免疫原理在计算机病毒检测中的应用

韦 芸[1,2] 章剑林[2] 徐慧剑[2]
[1](浙江大学计算机科学与技术学院 浙江 杭州 310012)
[2](浙江经贸职业技术学院 浙江 杭州 310018)

摘 要 主要探讨免疫原理在计算机病毒检测中的应用,并实现了一种将否定选择、克隆选择等生物免疫机制应用于传统的基于特征码的计算机病毒检测法。实验表明,该方法具有检测已知病毒和识别病毒的一些未知变种的能力,能够自动提取特征码,并且生成的病毒特征码具有很低的误别率,是一种实用的计算机病毒检测的方法。

关键词 免疫原理 病毒免疫系统 病毒检测 选择算法

THE APPLICATION OF IMMUNE THEORY TO VIRUS DETECTION

Wei Yun[1,2] Zhang Jianlin[2] Xu Huijian[2]
[1](*College of Computer Science and Technology, Zhejiang University, Hangzhou 310012, Zhejiang, China*)
[2](*Zhejiang Economic and Trade Polytechnic, Hangzhou 310018, Zhejiang, China*)

Abstract The Application of immune theory to virus detection is proposed. Some immune mechanisms, such as Negative Selection and Clone Selection, are applied to traditional signature-based virus detection. Experimental result shows that the method can detect known virus and some of their mutations, and it has the ability to extract signature automatically. Moreover, the signature extracted has low false-positive rate. The method is a practical way for virus detection.

Keywords Immune theory Virus immune system Virus detection Selection algorithm

0 引 言

计算机病毒检测是计算机安全领域的一个主要分支。当前使用的病毒检测方法[1]主要有特征码法,校验和法,行为检测法以及软件模拟法等,其中,开销最小、使用最广泛的是特征码法。这种方法需要建立一个病毒特征码库,通过使用这个特征码库来检测病毒。这种方法有两个明显的缺点:一是无法检测多态性病毒;二是新病毒的特征码提取工作需要由病毒专家人工完成。最近几年,网络的飞速发展使病毒的传播更加便捷,简单的病毒编写工具的出现使病毒的书写成为并不困难的事,这些都使人工提取变得越来越不堪重负。本文针对第二个缺点讨论了免疫原理在计算机病毒检测中的应用,提出了一种基于免疫原理的计算机病毒检测方法。该方法具有检测已知病毒和识别病毒的一些未知变种的能力,能够自动提取特征码,并且生成的病毒特征码具有很低的误别率。

1 病毒免疫系统

人体的免疫系统天生具有检测和识别已知和未知病毒的能力。计算机病毒防御与人体免疫有许多相似的地方。人体免疫系统的许多机制[2,3]可以用于计算机病毒检测。本文提出的基于免疫原理的病毒检测方法使用了克隆选择和否定选择等免疫机制[4],涉及的主要概念有自体与非自体、免疫检测细胞、亲和力、克隆选择、否定选择,免疫记忆。系统的功能:提呈一个可疑文件,系统使用克隆选择、否定选择等机制,通过免疫检测细胞来识别病毒,提取并记忆识别病毒的检测细胞(特征码)。

1.1 自体与非自体

生物免疫系统的自体和非自体是形态不同的蛋白质链。在本文的病毒检测系统中自体 S 为正常程序代码集合,$S\subseteq U$;非自体 T 为病毒程序代码集合,$T\subseteq U$。满足 $S\cap T=\Phi, S\cup T=U$。其中 $U=\bigcup_{i=16}^{+\infty} G^i$,代表问题域集合,$G=\{0,1,\cdots,9,a,\cdots,f\}$,代表可选十六进制字符集合。

1.2 检测细胞

检测细胞用长度在 16 位到 24 位之间的十六进制的病毒特征码定义:检测细胞 $b, b\in\bigcup_{i=16}^{24} G^i$。它模拟淋巴细胞,融合了 B 细胞 T 细胞和抗体的性质,用于检测和识别病毒。比如 Tiny_143 病毒的免疫检测细胞就是它的特征码 4db8004ccd21eb029090beff。

