基于免疫的Windows未知病毒检测方法

张 瑜[1,2]，李 涛[1]，吴丽华[1]，夏 峰[1]

(1. 海南师范大学信息学院 海口 571158; 2. 四川大学计算机学院 成都 610065)

【摘要】Windows未知病毒令传统反病毒技术疲于应付、防不胜防，且查杀效果不佳。该文借鉴人工免疫思想，在深入剖析Windows PE病毒逻辑结构基础上，提出利用病毒重定位模块作为病毒基因来生成抗体以检测病毒的方法，且建立了自体与非自体、抗原提呈以及抗体生成的动态演化数学模型。实验表明，该方法对于未知Windows PE病毒的检测率较高，且具有自适应、自学习能力。

关 键 词 人工免疫； 重定位； 未知病毒； 病毒基因

中图分类号 TP393 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.01.019

Immun-Based Approach for Detection of Unknown Windows Virus

ZHANG Yu[1,2], LI Tao[1], WU Li-hua[1], and XIA Feng[1]

(1. College of Information Science and Technology, Hainan Normal University Haikou 571158;
2. College of Computer Science, Sichuan University Chengdu 610065)

Abstract To effectively detect unknown Windows PE viruses, a novel approach that roots in artificial immune system and uses the self-relocation module to generate antibodies is presented. The logical structure of Windows PE virus is briefly described. The dynamic evolution of self and nonself, the presentation of antigen, and the generation of antibody are proposed. The experiment results indicate that this approach not only has relatively high detection rate of unknown Windows PE virus, but also has better capability of self-adaptive and self-learning.

Key words artificial immune; self-relocation; unknown virus; virus gene

传统的基于病毒特征码的反病毒技术，对于已知病毒的检测和杀灭非常有效，但对于已知病毒变种或未知病毒，其病毒库缺乏相应特征码，导致查杀效果不佳。

由于计算机病毒的系统相关性，当系统从DOS转为Windows后，病毒也相应地由DOS病毒转为Windows病毒。在Windows病毒中，宏病毒因微软Office套件的安全性增强而逐渐减少；脚本病毒也因网络浏览器的安全性增强而失去发作的条件。但Windows PE(portable executable)病毒利用Windows PE可移植执行文件格式设计，且具有不同硬件平台可移植性的特点，对Windows系统形成安全威胁。为此，本文将以Windows PE病毒为目标提出新的检测方法。

对于未知Windows PE病毒的检测，反病毒研究者提出了各种不同的方法。文献[1]提出了基于API调用序列的检测方法，文献[2]提出了基于N-图形的方法，文献[3]提出了基于神经网络的方法，文献[4]提出了基于贝叶斯理论的方法，文献[5]提出了基于支持向量机的方法，文献[6]提出了基于程序行为的方法，文献[7]提出了基于数据挖掘技术的方法。上述方法从不同的技术角度对PE病毒的检测进行探讨，尽管有的方法具有一定的智能性，但因实现复杂、训练代价高、系统资源占用率高，对未知病毒的检测效率不尽如人意，难以在实际中加以推广应用。

本文借鉴人工免疫思想，在对Windows PE病毒逻辑结构进行透彻分析的基础上，利用Windows PE病毒重定位模块的独特性，提出了一种基于免疫的Windows未知病毒检测方法。实验表明，该方法对于已知和未知的Windows PE病毒都有很高的检测效率。

1 Windows PE病毒逻辑结构分析

PE文件格式是Windows的执行体文件格式，PE病毒利用格式在不同Windows硬件平台的可移植性

收稿日期：2008－03－10；修回日期：2008－10－25
基金项目：国家自然科学基金(60573130、66873246)；国家863计划(2006AA01Z435)；教育部博士点基金(20070610032)
作者简介：张 瑜(1975－)，男，博士，副教授，主要从事网络安全和计算智能等方面的研究.

