i x_k}\right)\alpha_i(t)$$
$$\alpha_i(t+1) = \frac{1}{1+\exp(0.5 - A_i(t+1))} \tag{5}$$

其中,r 为与记忆检测器 x_i 相连的所有记忆检测器占免疫网络中记忆检测器总数的比率;q 为与 x_i 作用的抗原占抗原总数的比率;$A_i(t)$,$\alpha_i(t)$分别为 t 时刻 x_i 的激励水平和浓度;$\boldsymbol{M}$ 为记忆检测器集;$\boldsymbol{T}$ 为抗原集;$f_{x_i x_j}$ 表示 x_i 与记忆检测器 x_j 之间的亲和力;$f_{x_i x_k}$表示 x_i 与抗原 x_k 的亲和力.

式(5)中浓度的计算考虑了不同免疫细胞之间的相互协作.根据浓度值划分区间将记忆检测器集划分为若干子集,消除每个子集中检测器之间亲和力高于抑制阈值 θ_s,即过于相似的记忆检测器.协作免疫网络进化算法具体描述如算法 2.

算法 2 协作免疫网络进化算法 Co-Evolution($\boldsymbol{M}$)

1. 读取非我集 $\boldsymbol{N}$;
2. Set(Timer); //设置计时器 Timer;
3. While(记忆检测器集 $\boldsymbol{M}$ 中细胞数不足设定的数量或未达到最大迭代次数)
   $\forall x_i \in \boldsymbol{C}$,根据式(2)计算 fitness($x_i$, $\boldsymbol{N}$);
   按适应度对成熟检测器集 $\boldsymbol{C}$ 分级克隆得到 $\boldsymbol{C}'$
   对 $\boldsymbol{C}'$按自适应变异策略进行变异得到集合 $\boldsymbol{C}''$;
   对 $\boldsymbol{C}''$基于浓度分区进行网络抑制得到 $\boldsymbol{M}$
   If(Count($\boldsymbol{M}$) < Num ‖ Timer = 0)then //注入新检测器
   调用基于基因库的成熟检测器生成算法;
   Set(Timer);
4. End while;
5. Return($\boldsymbol{M}$);

2.3 协作免疫网络的病毒检测过程

利用进化稳定的协作免疫网络检测病毒的过程如下:计算协作免疫网络中的所有记忆检测器与样本文件特征码之间的亲和力,若亲和力大于匹配阈值 σ,则该样本文件为病毒;否则,为正常文件.

算法 3 基于协作免疫网络的病毒检测算法 Detection($\boldsymbol{M}$, m)

输入:协作免疫网络 $\boldsymbol{M}$,样本文件特征码 m

1. 设置匹配阈值 σ;
2. 读取协作免疫网络 $\boldsymbol{M}$ 中的记忆检测器;
3. Repeat
   从 M 中取一个记忆检测器 x_i;
   计算 $f_{x_i m}$;
   if($f_{x_i m} \geqslant \sigma$) // $\boldsymbol{M}$ 中存在某个检测器 x_i 识别 m
   then Return 为病毒;
4. Until M 中没有新的检测器;
5. Return 正常文件; //没有检测器识别 m

3 实验及其结果分析

实验采用文献[2]的数据集.数据集中共有 4266 个程序,其中 3265 个恶意程序,1001 个正常程序.该数据集由哥伦比亚大学采集,所有样本均来自于真实的计算机和网络环境.正常程序中绝大多数来自新安装好的 Windows 系统,少量从 Internet 下载;恶意程序收集自几个 FTP 服务器.数据集中的每个程序都通过病毒扫描器被正确地标记.从数据集中提取的实验数据包括自我集 $\boldsymbol{S}$、非我集 $\boldsymbol{N}$ 和测试集 $\boldsymbol{T}$.实验中特征码的二进制串长为病毒检测工业的标准长度 128 位.用 Matlab 7.0 实现本算法和对比算法.

本文的病毒检测主要由免疫网络中的记忆检测器实现,记忆检测器的优化与匹配阈值 σ 有关,因此首先比较了不同匹配阈值 σ 对病毒检测性能的影响.从数

据集中选取 100 个正常程序为自我集,50 个病毒的特征码为非我集,由自我集、非我集、非我集中已知病毒的 50 个变种和 2 个新病毒共 202 个样本组成测试集.本文算法在不同匹配阈值 σ 下的检测率、误检率和漏报率如图 1 所示.