1.3 亲和力

设检测细胞 $b=c_1c_2\cdots c_{l_b}$,病毒 $v=g_1g_2\cdots g_{l_v}$,c_i, g_j 为十六进制字符,$1\leqslant i\leqslant l_b$,$1\leqslant j\leqslant l_v$,$l_v>l_b$。那么式(1)定义了 b 对 v 的识别,式(2)定义了 b 和 v 之间的亲和力:亲和力越高,b 和 v 之间越匹配,当亲和力与检测细胞长度(也就是最大的亲和力值)比值达到一定的门限比例时,b 就识别 v。

收稿日期:2006-12-08。浙江省科技计划项目(2007C33071)。
韦芸,硕士生,主研领域:网格技术,管理信息系统,电子商务。

$$f_r(b,v)=\begin{cases}1 & \text{iff } f_{affinity}(b,v)/l_b \geq \lambda_r \\ 0 & \text{otherwise}\end{cases} \quad (1)$$

其中 1 表示 b 识别 v,0 表示不识别。λ_r 为设定的识别门限比率,$0 \leq \lambda_r \leq 1$。l_b 为检测细胞的长度。

$$f_{affinity}(b,v)=\max(a_1,a_2,\cdots,a_{l_v-l_b+1}) \quad (2)$$

其中 $a_i=\sum_{j=1}^{l_b}\delta_{ij}$　$\delta_{ij}=\begin{cases}1 & \text{iff } c_j=g_{i+j-1} \\ 0 & \text{otherwise}\end{cases}$

$1 \leq i \leq l_v-l_b+1$　$1 \leq j \leq l_b$

1.4 否定选择过程

在病毒检测中不能把正常的程序识别成病毒,所以检测细胞在参与病毒检测前必须经历自体耐受过程。检测细胞集合的自体耐受过程用式(3)定义:消除检测细胞集合中所有识别自体的检测细胞。

$$f_{tolerance}(B)=B-\{d \mid d \in B \wedge \exists y \in S(f_r(d,y)=1)\} \quad (3)$$

其中 S 为自体集合,$B \subseteq \{b \mid b \in \bigcup_{i=16}^{24} C^i\}$ 为接受耐受训练的检测细胞集合,f_r 为识别函数,如式(1) 所示。

1.5 克隆选择过程

生物免疫系统通过克隆选择过程来识别未知抗原。本文的病毒检测系统同样采用克隆选择原理来识别变种病毒。克隆选择算法如下:

```
procedure clone_selection {
    设置初始检测细胞集合 S(忆检测细胞集合);
  while (S 中的所有检测细胞都不能识别病毒且克隆选择的轮数小于 r) {
        计算所有检测细胞与病毒的亲和力;
        选择亲和力高于 λs 一个子集;
        根据亲和力克隆这个检测细胞子集形成 S1;
        根据亲和力变异 S1 中的检测细胞;
        随机产生(|S1|)个新的检测细胞加入到 S1 中;
        对 S1 进行自体耐受处理;
        置 S 为新的检测细胞集合 S1;
  }
if(识别病毒)
  保存识别病毒的检测细胞 D;
    else
      无法通过自体耐受,没有病毒,结束病毒检测;
}
```

其中某个检测细胞 b 的克隆数目 $n_b=[\theta * f_{affirnity}(b,v)/l_b]$,$v$ 为正在检测的病毒,θ 是克隆数目的上限值,l_b 为 b 串长,$f_{affinity}(b,v)$ 为 b 和 v 之间的亲和力。对检测细胞 b 的变异就是把 b 中随机选取的 $(l_b-f_{affinity}(b,v))$ 位字符用随机选取的十六进制字符代替。

1.6 免疫记忆

生物免疫系统可以记忆以前入侵过的抗原,在下次入侵时可以进行快速的反应。同样,在本文的病毒检测系统中,一旦识别一个新的计算机病毒就把对应的检测细胞加入到记忆检测细胞集合中成为记忆检测细胞。病毒识别算法如下:

```
procedure detect_algorithm {
    初始设置门限比率等参数;
    读取自体集合;
    初始化记忆检测细胞集合 Sm;
    读可疑样本文件 F 用于识别;
    if (Sm 中所有的检测细胞与 F 的亲和力都小于 λi)
      没有病毒,检测完成。
    if (Sm 中的某个检测细胞 D 识别 F 中的病毒)
      已知病毒,检测完成;
    else{
      调用 clone_selection 算法学习识别;
      把识别病毒的新检测细胞 D 存入记忆检测细胞集合 Sm 中;
    }
}
```

