高技术通讯 2012 年 第22卷 第9期:897~903

doi:10.3772/j.issn.1002-0470.2012.09.002

一种基于行为关联的主机系统入侵检测方法①

王映龙②* 李京春** 王少杰** 锁延锋*** 梁 利** 郭瑞龙*

(*江西农业大学软件学院 南昌 330045)
(**国家信息技术安全研究中心 北京 100084)
(***北京科技大学信息工程学院 北京 100083)

摘 要 提出了一种基于主机行为解析和行为关联分析的主机系统入侵检测方法,对嵌入式恶意软件具有较高的检测效率,可应用于基于网络行为的入侵检测系统。通过对行为进行深层次的解析,建立了行为间的关联关系模型,在降低存储异常行为样本规模的同时,提高了该方法的灵活性和应用范围。实验结果显示,与现有的异常行为检测方法相比,该方法需要较长的训练时间,但是,通过调整行为粒度,该方法可以使训练时间保持在合理的范围之内。随着时间的推进,该方法的性能将逐步提高,在漏报率、误报率及更新效率上,较现有系统都有较大的提高。

关键词 关联分析,主机系统,入侵检测系统(IDS)

0 引 言

网络的日益普及,使得人们对计算机和网络的依赖程度逐步提高,例如,越来越多的应用软件来源于网络,越来越多的人习惯把个人或企业的隐私信息存储在本地主机或网络上。但是计算机的安全防护水平参差不齐,其中多数习惯于依赖安全防护软件(如防火墙、入侵检测系统(IDS)、杀毒软件等)来保护主机的安全,因此,安全软件的工作效率在一定程度上直接影响网络系统的安全,致使关于系统异常行为检测和入侵检测的研究一直成为学术界和企业界研究的热点。

异常检测方法和技术已成为异常行为检测和入侵检测的主要研究方向,基本轮廓包括以下两个方面:(1)建立系统或用户的正常行为模式;(2)通过被监测系统或用户的实际行为模式和正常模式之间的比较和匹配来检测入侵。异常检测方法不需要有关系统缺陷的知识,具有较强的适应性,并且能够检测出未知的入侵模式。近年来,关于以 shell 命令或系统调用为审计数据的异常检测的研究和应用[1-10],已形成了一系列细粒度的系统行为解析方法,如基于虚拟机的系统指令解析,系统调用序列采集等。但是,由于该方法需要存储大量的正常行为模式,并具有较高的误报率,其扩展性(低重量级的正常行为表示方式)和更精准的异常分析方法一直是目前研究的重点。在总结现有异常行为检测和系统行为解析方法的基础上,本文提出了一种基于 Markov 模型的异常行为检测方法。该方法通过对正常行为进行分解,建立各原子行为间的转移关系,确定正常行为的行为特征(状态分布函数),并以状态分布表现正常行为。其后,通过给出采样窗口和采样频率的确定方法,对系统行为进行实时采样和分析,计算其特征值,实现对异常行为的判断和检测。实验验证了该方法的有效性,本文也分析了该方法的局限性。

1 相关工作和研究现状

关于主机入侵检测的方法可以分为误用特征检测和异常检测两类,所谓误用特征检测是通过比对收集的有关入侵或攻击的特征进行检测,异常检测是通过比对正常状态下系统的观察结果进行检测,其中,行为特征提取的准确度和检测窗口规模的选择直接影响到入侵检测效果。下面,我们从多个角度对相关研究工作做简要的综述。

① 国家发改委信息安全专项(发改办高技[2010]3044号)和江西省科技厅国际合作计划(2009BHB15100)资助项目。
② 男,1970年生,博士,教授;研究方向:入侵监测,主机防护,数据挖掘,知识工程等;联系人,E-mail: wangyinglong@sohu.com
(收稿日期:2011-08-08)

1.1 行为特征提取

行为特征的提取主要包括两种方法：系统数据审计分析和基于行为的模型验证。主机数据的审计分析包括以系统调用为审计数据的分析和以 shell 命令为审计数据的分析。程序行为是目前主要的数据审计分析对象之一，其主要原因在于程序行为主要依赖于系统的调用指令，相对比较单一，具备了审计分析的行为条件[1]。利用系统调用对特权程序行为分析已有较好的应用，但是基于系统调用的程序行为审计分析不能用于检测用户帐号假冒等攻击行为[2]。另外，shell 命令是另外一个重要的数据审计分析对象，如文献[11-13]，其主要原因在于：(1)用户的多数活动需利用 shell 完成，反映出用户的行为模式；(2)shell 命令较容易获取。

