基于行为特征库的木马检测模型设计

李焕洲[1], 陈婧婧[1,2], 钟明全[1], 唐彰国[1]
(1. 四川师范大学 网络与通信技术研究所,四川 成都 610101; 2. 四川师范大学 计算机科学学院,四川 成都 610101)

摘要:目前木马检测的主流技术主要是特征码检测技术,而该技术提取特征码滞后,无法检测未知新型木马. 为了更好的检测新型木马,详细归纳总结木马的行为特征,同时在此基础上提取木马通适性行为特征,构建木马行为特征库,设计了基于行为特征库的木马检测模型,并应用模糊模式识别方法判断木马程序. 通过实验证明此模型可以对可疑程序的行为特征进行分析判断,较准确地识别木马程序. 该检测模型是对基于特征码检测技术的强有力补充,在新型木马不断涌现的今天,基于木马行为特征检测技术具有重要的应用意义.

关键词:木马; 行为特征; 模糊模式识别
中图分类号:TP393.08 文献标志码:A 文章编号:1001-8395(2011)01-0123-05
doi:10.3969/j.issn.1001-8395.2011.01.024

随着互联网技术的发展和普及,利用广泛开放的网络环境进行全球通信已经成为时代发展的趋势,人们日常的经济和社会生活也越来越依赖互联网,但是网络技术给人们带来巨大便利的同时也带来了各种各样的安全威胁,例如黑客攻击、特洛伊木马泛滥等. 目前市面上的许多杀毒软件都支持木马查杀,而且还有一些专业的木马查杀工具可供选择,但是这些杀毒软件和专业木马查杀工具普遍采用特征码检测.

基于特征码检测有一定的局限性,需要对木马进行跟踪、反汇编及其它分析,比较复杂,而且基于此技术的查杀机制对新型木马的检测迟滞或束手无策. 而基于木马行为特征的检测技术可以解决这一问题,根据木马行为特征的分析可以检测出新型木马. 此外,基于特征码检测所提取的特征码本身占用的空间开销较大,同时随着新型木马的不断涌现,提取的特征码持续更新,影响检测效率. 而木马的行为特征本身占用开销较小,检测效率较高,同时仍然具有可扩展性. 目前,一些主流的杀毒软件,如 NOD32、卡巴斯基、瑞星等,在使用特征码检测技术的同时也加入了对木马行为特征的分析,虽然只是作为辅助检测技术,对木马行为特征的归纳也不够全面,但这种检测机制正逐渐成为木马检测领域的发展趋势. 基于木马当前检测现状,我们搜集了大量的木马,并对其行为特征进行分析、提取,详细归纳了木马的行为特征,同时在此基础上提取木马通适性行为特征,构建了木马的行为特征库,设计了基于行为特征库的木马检测模型,并应用模糊模型识别方法判断木马程序.

1 木马行为特征库

1.1 木马行为特征分析 木马从入侵目标系统开始后,使用各种手段欺骗用户、对自身进行隐藏伪装,掩盖在目标系统中留下的任何蛛丝马迹,但是无论其如何伪装隐藏,都会表现出许多行为特征,并在目标系统中留下痕迹. 本文对木马植入阶段、木马安装阶段、木马进/线程启动运行阶段、木马网络通信阶段分别归纳了相应的木马行为特征,可参考文献[1-13],并在参考文献的基础上归纳融合了测试木马程序过程中的行为特征,具体如下.

1) 木马植入阶段. 木马进行攻击时,首先利用各种手段骗取用户的信任,将木马被控端植入到目标系统里. 在木马植入阶段,木马表现出的行为特征为:(a) 利用操作系统或一些常用软件的漏洞进行攻击植入;(b) 与病毒结合成复合的恶意程序植入;(c) 利用端口植入;(d) 利用交互脚本植入;(e)

收稿日期:2010-03-23
基金项目:四川省应用基础研究项目(07JY029-011)和四川省教育厅自然科学重点基金(08ZA043)资助项目
作者简介:李焕洲(1974—),男,副教授,主要从事网络监控和可信计算的研究

利用电子邮件植入;(f) 利用网络发送超链接,引诱用户点击等.

2) 木马安装阶段. 和其它阶段相比,木马在安装阶段存在显著的区别于一般合法程序的行为特征. 在该阶段,木马行为作用的主要对象之一是木马程序本身,因而,木马的安装阶段是检测与清除木马的最佳时机. 在木马安装阶段,木马表现出的行为特征为:(a) 自动压缩或者解压缩文件;(b) 文件自我删除;(c) 设置自启动;(d) 修改系统时间;(e) 关闭、增加或修改服务;(f) 修改系统配置文件;(g) 修改文件关联等.