而不断传播。Windows PE病毒通常具有6个逻辑模块。

1.1　重定位模块

正常程序不关心变量和常量的位置，因为它在内存中的位置在编译源程序时就已被计算好。程序装入内存时，系统不用为它重定位，只需在使用变量或常量时直接用其名字访问。同样，病毒也要用到变量和常量，当病毒感染宿主程序后，由于其依附在宿主程序的位置各有不同，病毒随着宿主载入内存后，病毒中的各个变量及常量在内存中的位置自然也不相同。既然这些变量或常量没有固定的地址，病毒在运行过程中只有靠重定位才能正常地访问与自己相关的资源。因此，Windows PE病毒都需要重定位模块才能在Windows平台上正确执行。通常，病毒的重定位模块位于病毒程序开始处，且代码少、变化不大。鉴于此，本论文提出将病毒重定位模块作为病毒基因提取以生成抗体检测病毒的方法。

1.2　获取Windows API函数地址模块

Windows程序一般运行在Ring 3级，处于保护模式中。Windows中的系统调用通过动态链接库中API函数实现。普通Windows PE程序有引入函数表，该函数表对应于代码段中所用到的API函数在动态链接库中的真实地址，调用API函数时就可以通过该引入函数表找到相应API函数的真正执行地址。

同样，Windows PE病毒也需调用API函数。但Windows PE病毒只有一个代码段，并不存在引入函数表，病毒就无法像普通Windows PE程序那样直接调用相关的API函数，而应先找出这些API函数在动态链接库中的地址。因此，Windows PE病毒必须具备获取Windows API函数地址的模块。

1.3　目标文件搜索模块

病毒要扩大影响，就必须进行传播。而要进行传播，就需要在搜索到目标文件后再执行感染功能。因此，Windows PE病毒需有目标文件搜索模块。

1.4　内存映射文件模块

内存映射文件提供了一组独立的函数，使应用程序能够通过内存指针像访问内存一样对磁盘上的文件进行访问。对文件中数据的操作便是直接对内存进行操作，大大地提高了访问速度，对计算机病毒非常重要。因此，Windows PE病毒一般具有内存映射文件模块。

1.5　添加新节以感染其他文件模块

Windows PE病毒感染其他文件的方法是在文件中添加一个新节，然后向该新节中添加病毒代码和病毒，执行后返回Host程序的代码，并修改文件头中代码开始执行位置(address of entry point)指向新添加的病毒节的代码入口，以便程序运行后先执行病毒代码。

1.6　返回宿主程序模块

为了提高自己的生存能力，病毒不应该破坏HOST程序，应在执行后立刻将控制权交给HOST程序，跳到HOST程序继续执行。

2　基于免疫的检测理论

生物免疫系统的主要功能是如何识别自体与非自体[8-11]，在此基础上对自体予以保护，对非自体进行杀灭。而计算机病毒检测系统的功能是识别正常文件和病毒文件，清除病毒保护文件。鉴于两个系统的功能相似性，计算机病毒检测系统的设计可借鉴生物免疫系统原理，以达到较好的检测效率。

2.1　检测逻辑流程

首先，从病毒文件中提取其重定位模块组成病毒基因库；其次，对病毒基因进行否定选择，以生成合格的抗体；最后，用抗体识别病毒文件。检测逻辑流程如图1所示。

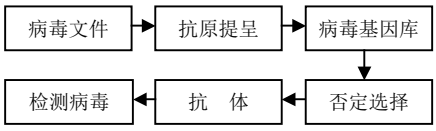

图1　病毒检测流程

2.2　自体与非自体的演化

假设问题域为AG，即$AG=\bigcup_{i=1}^{\infty}H^i$，其中$H$={0,1,2,…,9,*A*,*B*,*C*,*D*,*E*,*F*}是一个十六进制数集合，i为正整数。自体Self定义为受保护的文件，非自体Nonself定义为可疑文件。自体和非自体分别满足如下条件：Self⊂AG、Nonself⊂AG、Self∪Nonself=AG、Self∩Nonself=∅。