行为模型验证的研究，是一种较为经济的方法之一，模型方法的引入在一定程度上减弱了对实际数据的依赖程度，为进一步分析异常提供了基础。1996 年，Forrest 等提出了基于特权进程监控的入侵检测模型，通过对特权进程的系统调用序列进行实时检测和分析来实现入侵行为的检测[3]，减少了被监测对象的规模，实验证明其检测效果仍保持在较高的水平。Warrender 等[4]提出了基于隐 Markov 链的模型，把计算机系统行为模型为一个不可观测的 Markov 链和一个与之相关联的可观测链，通过采用统计分析较好地刻画了系统的可观测行为特征，使得使用该模型建立的入侵检测系统具有较高的检测效率。谭小彬等[5]在以往对有关基于隐 Markov 链的模型的研究的基础上，简化了该模型，并引入了遗忘因子来解决不同时刻观察值权重设置问题。尹清波等[6,7]为了克服 Markov 链刻画行为粗糙的弱点，针对多输入多输出状态，定义了状态克隆，提出了动态 Markov 链模型和线性预测与 Markov 模型。

此外，关于行为特征的提取还有贝叶斯网络、贝叶斯聚类、频率统计和数据挖掘等方法[8-10]，由于篇幅限制，在此不再赘述。

1.2 抽样窗口选择

抽样窗口是影响入侵检测系统准确性和效率的一个重要因素，它决定了一次采样可获得的信息量。窗口宽度过大势必导致大的性能开销和引入过多的冗余信息，过小则无法保证信息的完整性，导致漏报率的增加。关于抽样窗口的选择已成为设计入侵检测系统不可回避的一个问题，在文献[7]中，作者论述了窗口选择应满足准确性和随机性的要求，给出了一种基于 Markov 信源熵的选择方法；在文献[6]中，作者利用滑动窗口生成相同长度的短序列用于检测系统的数据输入，说明了通过恰当选择窗口大小，可以较高置信度地反映原始系统的行为状态，但缺少关于窗口大小如何选择的论述；文献[5,12,13]指出了窗口选择对检测结果的影响，并说明了应考虑多种因素的影响，但对于如何选择窗口没有给出论述。据我们所知，目前关于抽样窗口的研究还相对薄弱，关于窗口选择的形式化分析还是一个公开问题。

1.3 主机入侵检测

根据分析数据来源的不同，目前基于主机数据的入侵检测的研究可以分为基于系统调用的异常行为检测和基于 shell 命令的异常行为检测[1,6,7,13]。事实上，相对复杂的用户行为，程序运行所产生的行为序列较为单一，如何从简单的程序序列中挖掘出所对应的用户行为已成为了影响入侵检测效果的重要因素之一。在文献[1]中，作者通过确定系统的特权行为，建立了一种较为简便的基于系统调用行为的数据的用户异常行为分析模型，但是，由于系统调用行为来自系统的核心层，其获取复杂度较高，且占用较大的系统资源，并且其准确程度依赖于特权行为的选取，因此，该模型的实用效果较差。为摆脱主观因素的影响，隐 Markov 链模型[5]、时间序列模型[8]、状态向量机模型[10,14]等方法得到了广泛的研究。针对基于系统调用序列难以检测出类似用户帐号假冒等行为的不足，以 shell 命令为审计数据的异常检测得到了广大学者的重视，在文献[13]中，作者针对用户假冒行为给出了一套较为可行的分析方法。

尽管关于主机入侵检测的研究取得了较多的成果，但主要的研究成果多基于较为完整的审计事件，对系统资源的耗用开销较大。基于此，我们在借鉴已有研究成果的基础上，从数据的精简处理和行为关联模型两个方面综合考虑，给出一种基于行为关联关系的入侵检测方法。

2 基于 Markov 链的异常行为关联分析

Markov 链模型是一种高效、简便且应用广泛的随机模型之一，得到了数学家们的广泛研究，给出了许多较好的随机性质，形成了较为完善的理论体系。由于其支持稳态分析(时间 $t \to \infty$)和瞬态分析($t = T^*, T^* < \infty$ 为一选定值)，因此，被广泛应用于量化分析。在主机入侵检测领域，各种 Markov 链模

型被引入,并取得了较好的效果[6,7]。

通过实验我们发现,假设相邻行为满足 Markov 条件,Markov 链模型能较高效率地用于行为间关联性的判断,是建立高效主机入侵检测系统的可行选择之一。为后续叙述的方便,我们首先引入 Markov 的形式化定义和相关术语。