3) 木马进/线程启动运行阶段. 木马被控端被成功安装到目标系统以后,在目标系统的运行空间中必须有一定的运行形式,如进程、线程等. 一般情况下,进程是能够从用户的角度观察到的,但是从用户的角度观察不到线程,因此,木马为了不被目标系统用户及管理员发现,会将其运行形式进行隐蔽. 在木马进线程启动运行阶段,木马表现出的行为特征为为:(a) 隐藏进程;(b) 调用 cmd 进程;(c) 关闭特定进程;(d) 利用远程线程等技术注入其它进程等.

4) 木马网络通信阶段. 木马被控端通常需要利用一定的通信方式与控制端进行信息交流(如接收控制者的指令、向控制端传递信息等),以达到控制目标系统、窃取系统文件等目的. 在木马网络通信阶段,木马表现出的行为特征为:(a) 采用 1024 以上的高端口;(b) 采用端口复用技术或端口寄生实现端口隐藏;(c) 反向连接;(d) 使用 ICMP 协议进行通信;(e) 采用“HTTP 隧道技术”,建立隐蔽通道,实现穿墙;(f) 木马被控端处于监听状态,等待其它进程通信;(g) 进程通信发送大量 SYN 包等.

1.2 木马通适性行为特征 木马程序为了达到伪装隐藏的目的,在植入阶段、安装阶段、进/线程启动运行阶段和网络通信阶段所表现的许多行为特征与许多常用合法软件的行为特征是一致或相似的,因此经过对大量木马程序与合法程序的对比实验及分析,并利用概率统计学中的均方差提取了明显有别于合法程序的木马行为特征. 计算过程如下.

第 1 步,提取所有训练样本的行为特征 $BC=\{BC_1,BC_2,\cdots,BC_i\}$,统计各行为特征 BC_i 在木马程序 T 中出现的频率 $E(BC_i^T)$ 及在合法程序 N 中出现的频率 $E(BC_i^N)$;

第 2 步,计算各行为特征的总出现频率均值

$$E(BC_i)=\frac{E(BC_i^T)+E(BC_i^N)}{2};$$

第 3 步,计算各行为特征的类间频率均方差

$$D(BC_i)=\{[E(BC_i)-E(BC_i^T)]^2+[E(BC_i)-E(BC_i^N)]^2\}^{1/2};$$

第 4 步,按照均方差大小排序,选取前 15 个组成模糊特征集,即木马通适性行为特征,同时划分了行为特征危险等级,如表 1 所示,并在此基础上构建木马行为特征库.

表 1 木马通适性行为特征

Table 1 The common trojan behavioral characteristics

序号	行为特征	危险级别
1	文件自我删除	高
2	自动压缩或者解压缩文件	高
3	拷贝或创建文件到系统目录	中
4	设置自启动	中
5	修改系统时间	高
6	文件关联	高
7	使用 ICMP 通信	高
8	隐藏进程	高
9	关闭特定进程	高
10	调用 cmd 进程	中
11	自动发送邮件	高
12	伪装系统进程或路径通信	高
13	绑定侦听端口	中
14	注入系统进程或 IE 进程	高
15	关闭、增加或修改服务	高

2 基于木马行为特征库检测模型

2.1 模糊模式识别数学模型[14-18] 模糊模式识别数学模型的基本思想:设有 n 个程序样本,组成程序样本集合 $X=\{x_1,x_2,\cdots,x_n\}$,每个程序样本都会有 m 个指标特征向量,表示为 $\boldsymbol{x}_j=\{x_{1j},x_{2j},\cdots,x_{mj}\}^{\mathrm{T}}$,则程序样本集可用 $m\times n$ 阶指标特征矩阵表示如下

$$\boldsymbol{X}_{m\times n}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\cdots&\cdots&\cdots&\cdots\\x_{m1}&x_{m2}&\cdots&x_{mn}\end{bmatrix}=(x_{ij}),$$

其中,x_{ij} 表示程序样本 j 指标 i 的特征值,$i=1,2,\cdots,m$;$j=1,2,\cdots,n$. 为了消除 m 个指标特征值在量纲和量级上的差异,在进行识别时要先消除指标特

征值量纲的影响,使指标特征值规格化,即将指标特征值矩阵变换为指标特征值的相对隶属度矩阵如下

$$\boldsymbol{R}_{m\times n}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}=(r_{ij}),$$

其中,r_{ij}表示指标特征值规格化数或相对隶属度,$0\leqslant r_{ij}\leqslant 1$.