其中 λ_i,$(\lambda_i \leq \lambda_s < \lambda_r)$ 判断是否存在有病毒的门限比率。

2 实验结果

本实验具体的检测算法采用了 detect_algorithm 算法,其中的自体集合就是 20 多个系统文件,记忆检测细胞集合 Sm 包括了三个分别能够检测 Jerusalb(Jerusalem 病毒家族中的一个),Tiny-143(Tiny 病毒家族中的一个),by-vienna(Vienna 病毒家族中的一个)的检测细胞,可疑样本文件 F 是 windows 目录下面的所有文件,其中包括了感染了病毒的 100 多个文件(病毒感染通过可控的途径实现)。

实验中的一些常数的设定:检测细胞最大的长度为 24 位,病毒产生识别的门限比率为 $\lambda_r=90\%$,检测细胞的筛选门限比率 $\lambda_s=50\%$,判断是否有病毒的门限比率 $\lambda_i=30\%$,检测细胞最大克隆数目 $\theta=6$,克隆选择的最大轮数 $r=10^5$。

表 1 是算法对三个家族的已知病毒、变异病毒的识别效果,其中识别率表示了对感染病毒的 100 多个文件的识别效果。误别率表示没有感染病毒但是被认为有病毒的文件和 Windows 下所有没有感染病毒的文件的比值,在表中的误别率测试只对 Window 系统文件进行测试。已知病毒的检测细胞在最初的记忆检测细胞集合中,变种病毒的检测细胞通过算法学习生成。表 2 给出了识别变异病毒的一些算法性能数据。其中变异轮数指识别病毒的变种需要经过多少轮变异,该指标说明变异算法是否足够有效率,产生检测细胞数目和成熟检测细胞数目说明克隆选择算法的有效性,即变异后有多少检测细胞是有效的检测系统。

表 1　对病毒的识别效果

病毒名称	检测细胞	识别率	误别率	备注
Jerusalb	470033c08ec026a1fc03	100%	0	已知病毒
Tiny-143	4db8004ccd21eb029090beff	100%	0	已知病毒
by-vienna	8bfe83c71f908bde83c61f90	100%	0	已知病毒
By-volat	8bfe83c73b908bde83c60290	100%	0.013%	by-vienna 变种
Sub-zero	470033c08ec08089318b	100%	0	Jerusalb 变种
Sunday	470033c08ec0bb3c0320	100%	0	Jerusalb 变种
Tiny-167	4db8004ccd21eb0290900e08	100%	0.021%	Tiny-143 变种
Tiny-198	4db8004ccd21e800000e83ed	100%	0.021%	Tiny-143 变种
Jerusalb	470033c08ec026a1fc03	100%	0	已知病毒

(下转第 108 页)

$R\text{-}X_i$ 分别为产生 X_i 的子模块，由 Product 模块分别计算各服务满意度得分 X_i 和 Y_i 的乘积，再由 sum 模块把 6 个计算结果相加得到的即为服务系统总体满意度 TSI；TSI 的数值再通过 Out 端口输出到 MATLAB 的 workspace。整个模型采用 MC 方法抽取随机数实现 TSI 模型的仿真算法。

仿真模型图 1 中产生 Y_i 的模块 Subsystem，功能是实现归一化权重 Y_i 的仿真计算，由 6 个抽取随机数模块、2 个数学计算模块以及 Mux 模块组成，如图 2 所示，名为 R_i 的模块分别模拟调查问卷中数据对于交通、游览、餐饮、住宿、娱乐、购物等 6 个指标重要度的打分情况进行随机数的抽取，对应的产生数据对各指标重要度的打分 A_i 等数值输出，经由 *Mux* 模块使这 6 个输出值合并成为一个向量输出，通过 sum 把向量中的元素相加求得 $\sum A_i$，然后同时把 sum 模块的计算结果和 A_i 向量输入到 Product 模块中，实现归一化权重 Y_i 的式计算。