2.1 Markov 链及相关术语

定义 1 (Markov 链,MC)已知 $X(t)$ 为一随机过程,表示在时刻 t 的状态分布,若 $\Pr\{X(t+1) \mid \bigwedge_{k\le t} X(k)\} = \Pr\{X(t+1) \mid X(t)\}$,那么随机过程 $X(t)$ 被称为 Markov 链。若 t 为离散时间,称为离散的 Markov 链;若 t 为连续,那么称为连续的 Markov 链。

定义 2 (状态,State)已知 $X(t)$ 为一 Markov 链,那么 $X(t)$ 的取值称为该 Markov 链的状态,所有状态构成该 Markov 链的状态空间 S。若 $|S| < \infty$,称该 Markov 链为有限 Markov 链;若 $|S| = \infty$,称该 Markov 链为无限 Markov 链。

已知状态 s,若 $\lim_{t\to\infty}\Pr\{X(t) = s\} > 0$,那么,状态 s 称为常返的,否则,称为消失状态或滑过状态。

定义 3 (时间齐次 Markov)已知 $X(t)$ 为一 Markov 链,且初始状态为 $X(0)$,若存在时刻 $t_0 > 0$,使得对于 $t \ge t_0$,都有 $\Pr(X(t+2) \mid X(t+1)) = \Pr(X(t+1) \mid X(t))$,那么,$X(t)$ 被称为齐次 Markov 链。

定理 1 (1)若 $X(t)$ 为一有限的 Markov 链,那么,$X(t)$ 为齐次 Markov 链;(2)若 $X(t)$ 为一无限的 Markov 链且仅存在有限个常返状态,那么,$X(t)$ 为齐次 Markov 链;(3)齐次 Markov 链 $X(t)$ 存在稳态分布,可用常返状态的无限生成矩阵获得。特殊的,若仅存在一个常返状态,那么,稳态分布即为该状态。

证明:参见随机过程相关教材[14]。

2.2 基于 Markov 链的行为关联关系分析模型

根据微机原理,我们知道系统行为对应着由有限系统调用指令组成的有序指令列,而用户的行为可以表示为若干由于系统行为组成的序列,因此,如果以有限的指令集为原子行为,那么,针对用户行为可以表示为由原子行为构成的有序序列,可以通过调整采样窗口来解析系统行为,如图 1 所示。

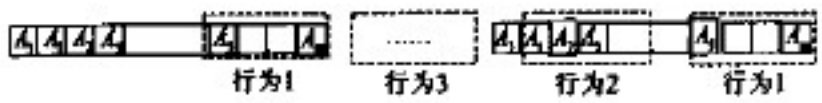

图 1 用户行为与系统行为

针对解析出的用户行为,根据系统设计和用户行为的目的可以发现正常行为间或异常行为间相互的关联关系。关于系统行为间的关联关系的分析,一般可以采用以下两种方式判断:(1)基于行为关联假设的判断方法;(2)基于行为独立假设的判断方法。

基于行为关联假设的判断方法是指在假设行为存在关联关系的基础上,即行为间存在一定的相互转移关系,以行为为状态建立状态间相互转换的 Markov 链模型,其中状态间的转移速率(或概率)表示两行为之间的关联强度。

事实上,通过对正常用户行为的解析不难发现该行为对应的系统行为序列,并假设它们之间满足关联关系;通过训练所有的已知正常用户行为,可以得到所有涉及系统行为的关联关系集合,通过采用数据挖掘、神经网络等方法得到基于系统行为的关联模型,这里我们采用各系统行为出现的频率关系表示。接下来,我们给出基于系统行为原子序列构造转移关系规则,并简单描述基于用户行为原子序列的 Markov 链行为关联分析的模型。

规则 1 若系统行为 X 对应的原子序列为 $x_1x_2\cdots x_n$,那么,x_i 与 x_{i+1} 相关联,其中 $1 \le i < n$,称 x_ix_{i+1} 为一关联对。

规则 2 若系统行为 X 对应的原子序列包括为 $x_1x_2\cdots x_{i-1}x_ix_{i+1}x_{i+2}\cdots x_n$ 和 $x_1x_2\cdots x_{i-1}x_{i+1}x_ix_{i+2}\cdots x_n$,那么,$x_i$ 与 x_{i+1} 不相关联。