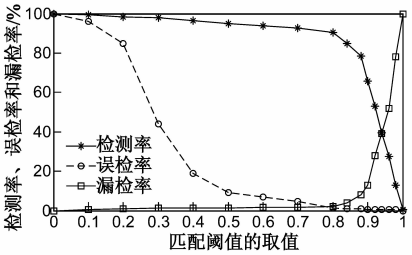

图1　σ取值对病毒检测效果的影响

从图 1 可以看出,σ 取值越小,算法的识别能力越强,但误检率也较大,需要适当提高 σ 值,当 σ 取 0.8 ~ 0.9 之间可在减少误检率的同时保证检测效率;另一方面,增加 σ 虽降低了误检率,但同时也增加了漏检率,当 σ 大于 0.8 后对病毒的误检率影响较小,但漏检率明显增大.从图中结果可见,当 $\sigma = 0.8$ 时,算法的检测效果最为理想.

其次,评估了两种免疫网络算法对记忆检测器的优化效果,控制免疫网络中记忆检测器的数量从 10 到 100 递增,比较了在相同记忆检测器规模下,两种算法检测病毒的正确率,结果如图 2 所示.

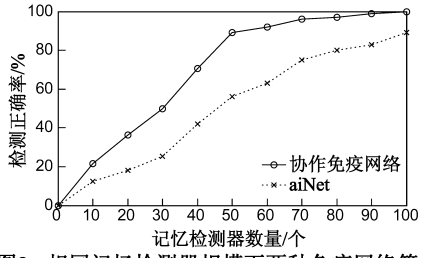

图2　相同记忆检测器规模下两种免疫网络算法检测病毒的正确率比较

从实验结果可以看出,当控制两个免疫算法的记忆检测器数量相同时,本文算法的病毒检测正确率一直明显高于 aiNet,说明本文算法通过引入非我集和加强免疫细胞之间的激励与协作能够优化记忆检测器、增加多样性,从而有效提高病毒检测能力.

最后,比较了本文算法与病毒签名法、多贝叶斯分类器法[1]、支持向量机法[2]和 aiNet 免疫网络算法的病毒检测能力,并采用了与文献[2]相同的评价指标:病毒的检测率、正确率和误检率.实验采用病毒检测中广泛使用的 5 次交叉验证法,本文算法的参数取匹配阈值 $\sigma = 0.8$,记忆检测器个数 $|M| = 500$,实验结果如图 3、4 所示,其中病毒签名法、多贝叶斯方法及支持向量机方法的实验结果来自文献[1,2],而 aiNet 免疫网络及协作免疫网络算法的实验结果由本文实验所得.

从图 3 可以看出,病毒签名法的检测率和正确率均较低;贝叶斯、支持向量机和 aiNet 免疫网络的检测率和正确率基本持平;而本文算法由于在进化过程中引入非我集,并具备较好的协作进化能力,优化了免疫网络中的检测器,使检测率和正确率分别达到 98.84% 和 98.31%,均高于其他方法.

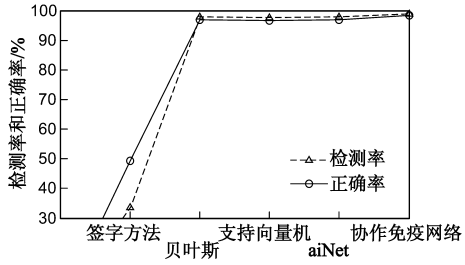

图3　不同检测方法的检测率和正确率

图 4 是不同检测方法对应的误检率.除了病毒签名检测法外,协作免疫网络的误检率为 3.39%,明显低于其他检测方法.因此与其他算法相比,本文提出的基于协作免疫网络的病毒检测算法在提高了检测率及正确率同时降低了误检率.