设将 n 个程序样本依据样本的 m 个指标特征按 c 个类别加以识别,其模糊识别矩阵为

$$\boldsymbol{U}_{c\times n}=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ u_{c1} & u_{c2} & \cdots & u_{cn} \end{bmatrix}=(u_{hj}),$$

其中,u_{hj}表示样本 j 从属于类别 h 的相对隶属度. 另外矩阵满足条件

$$\begin{cases} \sum_{h=1}^{c} u_{hj}=1, \quad \forall j, \quad 0\leqslant u_{hj}\leqslant 1, \\ \sum_{j=1}^{n} u_{hj}>0, \quad \forall h. \end{cases}$$

2.2　数学模型在基于木马行为特征库检测中的应用

基于木马行为特征检测是一个二分值分类问题,类别分为木马程序与合法程序. 应用模糊模式识别数学模型,首先应收集一定数量的木马程序样本与合法程序样本,根据木马通适性行为特征,提取木马程序集(定义为 T)与合法程序集(定义为 N)相应的指标特征矩阵,并利用行为特征出现的频率,结合隶属函数,计算求得其隶属度矩阵. 然后提取待测程序样本(定义为 M)的行为特征,计算求得待测程序样本的隶属度矩阵. 最终,在木马程序样本与合法程序样本的隶属度矩阵的基础上,构造待测样本隶属于木马程序样本与合法程序样本的模糊识别矩阵,判别该待测程序样本是木马程序还是合法程序.

现有 m 个木马程序样本,n 个合法程序样本(m,n 均为正整数),参照表 1 的 15 个通适性行为特征,每个程序样本都会具有其中的 $l(0\leqslant l\leqslant 15)$ 个行为特征. 如果程序样本具有第 $k(1\leqslant k\leqslant 15)$ 个行为特征则定义为 1,不具有则定义为 0,因此木马程序样本集 T、合法程序样本集 N 可用指标特征矩阵分别表示为

$$\boldsymbol{T}_{15\times m}=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ t_{15,1} & t_{15,2} & \cdots & t_{15,m} \end{bmatrix}=(t_{ij}),$$

$$\boldsymbol{N}_{15\times n}=\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1n} \\ n_{21} & n_{22} & \cdots & n_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ n_{15,1} & n_{15,2} & \cdots & n_{15,n} \end{bmatrix}=(n_{ij}),$$

其中,如果 $t_{ij}=1$($n_{ij}=1$)表示木马程序样本(或合法程序样本)j 具有第 i 个行为特征,反之 $t_{ij}=0$($n_{ij}=0$),表示不具有第 i 个行为特征,$i=1,2,\cdots,15$; $j=1,2,\cdots,m$(或 $j=1,2,\cdots,n$).

根据指标特征矩阵,利用每个行为特征出现的频率及正态分布偏大型模糊分布构造的隶属函数,计算出木马程序样本集与合法程序样本集对每种行为特征的隶属度,得出相对隶属度矩阵,表示为

$$\boldsymbol{TR}=[r_1^T,r_2^T,\cdots,r_i^T,\cdots,r_{15}^T],$$

$$\boldsymbol{NR}=[r_1^N,r_2^N,\cdots,r_i^N,\cdots,r_{15}^N],$$

其中,r_i 表示程序样本对于第 i 个行为特征的相对隶属度,$0\leqslant r_i\leqslant 1$.

提取待测程序样本 M 的行为特征,并计算得到隶属度矩阵,表示为

$$\boldsymbol{MR}=[r_1^M,r_2^M,\cdots,r_i^M,\cdots,r_{15}^M].$$

将待测的 x 个程序样本依据 15 个通适性行为特征的隶属度,按 2 个类别(类别分为木马程序和合法程序)加以识别,其模糊识别矩阵为

$$\boldsymbol{U}_{2\times x}=\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}=(u_{hj}),$$

其中,u_{hj}表示待测程序样本 j 从属于类别 h 的相对隶属度. 最后,根据隶属度判断待测程序样本为木马程序还是合法程序. 其中,获得相对隶属度矩阵过程中,借鉴了病毒程序计算隶属度的思想方法[8],主要利用行为特征出现的频率及正态分布偏大型模糊分布构造的隶属函数,具体如下.