规则 3 系统行为 X 对应的原子序列 $x_1x_2\cdots x_n$,关联对集合为 R,若 $s \in R$ 且在原子序列中存在 k_s 个 s 关联对,我们称 k_s 为 s 的强度,则 s 的归一化强度为 $\bar{k}_s = \dfrac{k_s}{\max_{\{t\in R\}}\{k_t\}}$;对于系统行为序列 $U = X_1X_2\cdots X_m$,基于 U 的 s 的归一化强度为 $\bar{k}_s^U = \sum_X \bar{k}_s^X \dfrac{|\{X_i \mid X_i = X, i = 1,2,\cdots,m\}|}{m}$,其中,$\bar{k}_s^X$ 为系统行为 X 的归一化强度。

规则 4 正常用户行为的集合为 $U = \{U_1, U_2, U_3, \cdots, U_p\}$,那么,基于该用户行为集合 s 的归一化强度为 $\bar{k}_s = \sum_U \bar{k}_s^U \dfrac{|\{U_i \mid U_i = U, i = 1,2,\cdots,p\}|}{p}$。

根据上述规则,我们可以得到基于正常用户行为的原子序列间的关联关系,由于关联对的有序性,我们不妨以得到的关联强度表示它们之间的转移强度,以原子行为为状态可以得到相应的随机模型,若假设原子行为间的转移满足指数分布,那么,我们便

得到了 Markov 链模型。由于系统原子行为的有限性，根据定理 1，我们得到在该种假设下得到所有状态的稳态分布，比照行为序列对应的系统序列，不难计算出各原子行为所占的比率，若在阈值范围内，则可以判断该行为间存在关联关系，否则，可以认为行为间相互独立。

基于行为独立的假设的判断方法是指在假设行为间相互独立的条件下，以行为为状态构造 Markov 链模型，仅考虑行为对应序列内部的原子行为存在转移关系（忽略用户行为对应系统行为的关联关系）；通过设置虚拟的初始位置连接个行为的首个原子位置和结束原子位置，并使得虚拟位置到连接位置的转移强度相同。根据定理 1，可以得到所有原子状态的稳态分布，比对行为序列对应的系统行为序列，若高于基于稳态分布的阈值，则判断行为间存在关联关系，否则，不存在关联关系。

上述两种方法中前一种依赖于较为复杂的计算和假设行为间的关联强度，而后一种，则需要慎重选择判断关联关系的行为对或行为序列。事实上，若是空间和时间允许，可以混合使用上述方法以提高判断的准确性，在本文中，我们采用了第一种方法。

2.3 基于 Markov 链的异常行为分析算法

通过对行为间关联关系的分析，我们不难得到异常或正常行为的特点，这里我们采用原子状态的稳态分布作为行为特征，通过采用抽样分析方法可以判断特定行为出现与否。若该方法应用于异常行为分析，我们称该方法为基于 Markov 链的异常行为分析算法，其详细算法设计见算法 1。

算法 1 基于 Markov 链的异常行为分析算法

输入：系统行为序列，判断阈值，异常行为关联模型稳态分布

输出：是否存在异常行为

1. 确定正常行为的关联模型；
2. 计算正常行为关联模型的稳态分布；
3. 根据阈值和稳态分布执行子算法 1 确定异常报警准则；
4. 对行为序列采样，若符合报警准则，执行子算法 2 激发异常行为警报。

子算法 1 异常行为判断准则确定

1.1 确定关联行为序列的长度；

1.2 计算 Markov 模型中稳态分布；

1.3 计算关联序列出现的频率，即长度除以关联序列长度乘以该序列状态的稳态概率；

1.4 遍历所有的关联序列，得到所有正常行为特征阈值。

子算法 2 异常行为报警

2.1 输入系统行为序列；

2.2 确定采样窗口，及各阈值出发条件；

2.3 if（采样窗口乘以采样次数小于关联行为序列长度）；

2.3.1 如果关联序列条件满足，出发异常警报；

2.3.2 否则，比对其他满足条件关联序列；

2.3.3 end if；

2.4 If（采样窗口乘以采样次数大于关联行为序列长度）

2.4.1 关联序列基数从该采样开始计数，执行 2.3；

2.5 结束。

分析上述算法，我们不难发现抽样策略的选择是影响异常行为检测效率的另一个主要因素，其采用窗口的大小及采样频率的选择将直接影响着分析结果的准确性。在本文中，我们采用了综合考虑时间和频率的方法，并通过实验分析的方法说明了抽样策略对分析结果的影响。