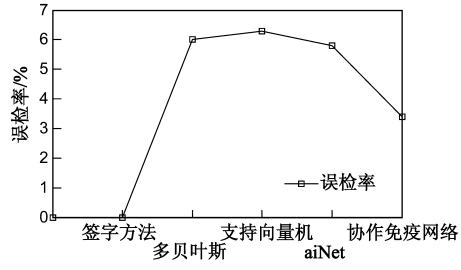

图4　不同检测方法的误检率

4　结束语

本文针对传统 aiNet 免疫网络模型及其改进模型中免疫细胞之间缺乏相互协作的不足,提出了一种用于病毒检测的协作免疫网络算法.通过成熟检测器集与非我集之间的相互协作,选择适应度高的成熟检测器进行克隆变异;根据进化代数自适应的调整检测器的变异方式,使得检测器向多样更优的方向变异;并根据整个免疫网络中抗原对抗体以及抗体之间的激励作用,提出基于浓度分区的网络抑制策略,保证记忆检测器多样性的同时降低冗余度.进化过程中不同免疫细胞之间的激励与协作,使免疫网络中的检测器节点逐渐进化成最优检测器,从而提高整个网络中记忆检测器的病毒检测效率.实验结果表明本文算法具有较好的收敛性,能有效提高整个免疫网络的病毒检测能力.

参考文献

[1] Matthew G S, Eleazer E, Erez Z, et al. Data mining methods for detection of new malicious executables[A]. IEEE Symposium on Security and Privacy[C]. Oakland, CA: IEEE, 2001. 1207 – 1217.

[2] 彭宏,王军.基于支持向量机的病毒程序检测方法[J].电子学报.2005,33(2):276 – 278.
Peng Hong, Wang Jun. Research of malicious executables detection method based on support vector machine[J]. Acta Electronica Sinica, 2005, 33(2): 276 – 278. (in Chinese)

[3] Fabio A. G, Juan C G, Diego A R, et al. A neuro-immune mode for discriminating and visualizing anomalies[J]. Natural Computing, 2006, 5(3): 285 – 304.

[4] Kim J, Bentley P J, Aickelin U. Immune system approaches to intrusion detection--a review[J]. Natural Computing, 2009, 6(4): 413 – 466.

[5] Ya He, Liang Yiwen, Li Tao. A model of collaborative artificial immune system[A]. CAR 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics[C]. Piscataway, NJ: IEEE Computer Society, 2010. 101 – 104.

[6] Gong B L, Junqho I, Giorqos M. An artificial immune network approach to multi-sensor land use/land cover classification[J]. Remote Sensing of Environment, 2011, 2 (15): 600 – 614.

[7] Castro D, Femando J. An evolutionary immune system net-work for data clustering[A]. Proceedings of the IEEE SBRN[C]. Piscataway, NJ: IEEE Computer Society, 2000. 84 – 89.

[8] 杨海东,郭建华,邓飞其.人工免疫系统集成与应用[M].北京:科学出版社,2010. 61 – 70.

[9] 赵云丰,尹怡欣,付冬梅,等.禁忌免疫网络算法及其在函数优化中的应用[J].智能系统学报,2008,3 (5):393 – 400.
Zhao Yun-feng, Yin Yi-xin, Fu Dong-mei, et al. Application of a Tabu immune network algorithm in function optimizations [J]. CAAI Transactions on Intelligent Systems, 2008, 3(5): 393 – 400. (in Chinese)

[10] Coelho G P, Von Zuben F J. A concentration-based artificial immune network for multi-objective optimization[A]. Lecture Notes in Computer Science. Proceeding of the Evolution Muti-Criterion Optimization 6th International Conference [C]. Berlin: Springer, 2011. 343 – 357.

[11] Coelho, G P, de Franga F O, Von Zuben F J. A concentration-based artificial immune network for combinatorial optimization [A]. Proceeding of IEEE Congress of Evolutionary Computation[C]. Piscataway, NJ: IEEE Computer Society, 2011. 1242 – 1249.

[12] Gasir, Fathi. An architecture for constructing fuzzy regression tree forests using opt-aiNet[A]. Proceeding of the IEEE International Conference on Fuzzy Systems[C]. Piscataway, NJ: IEEE Computer Society, 2011. 283 – 289.

作者简介

程春玲 女,1972 出生,陕西西安人.副教授、在职博士生.1989 年和 1993 年在南京理工大学分别获学士和硕士学位.现为南京邮电大学教师,主要从事信息安全、网络管理等方面的研究工作.

E-mail: chengcl@njupt.edu.cn

柴 倩 女,1986 出生,南京邮电大学计算机学院硕士生,主要研究方向为信息安全、计算机网络技术等.