某行为特征在木马文件中出现的频率为 $E(BC_i^T)$,在合法程序中出现的频率为 $E(BC_i^N)$,则用正态分布中的偏大型模糊分布来构造木马程序集 T 的隶属函数

$$r_i^T(E(BC_i^T))=$$

$$\begin{cases} 0, & E(BC_i^T) < 0, \\ 1 - \exp\left(-\dfrac{(E(BC_i^T))^2}{\sigma^2}\right), & E(BC_i^T) \geqslant 0, \end{cases}$$

式中,$\sigma = \max\{E(BC_1^T), E(BC_2^T), \cdots, E(BC_i^T)\}/q$(其中 q 为通适性行为特征的个数). 同样构造合法程序集 N 的隶属函数为

$$r_i^N(E(BC_i^N)) =$$

$$\begin{cases} 0, & E(BC_i^N) < 0, \\ 1 - \exp\left(-\dfrac{(E(BC_i^N))^2}{\sigma^2}\right), & E(BC_i^N) \geqslant 0. \end{cases}$$

对于待检测程序样本 M,其隶属函数为

$$r_i^M(E(BC_i)) =$$

$$\begin{cases} 0, & BC_i < 0, \\ 1 - \exp\left(-\dfrac{(BC_i)^2}{\sigma^2}\right), & BC_i \geqslant 0. \end{cases}$$

3 实验过程及分析

3.1 实验过程 实验是在 Windows 2000 操作系统下进行的,在获取程序行为特征时需要首先启动运行程序样本,由于运行过程中程序样本可能对计算机系统造成破坏,因此实验是在虚拟机环境下进行的,虚拟机选择的是目前实验较多的 VMWare Workstation 6.0. 实验运用的算法如下.

训练算法: 1) 通过运行启动程序及检测软件获得样本程序的行为特征; 2) 计算各行为特征的类间频率均方差 $D(BC_i)$; 3) 根据类间频率均方差大小,选取前 15 个作为通适性行为特征; 4) 根据各行为特征在木马程序样本集及合法程序样本集中出现的频率计算其相应的隶属函数,并获得相应的模糊集.

分类算法: 1) 通过运行启动程序及检测软件获得待测程序样本的行为特征; 2) 计算待测程序样本的隶属函数,并获得相应的模糊集; 3) 依据木马程序样本集及合法程序样本集的模糊集对待测文件进行分类.

3.2 实验分析 实验前共收集样本数 165 个(其中合法程序 79 个,木马程序 86 个). 实验根据训练算法,首先选取 80 个样本(其中合法程序 40 个,木马程序 40 个)进行训练并自学习. 根据训练样本获得数据完善训练算法,利用另外 85 个样本作为测试样本(其中合法程序 39 个,木马程序 46 个)进行分类测试,测试样本实验结果如图 1.

实验的目标主要是对模型的误报率和漏报率进行测试,其中误报是指将合法程序判断为木马程序,漏报是指将木马程序判断为合法程序. 实验时,随机选取 10 的倍数的样本数,并计算其误报率或漏报率,作为主要实验绘图数据,其中

误报率 = 被误判为木马程序样本数 / 样本总数,

漏报率 = 被漏判为合法程序样本数 / 样本总数.

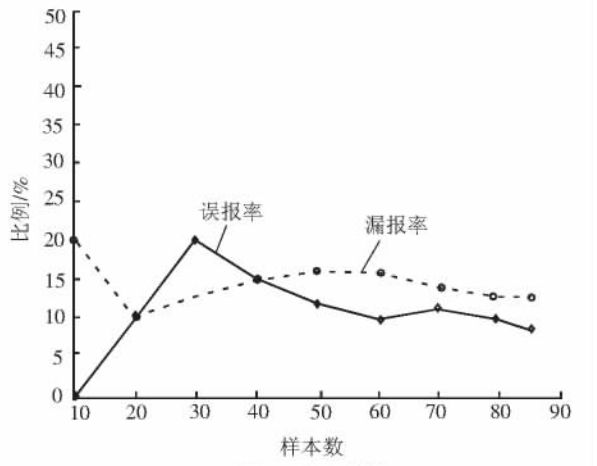

图 1 实验结果

Fig. 1 Experimental results

通过对实验结果的分析,选取的木马通适性行为特征较为准确,误报率与漏报率较低. 同时实验结果表明,选取样本数的多少对木马误报率与漏报率的影响较小. 在 30 个随机样本数以内,木马的误报率是随着样本数的增加而增长的,但之后随着样本数量的增加,木马的误报率波动基本不大; 木马的漏报率是在 20 个随机样本数时达到漏报率的低点,之后随着样本数的增加,漏报率基本趋于平缓. 因此,这对于在木马样本较难获得的现实情况下,对木马行为特征提取的准确性的验证是非常有利的,根据较有限的样本提取通适性行为特征,可以较为准确地设计木马行为特征库.