针对系统调用序列的抽样，我们从采样的频率和采样的时间长度两方面综合考虑，其准则如下：

（1）覆盖单用户行为对应的系统行为序列，这样可以保证抽样获得尽量多的完整行为，而且避免了过分的抽样资源的消耗。

（2）抽样频率不大于主机处理频率的 1/3，根据抽样原理，超过一定界限的抽样数据的增加不能获得更多的信息，那么采用三分之一原则基本保证了抽样的准确性。

（3）随机抽样，针对隐藏隧道攻击，攻击者把攻击行为均匀地分布于正常行为之中，随机抽样可以较好地保证对该类攻击的报警。

3 基于 Markov 链的主机异常检测系统设计

通过采用随机过程方法确定各原子行为间的相互关联关系，完成组合行为特征和属性判断，为实施异常检测确定了比对基准；随机抽样可以实时得到系统的运行状况，通过计算相应的属性值，并于基准比对，可以实现异常行为的检测。在该部分，我们给出本文使用的基于系统行为关联关系的主机行为异常检测系统的设计。

特征计算模块：采用文中提出的关联关系计算模型，实现正常行为的关联关系的分析，并以相关联行为的稳态分布为关联关系的特征指标。

数据收集模块:由于用户行为不仅可能是一个主机上的多个进程行为,也可以是在一个局域网(甚至是 Internet)中的多个终端上运行的进程行为,通过引入时间戳,使得数据具有时间特性,并扩充了分析的范围,其行为及分析流程如图 2所示。

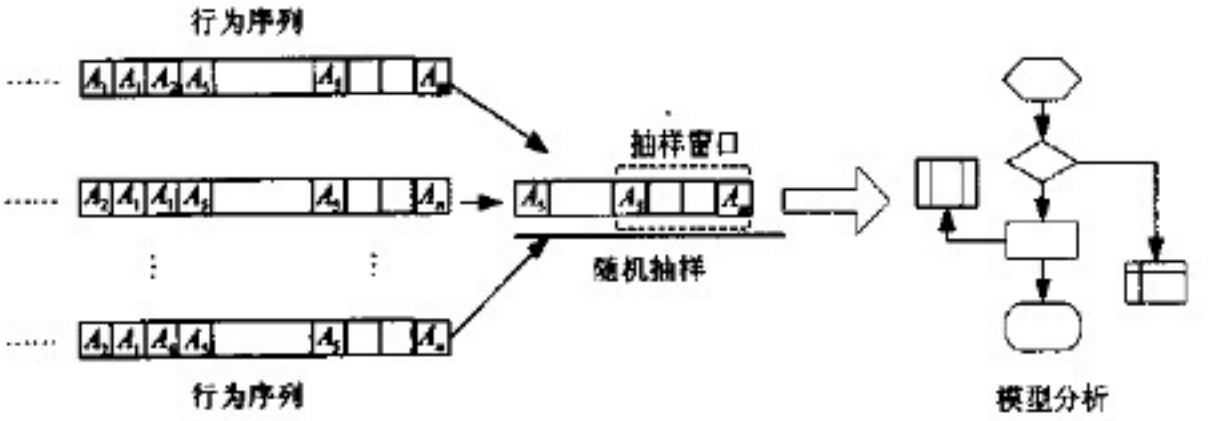

图2 异常行为检测系统设计

这里我们采用的模型为 Markov 链模型,针对原子行为的关联分析,以连续 Markov 链模型为例,其模型如图 3 所示,其中,针对原子行为 $Act = \{A_1, A_2, \cdots, A_n\}$,$A_m$ 为隐患原子行为,A_n 为恶意原子行为,其余为合法的原子行为。原子行为间合法行为序列的发生速率为 λ,α,μ 和 τ;隐患原子行为的回归合法行为的速率为 α;发生恶意行为的速率为 λ_2。我们可以得到如下连续 Makov 链模型,因此,在已知初始原子行为的条件下,可以通过计算模型得到各行为在时刻 t 发生的概率 $\Pr(A_i, t)$,若 $t \to \infty$,则为稳态概率分布,通过比对抽样值可以判断该行为是否为恶意行为。

