文章编号: 1001-9081(2008)11-2829-03

基于 Win32 API的未知病毒检测

陈 亮[1], 郑 宁[1], 郭艳华[1], 徐 明[1], 胡永涛[2]
(1.杭州电子科技大学 计算机学院, 杭州 310018; 2.公安部第三研究所, 上海 201204)
(mecury5460@163.com)

摘 要: 提出了一个基于行为特征向量的病毒检测方法。特征向量的每一维用于表示一种恶意行为事件, 每一事件由相应的 Win32应用程序编程接口(API)调用及其参数表示, 并实现了一个自动化行为追踪系统(Argus)用于行为特征的提取。试验中, 通过对样本数据的分析, 利用互信息对特征向量进行属性约简, 减少特征维数。试验结果表明, 约简后的模型对于发生行为事件数大于1的病毒程序仍有着较好的检测效果。

关键词: 恶意行为; Win32应用程序编程接口; 互信息; 属性约简
中图分类号: TP309.5 文献标志码: A

Unknown virus detection based on Win32 API

CHEN Liang, ZHENG Ning, GUO Yan-hua, XU Ming, HU Yong-tao
(1. School of Computer Science, Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China;
2. The Third Research Institute of Ministry of Public Security, Shanghai 201204, China)

Abstract: This paper proposed a virus detection method using behavior feature vector. Each dimension of the vector stood for a malicious behavior event represented by corresponding Win32 API calls and their certain parameters. An automatic executable behavior tracing system (Argus) was also implemented to dynamically capture the events. In the experiment, attribute reduction was applied to mutual information to decrease the number of dimension after an analysis of the sample dataset. Experimental result suggests that model after attribute reduction is still efficient in detecting unknown virus which has more than one event captured.

Key words: malicious behavior; Win32 Applications Programming Interface (API); mutual information; attribute reduct

0 引言

应用程序编程接口(Applications Programming Interface, API)作为操作系统为应用程序提供的服务接口, 完成对网络、文件系统和其他重要系统资源的访问。借助逆向工程和调试技术, 反病毒工程师通过分析可疑上报程序 API调用及参数情况, 依据经验判断程序是否为病毒。如何将人工分析的方法自动化, 并最终实现病毒的自动化识别, 已成为反病毒研究领域的一个研究热点。文献[1]将 API调用序列应用于多态病毒的检测, 通过遍历文件代码段, 获取目标 API, 该方法需从文件内容中获取 API, 因此依赖反汇编、解压缩、脱壳等技术的发展。文献[2]将模式识别应用于病毒检测, 将病毒行为特征定义为相关的 API调用, 但仅使用 API描述某一恶意行为并不精确, 同一个 API, 其危险程度会随着其参数的不同而不同。文献[3]采用 API和参数结合的方法用于病毒的检测, 但仅定义文件创建和注册表操作两类行为, 这无疑使得很多正常程序被误报为病毒。

鉴于此, 本文根据专家经验总结出 Win32环境下的病毒常见恶意行为, 将恶意行为表示为相应的 Win32 API(后文简称 API)调用及其参数形式, 并结合虚拟机技术设计了一个自动化行为追踪系统用于 API和参数的动态捕获。同时通过对样本数据的分析, 利用互信息对行为特征属性进行评价和约简, 去掉专家经验中区分能力较弱的行为特征。

1 病毒行为事件定义

API作为操作系统为应用程序提供的服务接口, 完成对系统资源的访问。但仅使用 API并不能准确描述一个恶意行为, 同一个 API调用, 其危险性随着其参数的不同而不同。拷贝一个可执行文件到系统敏感目录相对拷贝一个文本文件到用户目录的潜在危险性要高的多。因此本文采用 API与参数相结合的方式来描述恶意行为, 并依据专家经验定义了一个35维的特征向量, 其每一维代表一种恶意行为事件, 该事件由相应的 API及其参数表示, 仅当程序调用的 API和参数均满足条件时, 我们才认为发生了该行为事件。表1中给出部分行为事件的定义。

2 行为特征提取

设计了一个用于捕获本文所定义恶意行为的自动化行为追踪系统(automatic executable behavior tracing system, Argus), 其模块结构如图1所示。

1)样本入库模块。该模块用于将样本程序批量导入样本数据库。

2)虚拟机控制模块。该模块负责从样本数据库中读取样本文件, 然后将读出的样本文件放入 VMware虚拟机中运行, 并通过 API追踪模块捕获样本运行时产生的实时行为, 待行为捕获工作完成后, 重启虚拟机并处理下一样本文件。出