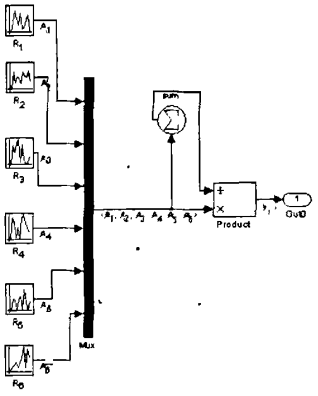

图 2 子系统 Subsystem 的模型图

图 1 中 $R\text{-}X_i$ 模块为产生正态分布随机数的模块，其模块参数 Mean 设置为 8.2，Variance 设置为 1.5，数据来自于调查问卷的数据统计 X_i 的样本均值和方差（见表 1）。

5 TSI 仿真计算结果与结论

所有参数设置完毕，即可运行仿真模型。仿真计算 1500 次，输出结果为 TSI 的 1500 次计算结果，以 yout 为数组名存放在工作空间中，并求得 TSI 的方差。上海旅游服务系统的顾客满意仿真结果的均值为 7.7579，方差为 1.9763，并绘出 TSI 的直方图，如图 3 所示。

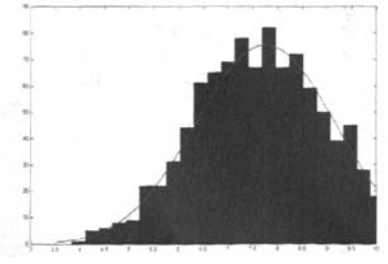

图 3 上海旅游服务系统的游客满意度仿真结果直方图

上海旅游服务权重得分 Y_i 以及游客满意度得分 X_i 都近似服从于正态分布，服务系统总体满意度也可认为近似于服从正态分布。上海旅游服务系统的游客满意度仿真结果均值为 7.7579，方差为 1.9763，所以目前上海旅游服务系统的游客满意度评价接近“较满意”的程度。评价结果不排除取值或计算过程中可能存在的误差导致评价结果的误差。本文通过对上海旅游服务系统的游客满意度进行仿真，对其现状进行评价，所建仿真模型在理论上具有一定科学性和合理性，在应用上具有广泛性和可操作性，其结果对于发展旅游业具有一定参考作用。

参 考 文 献

[1] 国家旅游局教育司. 旅游学概论[M]. 北京：中国旅游出版社，2001.
[2] 刘德艳. 上海旅游资源产品化初探[J]. 旅游学刊，2000(4)：38－42.
[3] 吴晓明，等. 顾客满意度的模糊综合评价方法[J]. 计量与测试技术，2003(3)：39－41.
[4] 许斌. 顾客满意度测评[J]. 农村金融研究，2004(7)：4－10.
[5] 马全恩，等. 顾客满意度评价指标体系设计[J]. 陕西工学院学报，2003，19(4)：67－69.
[6] 何大义，等. 构建中国顾客满意度指数的设想[J]. 世界标准化与质量管理，2000(10)：7－10.
[7] 范大祥，等. 顾客满意度(CSI)及其仿真[J]. 计算机仿真，2003，20(9)：130－134.
[8] 王庆庆. 房地产风险分析中的蒙特卡洛模拟[J]. 统计与决策，2005(11)：143－144.
[9] 左孝凌，等. 离散数学[M]. 上海：上海科学技术文献出版社，2002.
[10] 汪企新，等. 线性代数与概率统计[M]. 上海：上海世界图书出版公司，2001.
[11] Cui Xiangqun. Computer Simulation for the Active Support System of LAMOST Reflecting Schmidt Plate[J]. Nanjing Astronomical Instrument Research Center，2000：208－211.