生物免疫系统的自体和非自体具有动态变化性。与此类似，在计算机中，自体和非自体也具有动态变化性，自体(即受保护的正常文件)由于受病毒感染而成为非自体(即可疑文件)，而非自体由于消除了病毒感染而成为自体。对于自体与非自体动态演化的考虑，可提高病毒的检测率、降低误报率。在该检测方法中，自体与非自体的动态演化递推方程定义为：

$$\begin{cases} \mathrm{Self}(t)=\begin{cases} S_{\mathrm{initial}} & t=0 \\ \mathrm{Self}(t-1)-\mathrm{Self}_{\mathrm{del}}(t)+\mathrm{Self}_{\mathrm{new}}(t) & t\geqslant 1 \end{cases} & (1) \\ \mathrm{Self}_{\mathrm{del}}(t)=\{s\mid s\in \mathrm{Self}(t-1)\wedge \exists y\in \mathrm{SA}(t-1)\wedge\langle s,y\rangle\in \mathrm{Match}\} & (2) \\ \mathrm{Self}_{\mathrm{new}}(t)=\{s\mid s\in \mathrm{Nonself}(t-1)\wedge \forall y\in \mathrm{SA}(t-1)\wedge\langle s,y\rangle\notin \mathrm{Match}\} & (3) \\ \mathrm{Nonself}(t)=\begin{cases} \mathrm{NS}_{\mathrm{in\,itial}} & t=0 \\ \mathrm{Nonself}(t-1)-\mathrm{Nonself}_{\mathrm{del}}(t)+\mathrm{Nonself}_{\mathrm{new}}(t) & t\geqslant 1 \end{cases} & (4) \\ \mathrm{Nonself}_{\mathrm{del}}(t)=\mathrm{Self}_{\mathrm{new}}(t) & (5) \\ \mathrm{Nonself}_{\mathrm{new}}(t)=\mathrm{Self}_{\mathrm{del}}(t) & (6) \end{cases}$$

式(1)刻画了自体的动态演化过程，其中S_{initial}是由500个正常的Windows系统文件所组成的自体初始集合；$\mathrm{Self}_{\mathrm{del}}$是从自体集合中删除的已成为非自体的文件所组成的集合；$\mathrm{Self}_{\mathrm{new}}$是非自体集合中未匹配任何抗体的非自体所组成的集合。

式(4)刻画了非自体的动态演化过程，其中$\mathrm{NS}_{\mathrm{initial}}$是由已感染了病毒的100个可疑文件所组成的非自体初始集合；$\mathrm{Nonself}_{\mathrm{del}}$是从非自体集合中删除的未匹配任何抗体的非自体所组成的集合；$\mathrm{Nonself}_{\mathrm{new}}$是从自体集合中删除的已成为非自体的文件所组成的集合。

2.3 抗原提呈

抗原提呈是从疑似病毒文件中提取病毒基因(病毒特征码)以组成病毒基因库。本文将Windows PE病毒重定位模块作为病毒基因加以提取。原因如下：(1) Windows PE病毒重定位模块代码少，变化不大，且通常位于病毒程序开始处，易于提取；(2) Windows PE病毒获取Windows API函数地址模块、目标文件搜索模块、内存映射文件模块、返回宿主程序模块，在正常程序中也常会使用，故不宜作为特征基因提取；(3) 尽管Windows PE病毒添加新节以感染其他文件模块在普通的程序中不太常用，但其实现复杂代码量大，也不宜作为特征基因提取。可提取的病毒基因为：

E8 00 00 00 00 5B 81 EB 2E 1F 40 00

Windows PE病毒重定位模块如图2所示。

```
.text:00401F29   E8 00 00 00 00          call $+5
.text:00401F2E       loc_401F2E:
.text:00401F2E   5B                      pop ebx
.text:00401F2F   81 EB 2E 1F 40 00 sub ebx, offset loc_401F2E
```