收稿日期: 2008-05-26; 修回日期: 2008-08-01。 基金项目: 浙江省自然科学基金资助项目(Y106176); 浙江省科技厅计划项目(2007C33058); 上海市科委基金资助项目(06511502)。

作者简介: 陈亮(1984-), 男, 湖北鄂州人, 硕士研究生, 主要研究方向: 反计算机病毒、计算机取证; 郑宁(1961-), 男, 浙江杭州人, 研究员, 博士生导师, 主要研究方向: 网络与信息安全; 郭艳华(1963-), 女, 浙江杭州人, 副教授, 主要研究方向: 信息管理与计算语言学; 徐明(1970-), 男, 浙江杭州人, 副教授, 主要研究方向: 网络信息技术; 胡永涛(1976-), 男, 上海人, 副研究员, 主要研究方向: 反计算机病毒。

于虚拟环境安全和追踪结果准确性的考虑，每次样本处理之前，虚拟机将被还原到初始状态。

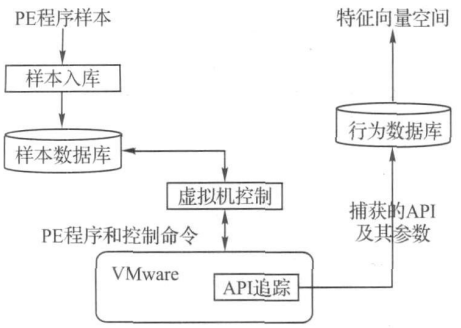

图 1 Argus模块结构

3）API追踪模块。该模块是 Argus系统的核心，用以分析样本进程运行时所加载的系统 DLL，然后根据事先定义的恶意行为来对特定 API进行监控。当样本进程调用到了特定 API时，API追踪模块会在该处产生调试中断，通过调试子系统获取中断发生时样本进程的执行环境（指令寄存器、堆栈寄存器等多个寄存器的值），并将中断时的堆栈参数交与特定 API相关的行为处理函数进行具体分析。

3 基于差分互信息的属性约简

文中定义的恶意行为事件，均为依据专家经验得来，这些行为事件在病毒程序大量出现的同时，在正常程序中也可能经常发生。对基于特征向量的模型来说，这类行为事件对类别的区分度不明显，不仅无益于模型的检测效果反而会增加检测的运算开销，因此需对这些行为事件进行评价，删除其中区分能力较弱的行为事件。互信息是信息论中的一个重要概念，用来度量一个消息中两个信号之间的相互依赖程度，常用于文本分类和垃圾邮件过滤中特征属性的约简和选择[4-6]。本文将互信息用于衡量行为事件属性和类别属性之间的统计关联程度，作为行为事件评价和约简的依据，其定义为：

$$I(a_k, d) = P(a_k \mid d) \lg \frac{P(a_k, d)}{P(a_k) P(d)} \quad (1)$$

其中 a_k表示某一行为事件属性，$P(a_k)$为该行为事件发生的概率，在实际应用中，表示为：训练集中发生该事件的样本所占比例 $P(a_k) = |a_k| / N$，N为样本总数；$P(d)$为样本程序属于类别 d（病毒或正常程序）的概率，$P(d) = |d| / N$，$P(a_k \mid d)$表示样本为类别 d时，行为事件 a_k发生的概率，$P(a_k \mid d) = |a_k \cap d| / |d|$，$|d|$表示类别为 d的样本数；$P(a_k, d)$为样本发生事件 a_k且属于类别 d的概率，$P(a_k, d) = |a_k \cap d| / N$；互信息 $I(a_k, d)$的大小表示了行为事件 a_k和类别 d的关联程度，值越大，关联程度越强。

在属性约简中，某属性对两个类别互信息都大并不意味着它对这两个类别的区分能力也强。而利用差分互信息则可以更好地衡量属性对类别的区分能力，能够过滤掉在两类间概率分布比较均匀、分布差异不大的属性，保留大量出现在某一类中，而在另一类中出现次数比较少的属性。

差分互信息定义为[4]：

$$MID(a_k) = |I(a_k, d_1) - I(a_k, d_0)| \quad (2)$$

其中 $I(a_k, d_1)$为行为事件属性 a_k与类别 d_1（病毒）的互信息，$I(a_k, d_0)$为属性 a_k与类别 d_0（正常程序）的互信息。$MID(a_k)$越大即该行为事件对类别的区分能力越强。

为避免一些发生概率较小，但在某一类中发生的概率远远大于另一类的行为事件属性因 $MID(a_k)$太小被约简，文中引入 $PD(a_k)$对差分互信息进行改进，其定义：

$$PD(a_k) = \max\left(\frac{P(a_k \mid d_1)}{P(a_k \mid d_0) + \alpha}, \frac{P(a_k \mid d_0)}{P(a_k \mid d_1) + \beta}\right) \quad (3)$$