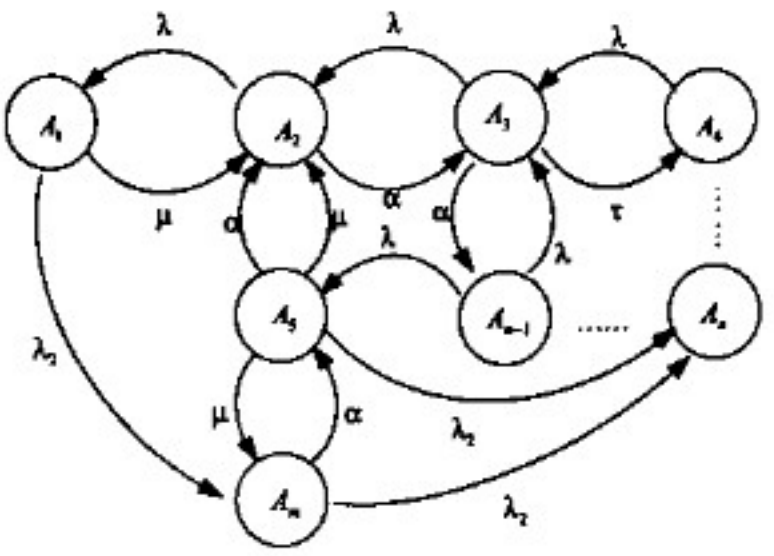

图3 原子行为间连续 Markov 模型

4 实验分析与讨论

通过收集大量的测试数据,我们对本文提出的入侵检测模型进行了实验数据分析和测试,将以收集到的 tel、telnet 和 telcmd 日志文件为数据源展开分析。

以 telcmd. log 为例,日志格式如下:CALL name:[C: Telnet Server telcmd. exe], cr3:[0xfc2a000] pid:[2036], NtRequestWaitReplyPort (0x7ec, 0xaffc70, 0xaffc70) RET 0x0 = name:[C: Telnet Server telcmd. exe], cr3:[0xfc2a000] pid:[2036], NtRequestWaitReplyPort (0x7ec, 0xaffc70, 0xaffc70)。每一条完整的日志记录了原地址、目的地址、访问端口、访问内容等信息。我们通过引入分隔符"&",提出格式信息得到的原子行为序列为 &&0xfc2a000&2036&, 0x7ec $ 0xaffc70 $ 0xaffc70& 0xfc2a000&2036&0x7ec $ 0xaffc70 $ 0xaffc70,其中"&&"表示新日志开始,"&"表示相应调用序列分割符,"$"表示内容分割符。

我们在不同的时间段、在不同的机器上采取了同一种行为的 3000 条数据作为该行为的训练数据,用以得到正常行为对应的原子行为关联特征。针对每一种命令文件日志,我们设计了多种入侵日志,通过随机选取 5000 条日志文件(其中包含入侵日志),采用本文提出的检测系统对插入的入侵日志进行检测。我们分别检验了采样窗口、采样频率对检测结果的影响,及该系统的实际性能。

用 W 表示抽样窗口的长度,F 表示抽样的频率(次/s)。在检测系统的原型系统中,我们设定每次抽样的起始位置标志为"&",即每次采样以该开始符为标志,采样相应长度的序列并记录。为了保证检测结果的可用性,我们选取的判断准则为模型计算值的浮动值,如 1 + 10% 表示抽样值在理论值左右 10% 范围内都为合理。本文的实验数据序列为正常序列中随机插入了两种攻击行为(分别用输出 1 和 -1 表示,正常行为用 0 表示)。