4 结语

本文引入基于木马行为特征库检测模型,同时提取木马通适性行为特征,是对基于特征码检测木马的一种强有力的补充,在效率、占用开销、更新率、检测未知木马等很多方面优于特征码检测. 木马行为特征的研究代表着未来木马检测技术的必然发展趋势,在新型木马不断涌现的今天,基于木

马行为特征检测技术具有重要的应用价值和意义.

参考文献

[1] 李伟斌,王华勇,罗平. 通过注册表监控实现木马检测 [J]. 计算机工程与设计,2006,27(12):2220 - 2222.
[2] 蔺聪,黑霞丽. 木马的植入与隐藏技术分析 [J]. 信息安全与通信保密,2008(7):53 - 55.
[3] 郝向东,王开云. 典型恶意代码及其检测技术研究 [J]. 计算机工程与设计,2007,28(19):4639 - 4641.
[4] 商海波. 木马的行为分析及新型反木马策略的研究 [D]. 杭州: 浙江工业大学图书馆,2005.
[5] 张超,孙晓燕,徐波. 基于主动防御的网络病毒防治技术研究 [J]. 网络安全技术与应用,2009(1):31 - 32.
[6] 康治平. 特洛伊木马可生存性研究及攻防实践 [D]. 重庆: 重庆大学图书馆,2006.
[7] 罗晓波,王开建,徐良华. 基于行为分析的主动防御技术及其脆弱性研究 [J]. 计算机应用与软件,2009,26(7):269 - 371.
[8] 侯明明. 浅析"木马"病毒及其防治措施 [J]. 广西轻工业,2009(3):79 - 80.
[9] 邓璐娟,刘涛,甘勇,等. 基于进程鉴别和隐藏的病毒主动式防御技术 [J]. 计算机工程,2007,33(5):117 - 119.
[10] 胡卫,张昌宏,马明田. 基于动态行为监测的木马检测系统设计 [J]. 火力与指挥控制,2010,35(2):128 - 132.
[11] 姜坚,袁家斌. 基于特征行为的远程访问型木马阻断技术 [J]. 计算机与数字工程,2008,36(11):90 - 93.
[12] 张文达. 基于行为识别技术的网络防御系统研究与实现 [D]. 上海: 同济大学图书馆,2008.
[13] 陈婧婧,李焕洲,唐彰国,等. 木马运行机制及行为特征分析 [J]. 计算机安全,2009(10):108 - 110.
[14] 顾雨捷. 用于行为分析反木马的模糊分类算法研究 [D]. 杭州: 浙江工业大学图书馆,2008.
[15] 谢季坚,刘承平. 模糊数学方法及其应用 [M]. 武汉: 华中科技大学出版社,2005.
[16] MarquesdeSá J P. 模式识别——原理、方法及应用 [M]. 吴逸飞,译. 北京: 清华大学出版社,2002.
[17] 张波云,殷建平,唐文胜等. 基于模糊模式识别的未知病毒检测 [J]. 计算机应用,2005,25(9):2050 - 2054.
[18] 郭旭亭. 基于程序行为模糊模式识别的病毒检测研究 [D]. 青岛: 青岛大学图书馆,2008.

Design of Trojan Horse Detection Model Based on the Behavioral Library

LI Huan-zhou[1], CHEN Jing-jing[1,2], ZHONG Ming-quan[1], TANG Zhang-guo[1]

(1. *Research Institute of Network and Communication Technology, Sichuan Normal University, Chengdu* 610101, *Sichuan*;
2. *College of Computer Science, Sichuan Normal University, Chengdu* 610101, *Sichuan*)

Abstract: At present, the mainstream technology of Trojan horse detection is the characteristic code detection technology, but this technology lags behind extracting characteristic code, it can not detect the new Trojans. In order to detect the new Trojans better, by summing up the Trojan behavioral characteristics in detail, and constructing the library of Trojan behavioral characteristics, the Trojan detection model which based on the behavioral library has been designed. It can judge Trojan making use of fuzzy pattern recognition. Experiments prove that this model can analyze and judge the behavior of suspicious program, identify the Trojan. This detection technolgy has great value in practical application.

Key words: trojan horse; behavioral characteristics; fuzzy pattern recognition

(编辑　李德华)