α、β分别取 $1/|d_0|$，$1/|d_1|$以避免计算机分母为零。仅当 $MID(a_k)$和 $PD(a_k)$均小于各自的阈值 TMID和 TPD时该行为事件属性才被约简。

表 1 常见病毒行为定义

行为类别	具体行为事件	对应的 API调用	参数描述
文件相关类行为	MODIFY_FILEATTR修改文件属性	SetFileAttributes	参数 dwFileAttributes为隐藏，系统，或只读
	COPY_FILE复制 PE文件到关键目录	CopyFile CopyFileEx	源文件为 PE文件，目标路径为系统路径或逻辑分区根目录
	WRITE_FILE 向敏感目录下的 PE文件写入数据	WriteFile WriteFileEx	写入文件句柄对应为 PE文件，目标路径为系统路径或逻辑分区根目录
	DELETE_FILE在关键目录删除文件	DeleteFile	删除的文件所在路径为系统路径或逻辑分区根目录
进程相关类行为	ISDEBUG检测是否处于被调试状态	IsDebuggerPresent CheckRemoteDebuggerPresent	任意参数
	SHELL_EXECUTE是否以 SHELL方式隐蔽启动外部程序	ShellExecute ShellExecuteEx	执行方式是否为 SW_HIDE方式（隐藏窗口方式）
	CREATE_PROCESS创建新进程	CreateProcess	任意参数
	…	…	…

4 病毒检测

4.1 数据集描述

试验数据集由 1865个黑样本（恶意的可执行程序）和 2821个白样本（正常的可执行程序）构成。黑样本由哈尔滨安天实验室提供，为避免数据集中存在过多来自同一病毒的变种，使得训练后的分类器带有倾向性，已依据病毒命名规则去掉黑样本中病毒程序的近似变种。白样本均从 Internet网上获取，为增强检测模型在实际应用中的实用性，白样本中包含一定数量介于白样本与黑样本之间的灰样本，这类样本和黑样本有着较为类似的恶意行为事件，但实际上仍然属于合法的正常程序。黑白样本均为 PE结构的可执行程序。白样本均通过卡巴斯基扫描杀毒，其中可能存在的重复样本也已通过 MD5值去除。

4.2 分类算法描述

朴素贝叶斯分类算法的基础是贝叶斯定理，其在分类性能上与决策树和神经网络均有可比性，在处理大规模数据库时，贝叶斯分类算法表现出较高的分类准确性和运算性能。

计算机病毒检测是一个二分类问题。对于样本空间的每一个样本可以看作是 n维空间的一个点 $X=(x_1, x_2, \cdots, x_n)$,其中 $x_1, x_2, \cdots, x_n$ 分别为行为事件属性 $a_1, a_2, \cdots, a_n$ 的取值。D为样本类别属性,$D=\{d_0, d_1\}$,d_1 表示病毒,d_0 表示正常程序。依据朴素贝叶斯定理,病毒检测判别式表示为:

$$C(X)=\arg\max_d\left(\frac{P(d)P(x|d)}{P(x)}\right) \quad (4)$$

根据朴素贝叶斯理论各条件属性相互独立的基本假设和 $P(x)$ 在判别式中对具体样本而言为常数,判别式可简化为:

$$C(X)=\arg\max_d\left(P(d)\prod_{i=1}^{n}P(x_i|d)\right) \quad (5)$$

4.3　试验及结果

为保证检测方法对典型病毒的检测效率,训练和测试仅对发生恶意行为事件数大于 1的黑白样本进行。如表 2所示,黑样本集合 O按比例 4∶1随机划分为集合 A和集合 B,白样本集合 O′随机划分成集合 A′和 B′,其中 A′的样本数等于集合 A,集合 A和 A′用于训练朴素贝叶斯分类器,集合 B和 B′作为测试集。模型检测效果由检出率(True Positive Fraction, TPF)和误报率(False Positive Fraction, FPF)评价。检出率为黑样本中正确检测为黑样本的样本比例,误报率为白样本中误报为黑样本的比例。检出率和误报率结果取 5次随机试验的平均结果。

表 2　试验中的样本

样本类别	总样本数	事件数大于 1的样本数	训练集
黑样本	1865	1200(O)	960(A)
白样本	2821	2444(O′)	960(A′)

阈值 TMID和 TPD取值原则上应保持属性约简前后模型检测效果相当,在试验中,分别取 0.12和 1.5。表 3列出属性约简前后的试验结果。

表 3　属性约简前后的试验结果　%

测试集	维数为 35		维数为 20	
	TPF	FPF	TPF	FPF
B∪B′	83.82	7.87	86.14	7.96

从表中可以看出,经过约简后的维数从 35维降到 20维,在减少运算量的同时,模型检测效果基本保持不变,且对发生事件数大于 1的样本有着较高的检出率和较低的误报率。