图2 Windows PE病毒重定位模块

经抗原提呈所得到的病毒基因库定义为：

$$V=\{v\mid v\in\bigcup_{i=8}^{32}H^i\wedge|v|=i\wedge v=\mathrm{Ap}(x\in \mathrm{Nonself})\} \quad (7)$$

式中 Ap(x)是抗原提呈函数，它将病毒重定位模块提取出作为病毒基因v，基因长度为8～32个十六进制编码。

2.4 抗体生成及病毒检测

生物获得性免疫系统一般通过接种疫苗而获得特异性抗体，进而使机体具有免疫功能。疫苗是对病毒基因库中的病毒基因进行基因突变和基因重组而得到的新病毒基因，抗体通过从病毒基因库中提取疫苗而生成。假设有抗体：

E8 00 00 00 00 5B

对其进行基因突变后可能得到的抗体形如：

E8 02 00 00 00 5F

另假设有抗体：

E8 00 00 00 00 5B

E8 10 00 00 00 5A

对它们进行基因重组后可能得到的抗体形如：

E8 10 00 00 00 5B

病毒基因操作如图3所示。

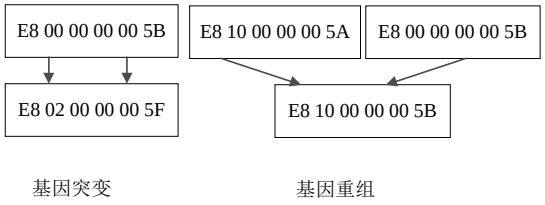

图3 病毒基因操作

利用抗体以及抗体与抗原的亲和力，可生成检测器检测病毒。检测器集合定义为：

$$D=\{\langle d,\mathrm{affinity}\rangle\mid d\in\bigcup_{i=8}^{32}H^i,\mathrm{affinity}\in N\} \quad (8)$$

式中 d为抗体；affinity为抗体与抗原的亲和力。

在利用抗体检测病毒抗原时，有的抗体在生成后由于缺乏亲和力而未能匹配病毒抗原，将导致死亡而被从检测器中删除。有的抗体由于亲和力大于阈值则会成为永久抗体。这种抗体的动态生成机制保证了该检测技术的动态自学习性。抗体的演化递推方程定义为：

$$\begin{cases} \mathrm{SA}(t)=\begin{cases}\varnothing & t=0\\ \mathrm{SA}(t-1)+\mathrm{SA}_{\mathrm{new}}(t) & t\geqslant 1\end{cases} & (9)\\ \mathrm{SA}_{\mathrm{new}}(t)=\{v\mid v\in D,\forall y\in \mathrm{Self},\langle v,y\rangle\notin \mathrm{Match}\wedge v.\mathrm{affinity}\geqslant\beta\} & (10)\\ \mathrm{Match}=\{\langle v,y\rangle\mid v\in D,y\in \mathrm{AG},f_{\mathrm{match}}(v.d,\mathrm{AP}(y))=1\} & (11)\\ f_{\mathrm{match}}(v,y)=\begin{cases}1 & f_{\mathrm{affinity}}(v,y)/L_v\geqslant\alpha\\ 0 & \mathrm{otherwise}\end{cases} & (12)\\ f_{\mathrm{affinity}}(v,y)=\max(x_1,x_2,\ldots,x_{|L_v-Ly|+1}) & (13)\\ x_i=\sum_{i=1}^{\min(L_v,L_y)}\theta_{ij} & (14)\\ \theta_{ij}=\begin{cases}1 & v_i=y_{i+j-1},1\leqslant i\leqslant|L_v-L_y|+1,1\leqslant j\leqslant L_v\\ 0 & \mathrm{otherwise}\end{cases} & (15)\end{cases}$$