通过观察图 4 发现,1 +5% 对应的曲线对应的报警一般早于 1 + 10% 对应的曲线,而且后者的报

警持续时间比前者更长，因此，我们知道宽松的准则势必导致漏检率的升高和异常报警的延迟增大，与预期一致。但是最佳准则的选择还不能给出相应的准则。

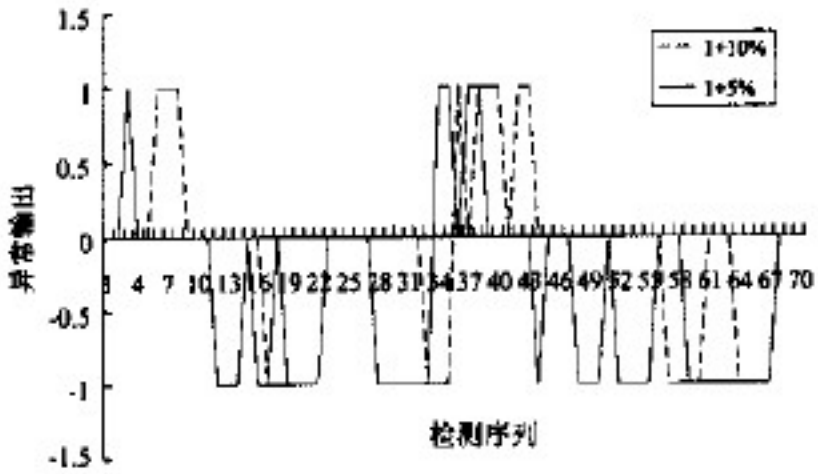

图 4 检测准则对检测结果的影响（$W=30, F=10$）

图 5、图 6 和图 7 显示了抽样窗口和抽样频率对检测结果的影响情况。图 5 说明对于固定的窗口长度，抽样频率与成功检测率成正比；图 6 说明对于固定的抽样频率，窗口越大，对应的成功检测率越高；图 7综合对比了 $W = 30$ 的情况下，抽样频率对成功检测率的影响。

另外，我们针对隐藏攻击也做了相应的实验，结果显示，通过随机抽样（$W = 10$, $F = 30$），可检测的概率保证在 40% 左右，但通过增加抽样序列的存储长度，其成功检测概率可以提高到75% 以上。通