(上接第 53 页)

表 2 对变种病毒的识别的算法效率

病毒名称	病毒家族	变异轮数	产生检测细胞数目	成熟检测细胞数目
By-volat	vienna	7	7	7
Sub-zero	Jerusalem	294	588	294
Sunday	Jerusalem	15	45	30
Tiny-167	Tiny	1	2	2
Tiny-198	Tiny	4	4	4

3 结 论

从实验结果可以看到，系统对已知病毒和它的一些变种病毒具有很好的识别能力，而且能够产生误别率很低的特征码。在未知病毒识别过程中系统体现了多样性，自学习，动态调节等多种特性。

参 考 文 献

[1] Kephart J O. A biologically inspired immune system for computers. Artificial Life IV，1994，130－139.
[2] Forrest S，Hofmeyr S，Somayaji A. Computer immunology. Communications of the ACM，1997，40(10)：88－96.
[3] Hofmeyr S A，Forrest S. Architecture for an artificial immune system. Evolutionary Computation，2000，7(1)：45－68.
[4] Perelson A S，Weisbuch G. Immunology for physicists. Modern Physics，1997，69(4)：1219－1267.

免疫原理在计算机病毒检测中的应用

作者： 韦芸， 章剑林， 徐慧剑， Wei Yun， Zhang Jianlin， Xu Huijian

作者单位： 韦芸,Wei Yun(浙江大学计算机科学与技术学院,浙江,杭州,310012;浙江经贸职业技术学院,浙江,杭州,310018)， 章剑林,徐慧剑,Zhang Jianlin,Xu Huijian(浙江经贸职业技术学院,浙江,杭州,310018)

刊名： 计算机应用与软件 ISTIC

英文刊名： COMPUTER APPLICATIONS AND SOFTWARE

年，卷(期)： 2008,25(9)

参考文献(4条)

1. Kephart J O A biologically inspired immune system for computers 1994

2. Forrest S;Hofmeyr S;Somayaji A Computer immunology[外文期刊] 1997(10)

3. Hofmeyr S A;Forrest S Architecture for an artificial immune system[外文期刊] 2000(01)

4. Perelson A S;Weisbuch G Immunology for physicists 1997(04)

本文读者也读过(10条)

1. 李慧颖.李志一.LI Hui-ying.LI Zhi-yi 计算机病毒检测模型的初步构建[期刊论文]-电脑知识与技术2009,5(23)

2. 李柳柏 计算机病毒检测程序设计[期刊论文]-微型机与应用2002,21(5)

3. 许春 人工免疫系统及其在计算机病毒检测中的应用[学位论文]2004

4. 赵红霞.张清华.牛之贤.ZHAO Hong-xia.ZHANG Qing-hua.NIU Zhi-xian 基于免疫原理的计算机病毒检测技术分析与展望[期刊论文]-茂名学院学报2009,19(6)

5. 高子渝.徐乃平 一种基于生物免疫系统的计算机病毒检测模型[期刊论文]-计算机应用研究2003,20(5)

6. 陈桓.刘晓洁.宋程.梁可心.CHEN Huan.LIU Xiao-jie.Song Cheng.LIANG Ke-xin 一种基于免疫的计算机病毒检测方法[期刊论文]-计算机应用研究2005,22(9)

7. 樊同科 一种基于人工免疫原理的计算机病毒检测方法[期刊论文]-商情2010(4)

8. 党齐民.宋丽丽.DANG Qi-min.SONG Li-li 基于免疫原理的计算机病毒检测研究应用[期刊论文]-电脑知识与技术2008,4(28)

9. 刘义.LIU Yi 计算机病毒木马程序的基本防御与解决[期刊论文]-电脑知识与技术2008,2(14)

10. 付麦霞.张天乐.Fu Mai-xia.Zhang Tian-le 信息融合技术在计算机病毒检测中的应用[期刊论文]-计算机安全2007(8)

引证文献(4条)

1. 邢小东.侯飞.李千路 一种应用免疫原理的计算机病毒检测方法研究[期刊论文]-计算机安全 2011(2)

2. 刘臣 医院局域网内发现的主要病毒特征与清除方法[期刊论文]-科技创新导报 2009(9)

3. 张文杰.李武鹏.李霞 基于二维混沌映射的免疫算法研究[期刊论文]-电脑开发与应用 2011(7)

4. 李超.刘以泓.邢丹丹 一种基于人工免疫的计算机病毒提取方法[期刊论文]-青岛大学学报（自然科学版）2011(4)