表 4列出的是属性约简后保留的 20个属性中 ISDEBUG、WRITE_FILE、DELETEFILE和 CREATE_PROCESS四个属性,其中 MI1表示行为事件属性与类别病毒的互信息,MI0表示行为事件属性与正常程序的互信息。

这四个属性的共同特点是其与正常程序的互信息要高于病毒程序。在实际的启发式病毒检测中,应降低这类行为在病毒判定中的决策权重,以减少误报率。

表 4　部分保留属性

行为属性	MI1	MI0	MID	PD
ISDEBUG	0.54	0.93	0.39	1.58
WRITE_FILE	0.12	0.30	0.19	1.85
DELETEFILE	−0.06	0.95	1.01	4.66
CREATE_PROCESS	0.27	0.51	0.24	1.64

5　结语

本文提出了一个基于行为特征向量的未知病毒检测方法。特征向量的每一维用于表示一种恶意行为事件,每一事件由相应的 Win32 API调用及其参数表示。该方法通过监视程序运行行为获取恶意行为特征,其特征提取不依赖于文件内容,可克服基于特征码扫描法在病毒检测上的不足。对于已定义的行为属性,利用互信息理论进行评价和约简,对专家经验进行反馈,去掉专家经验中区分能力较弱的行为属性。实验结果表明,属性约简后的模型仍保持原有模型的检测效果,对于发生事件数大于 1的样本,有着较高的检出率和较低的误报率。另外,从信息论角度对行为特征的评价也为启发式病毒检测中病毒行为决策权重大小的选取提供了一定的参考。今后的工作将关注以下两个方面:增加特征向量维数,定义更多在病毒和正常程序中具有区分能力的行为特征,使得程序在理论上能被捕获到更多的事件,增加模型可检测样本范围;另一方面,对于被删除的属性,探究其删除的原因,是因为参数定义的不够准确还是该 API函数本身在黑白样本中就不具有区分度。

参考文献:

[1] XU J Y, SUNG A H, CHAVEZ P, et al. Polymorphic malicious executable scanner by API sequence analysis[C]// Proceedings of the 4th International Conference on Hybrid Intelligent Systems. Washington, DC: IEEE Computer Society, 2004: 378−383.

[2] KOIKE R, NAKAYAN, KOI Y. Development of system for the automatic generation of unknown virus extermination software[C]// Proceedings of the 2007 International Symposium on Applications and the Internet. Washington, DC: IEEE Computer Society, 2007: 8−8.

[3] 张波云,殷建平,蒿敬波,等.基于多重朴素贝叶斯算法的未知病毒检测[J].计算机工程,2006,32(10):18−21.

[4] 张文良,黄亚楼,倪维建.基于差分贡献的垃圾邮件过滤特征选择方法[J].计算机工程,2007,33(8):80−82.

[5] 伍建军,康耀红.基于改进的互信息特征选择的文本分类[J].计算机应用,2006,26(S2):172−173.

[6] 王卫玲,刘培玉,初建崇.一种改进的基于条件互信息的特征选择算法[J].计算机应用,2007,27(2):433−435.

(上接第 2828页)

[6] HUANG Z, CHEN K, WANG Y. Efficient identity-based signatures and blind signatures[C]// Proceeding of the 4th International Conference on Cryptology and Network Security. Berlin: Springer-Verlag, 2005: 120−133.

[7] GALINDO D, HERRANZ J, KILTZ E. On the generic construction of identity-based signatures with additional properties[C]// Advances in Cryptology—Asiacrypt2006. LNCS 4284. Berlin: Springer-Verlag, 2006: 178−193.

[8] BONEH D, LYNN B, SHACHAM H. Short signatures from the Weil pairing[C]// Advances in Cryptology—Asiacrypt2001. LNCS 2248. Berlin: Springer-Verlag, 2001: 514−532.

[9] BOLDYREVA A. Efficient threshold signatures, multisignatures and blind signatures based on the gap-Diffie-Hellman-group signature scheme[C]// Proceeding of the 6th International Workshop on Theory and Practice in Public Key Cryptography. Berlin: Springer-Verlag, 2003: 31−46.

[10] BONEH D, GENTRY C, LYNN B, et al. Aggregate and verifiably encrypted signatures from bilinear maps[C]// Advances in Cryptology—Eurocrypt2003. LNCS 2656. Berlin: Springer-Verlag, 2003: 416−432.