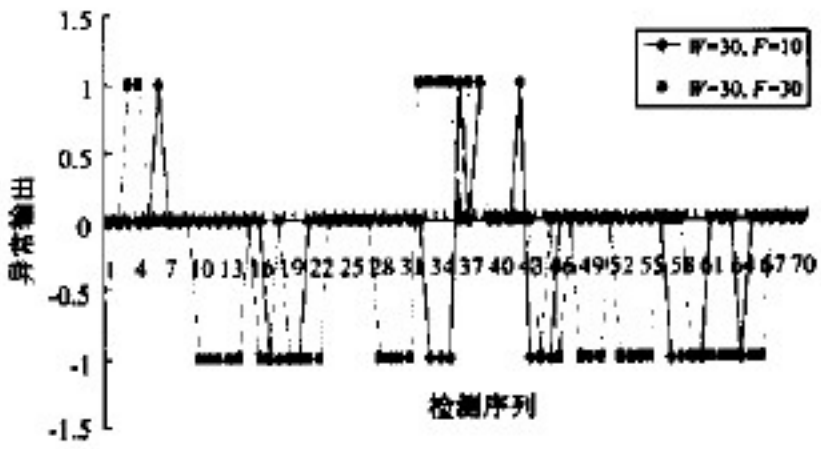

图 5 抽样频率对检测结果的影响

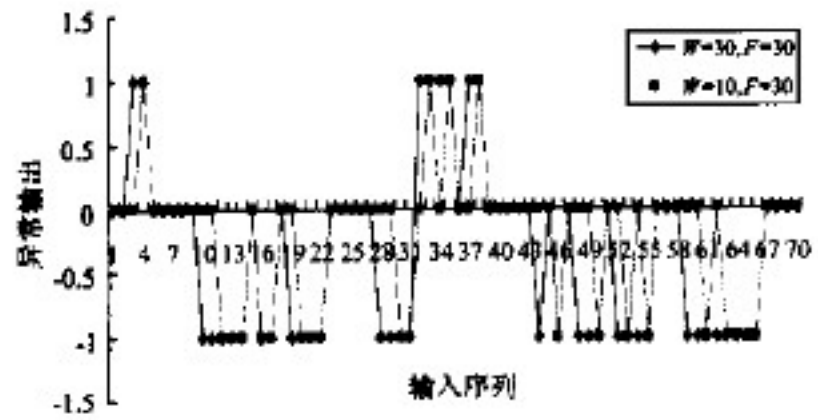

图 6 抽样窗口对检测结果的影响

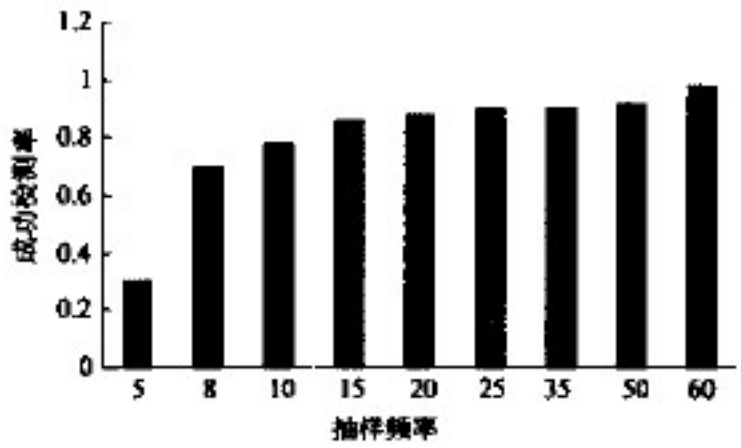

图 7 抽样频率对检测结果的影响

过对系统进行性能分析，我们发现该系统对资源的占用较少，为实际应用提供了较好的基础。

5 结 论

本文在提出一种原子行为间关联关系分析方法的基础上，设计了一种基于行为关联的主机入侵检测系统。在该系统中，我们仅采用了原子行为的期望概率分布作为判断特征，大大减少了正常特征存储的开销；并分析了采样策略对检测效果的影响。通过采用实验的方法，我们发现该系统的检测效率并没有因为采用的特征指标的简化而明显降低，但通过适当的调整采样的长度和频率基本可以实现 85% 以上的检测成功率，从而说明了关联关系分析算法的正确性和系统的可行性。

随着网络信息系统的普及，未来的入侵行为越来越趋于分布式，该趋势为本文提出的基于主机的入侵检测系统提出了挑战。为了使该系统能在保证分布式攻击的条件下保持较高的检测概率，基于分布式信息系统的行为序列关联分析方法和模型将是下一步研究的重点。

参考文献

[1] 田新广，高立志，孙春来等. 基于系统调用和齐次 Markov 链模型的程序行为异常检测. 计算机研究与发展，2007，44(9)：1538-1544

[2] Kim H S, Cha S D. Empirical evaluation of SVM-based masquerade detection using UNIX commands. *Computers and Security*, 2005, 24(2) : 160-168

[3] Forrest S, Hofmeyr S A, Somayaji A, et al. A sense of self for Unix processes. In: Proceedings of the 1996 IEEE Symposium on Security and Privacy, Oakland, USA, 1996. 120-128

[4] Warrender C, Forrest S, Pearlmutter B. Detecting intrusion using system calls: Alternative data models. In: Pro-

ceedings of the 1999 IEEE Symposium on Security and Privacy, Oakland, USA, 1999. 133-145

[5] 谭小彬,王卫平,奚宏生等. 计算机系统入侵检测的隐马尔可夫模型. 计算机研究与发展,2003,40(2):245-250

[6] 尹清波、张汝波,李雪耀等. 基于动态马尔科夫模型的入侵检测技术研究. 电子学报, 2004, 32(11): 1785-1788

[7] 尹清波,张汝波,李雪耀等. 基于线性预测与马尔可夫模型的入侵检测技术研究. 计算机学报, 2005, 28(5):900-907

[8] Lane T, Brodley C E. Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security*, 1999, 2(3): 295-331

[9] Lee W, Stolfo S J. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symopsium, SanAntonio, USA, 1998. 26-29

[10] Schonlau M, Mouchel W. Computer intrusion: Detecting masquerades. *Statistical Science*, 2001, 16(1): 58-74

[11] Szymanski B K, Zhang Y Q. Recursive data mining for masquerade detection and author identification. In: Proceedings of the 5th IEEE System, Man and Cybernetics Information Assurance Workshop, West Point, USA, 2004. 424-431

[12] Tian X G, Duan M Y, Li W F, et al. Anomaly detection of user behavior based on shell commands and homogeneous Markov chains. *Chinese Journal of Electronics*, 2008, 17(2): 231-236

[13] 田新广,段淑毅,程学旗. 基于 shell 命令和多重行为模式挖掘的用户伪装攻击检测. 计算机学报, 2010, 33(4): 697-705

[14] 林元烈. 应用随机过程. 北京:清华大学出版社. 2008. 244-246

An intrusion detection method for host systems based on behavior correlation

Wang Yinglong*, Li Jingchun**, Wang Shaojie**, Suo Yanfeng***, LiangLi**, Guo Ruilong*

(*School of Software College, University of Agricultural University Jiangxi, Nanchang 330045)

(**National Research Center for Information Technology Security, Beijing 100084)

(***School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

Abstract

A method for detecting intrusions through analyses of host behavior and behavior correlation is proposed. The method can efficiently find out the malicious software that is embedded with anomaly codes, and can be applied to behavior-based intrusion detection systems (IDS). By mining the characters of normal and anomaly behaviors of hosts, a way to build the Markov model of relationship of meta-behaviors and a method to detect intrusions are given. With them, the feasibility and scalability of the proposed method can be enhanced, and the store space can be reduced. The experimental results show that the loss-detection ratio, the error-detection ratio and the renew efficiency of the method are better than the existing methods, although it need more time to train datasets.

Key words: correlation analysis, host system, intrusion detection systems (IDS)