式(9)刻画了特异性抗体的动态演化过程，其中SA表示特异性抗体集合；$\mathrm{SA}_{\mathrm{new}}$表示新生成的抗体集，且与自体的亲和力大于阈值$\beta$；$f_{\mathrm{match}}$表示抗体与抗原的匹配度计算函数；$f_{\mathrm{affinity}}$表示抗体与抗原的亲和力计算函数；Match表示抗原、与抗原相匹配的抗体二元组所组成的集合；θ_{ij}表示抗体与抗原亲和力的计算值，如匹配则为1，否则为0；L_v表示疫苗长度；L_y表示抗原长度。

3　实　验

实验所用数据由500个正常的Windows系统文件和100个病毒文件组成。Wildlist是The WildList Organization International的一个采集自权威反病毒组织和专家的病毒列表，该列表中包含的病毒是那些当时有实际感染和传播行为被发现的病毒，是病毒感染情况的晴雨表。Wildlist通常每月发布一次，已成为国际上反病毒软件测评的权威病毒样本集。参考WildList，从http://vx.netlux.org下载100个Windows PE病毒样本。实验目的是测试本文方法对于(已知和未知)病毒的检测率以及对于正常文件的误报率。实验结果如图4所示。对于PE病毒的检测率为97%，漏报率为3%，而误报率只有3.6%。由于本文方法主要针对PE病毒不可或缺的重定位模块进行检测，所以检测率极高。当然，由于有些病毒对其重定位模块进行了变形、加花指令等操作，导致本文方法会有极低的漏报率。另外，正常的系统文件几乎不需要进行自我重定位，但由于IA-32体系数据和代码的不可区分性，如数据中恰好有与重定位指令机器码相同的数据，则会导致误报。这就是本文方法有极低误报率的根本原因。

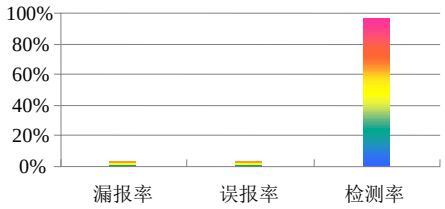

图4　本文方法实验结果

为了客观验证本文方法的检测效率，与当前成熟的反病毒产品(瑞星2008、金山毒霸2008、江民KV2008、卡巴斯基7.0以及Eset NOD32)进行了对比实验，对比实验结果如图5所示。

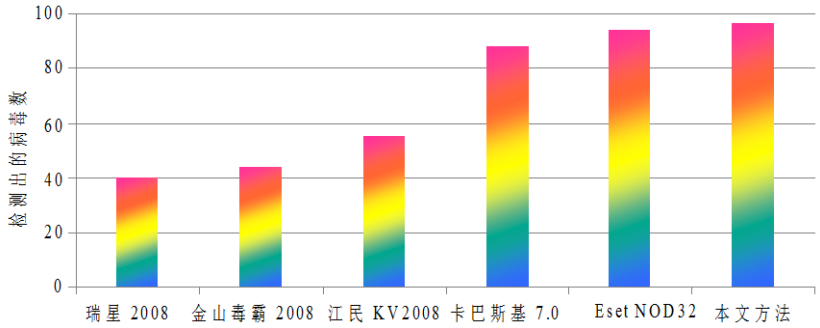

图5　不同技术的实验结果对比

从实验结果来看，本文方法的检测率为97%，高于瑞星2008的40%、金山毒霸2008的44%、江民KV2008的55%、卡巴斯基7.0的88%、Eset NOD32的94%。因此，从本文方法的单独实验和与其他技

术的对比实验结果来看，本文方法的检测效果都非常理想。因为传统反病毒技术是基于病毒特征码，对于已知病毒的查杀非常有效，而对于未知病毒或已知病毒的变种，由于其病毒库中无此特征码，故不能检测到病毒。从本质上说，本文方法是启发式查毒，是基于病毒所共有的重定位模块提取病毒基因，运用免疫思想改变基因，从而能检测出未知病毒或已知病毒的变种。

4 结 论

未知Windows PE病毒不断涌现令传统基于特征码的反病毒技术防不胜防、疲于应付，且查杀效果不佳。本文提出的基于免疫的未知Windows PE病毒检测方法，在对Windows PE病毒逻辑结构深入分析的基础上，结合人工免疫思想，通过提取病毒重定位模块作为病毒基因，再利用生成的抗体对未知Windows PE病毒进行检测。实验表明，该方法不仅具有检测率高、误报率和漏报率低的特性，且查毒效率均高于当前成熟的反病毒产品。

当然，本文方法目前只能对未加壳的Windows PE病毒进行检测实验。病毒为更好地生存，为防止被反病毒软件反跟踪查杀以及被动态调试和被静态分析，会增加自身的变形能力，还会与程序加壳压缩技术结合。病毒技术和程序加壳技术的紧密结合，将对反病毒技术提出更高的挑战。对于加壳的Windows PE病毒的检测将是下一步要进行的工作。

本文研究工作得到了海南师范大学引进博士科研启动项目(00203020214)的支持，在此表示感谢。

参 考 文 献

[1] XU J Y, SUNG A, CHAVEZ H P. Polymorphic malicious executable scanner by API sequence analysis[C]//Fourth International Conference on Hybrid Intelligent Systems. Washington, USA: IEEE Computer Society Press, 2004: 378-383.

[2] REDDY D K S, PUJARI A K. N-gram analysis for computer virus detection[J]. Journal in Computer Virology, 2006, 2: 231-239.

[3] YAO Yu, YU Ge, GAO Fu-xiang. A neural network approach for misuse and anomaly intrusion detection[J]. Wuhan University Journal of Natural Sciences, 2005, 10(1): 115-118.

[4] SHIH Dong-her, CHIANG Hsiu-sen, DAVID C Y. Classification methods in the detection of new malicious emails[J]. Information Sciences, 2005, 172(1): 241-261.

[5] WANG Shuo, ZHOU Ji-liu, PENG Bo. Unknown virus detection based on API sequence and support vector machine[J]. Journal of Computer Applications, 2007, 27(8): 1942-1943.

[6] MICHAEL C C. Finding the vocabulary of program behavior data for anomaly detection[C]//Proceedings of DARPA Information Survivability Conference And Exposition. Washington, USA: IEEE Computer Society Press, 2003: 152-163.

[7] SCHULTZ M G, ESKIN E, ZADOK E. Data mining methods for detection of new malicious executables [C]//IEEE symposium on security and privacy. Oakland, California, USA: IEEE Computer Society Press, 2001: 1-38.

[8] FORREST S, PERELSON A S. Self-nonself discrimination in a computer[C]//IEEE Symposium on Security and Privacy. Los Alamitos, CA: IEEE Computer Society Press, 1994: 202-213.

[9] ZHANG Yu, LI Tao, QIN Ren-chao. A dynamic immunity-based model for computer virus detection[C]// International Symosium on Information Processing. Moscow, Russia: IEEE Computer Society, 2008: 515-519.

[10] ZHANG Yu, LI Tao, SUN Jia, et al. An FSM-based approach for malicious code detection using the self-relocation gene[C]//Fourth International Conference on Intelligent Computing. Shanghai, China: Springer LNCS 5226, 364-371.

[11] 刘才铭, 赵 辉, 张 雁, 等. 受人工免疫启发的脚本病毒检测模型[J]. 电子科技大学学报, 2007, 36(6): 1219-1222.
LIU Cai-ming, ZHAO Hui, ZHANG Yan, et al. Artificial immunity-inspired script-virus detection model[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(6): 1219-1222.

编 辑 蒋 晓