基于 Shell 命令和 DTMC 模型的用户行为异常检测新方法

肖 喜[1,2] 翟起滨[1] 田新广[3] 陈小娟[4]
(中国科学院研究生院信息安全国家重点实验室 北京 100049)[1]
(清华大学深圳研究生院 深圳 518055)[2]
(中国科学院计算技术研究所网络科学与技术重点实验室 北京 100190)[3]
(北京工商大学计算机与信息工程学院 北京 100037)[4]

摘 要 提出一种新的基于离散时间 Markov 链模型的用户行为异常检测方法,主要用于以 shell 命令为审计数据的入侵检测系统。该方法在训练阶段充分考虑了用户行为复杂多变的特点和审计数据的短时相关性,将 shell 命令序列作为基本数据处理单元,依据其出现频率利用阶梯式的数据归并方法来确定 Markov 链的状态,同现有方法相比提高了用户行为轮廓描述的准确性和对用户行为变化的适应性,并且大幅度减少了状态个数,节约了存储成本。在检测阶段,针对检测实时性和准确度需求,通过计算状态序列的出现概率分析用户行为异常程度,并提供了基于固定窗长度和可变窗长度的两种均值滤噪处理及行为判决方案。实验表明,该方法具有很高的检测性能,其可操作性也优于同类方法。

关键词 网络安全,入侵检测,shell 命令,异常检测,离散时间 Markov 链
中图法分类号 TP393.08 **文献标识码** A

Novel Method for Anomaly Detection of User Behavior Based on Shell Commands and DTMC Models

XIAO Xi[1,2] ZHAI Qi-bin[1] TIAN Xin-guang[3] CHEN Xiao-juan[4]
(State Key Laboratory of Information Security, Graduate University of Chinese Academy of Sciences, Beijing 100049, China)[1]
(Graduate School at Senzhen, Tsinghua University, Shenzhen 518055, China)[2]
(Key Laboratory of Network Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[3]
(College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100037, China)[4]

Abstract This paper presented a novel method for anomaly detection of user behavior based on the discrete-time Markov chain model, which is applicable to intrusion detection systems using shell commands as audit data. In the training period, the uncertainty of the user's behavior and the relevance of the operation of shell commands in short time were fully considered. This method takes the sequences of shell commands as the basic processing units. It merges the sequences into sets in terms of their ordered frequencies and then constructs states of the Markov chain on the merged results. Therefore this method increases the accuracy of describing the normal behavior profile and the adaptability to the variations of the user's behavior and sharply reduces the number of states and the required storage space. In the detection stage, considering the real-time performance and the accuracy requirement of the detection system, it analyzes the anomaly degree of the user's behavior by computing the occurrence probabilities of the state sequences, and then provides two schemes, based on the probability stream filtered with single window or multi-windows, to classify the user's behavior. The results of our experiments show that this method can achieve higher detection performance and practicability than others.

Keywords Network security, Intrusion detection, Shell command, Anomaly detection, Discrete-time Markov chain

1 引言

入侵检测系统(Intrusion Detection System, IDS)根据检测方法可划分为异常检测和误用检测。

异常检测利用被监控系统正常行为的信息作为检测系统中入侵、异常活动的依据;误用检测是根据已知入侵攻击的信

到稿日期:2010-12-17 返修日期:2011-03-30 本文受国家"863"高技术研究发展计划基金项目(2006AA01Z452),国家 242 信息安全计划基金项目(2005C39)资助。

肖 喜(1979—),博士生,主要研究方向为入侵检测、信息安全和密码应用技术,E-mail:xiaoxi_ac@163.com;翟起滨(1947—),教授,博士生导师,主要研究方向为密码学、信息安全和入侵检测;田新广(1976—),博士后,CCF 高级会员,主要研究方向为网络安全、入侵检测、智能信息处理;陈小娟(1977—),实验师,主要研究方向为通信工程、数字信号处理。

息(知识、模式等)来检测系统中的入侵和攻击[1]。异常检测的主要优点是不需要过多专业知识,能够检测出未知的攻击类型,有较强的适应性。近年来,用户行为异常检测作为入侵检测的一个重要分支得到了广泛研究,并在网络安全工程中发挥着越来越大的作用[2-7]。

用户行为异常检测面临的主要困难是用户行为具有多变性和复杂性,即用户行为会随着工作内容、用户兴趣、工作时间和其它不确定性因素的变化而变化[2,8,9]。基于 shell 命令的用户行为异常检测在最近 10 年受到了较多研究者的关注。T. Lane 等人[8,10]开展了基于实例学习和隐 Markov 模型(Hidden Markov Model,HMM)的两种用户行为异常检测方法的研究。基于实例学习的方法用特定的相似度函数刻画当前行为与正常行为模式之间的相似性,原理较为简单,有较强的适应能力,但没有考虑行为模式在训练数据中的出现频率和不同行为模式之间的相关性,因此检测准确率较低。基于 HMM 的方法虽然准确率高,但需要的计算量比较大,检测效率较低。孙宏伟等人[11]在基于实例学习方法的基础上改进了用户行为模式的表示方式,以 shell 命令为单位进行相似度赋值,改善了检测性能。M. Schonlau 等人[12]研究了基于统计理论的异常检测方法,综合分析了 6 种不同方法的优势和局限性。R. A. Maxion 等人[9]对 M. Schonlau 的方法进行了改进,引入了贝叶斯分类算法,提高了检测准确率。最近,S. K. Dash 等人[2]提出延迟检测概念(deferred detection concept),运用适应性朴素贝叶斯方法进行用户行为异常检测,使检测性能得到进一步提高。X. G. Tian 等人[7]提出了基于 Markov 链模型的检测方法,该方法有良好的检测性能,但是把不同的 shell 命令符号当作不同的状态,存在的问题有状态数目过多、计算复杂度大、容错能力和泛化能力不强等。

此外,B. K. Szymanski 等人[13]和 H. C. Wu 等人[3]研究了数据挖掘的方法,H. S. Kim 等人[14]提出了支持向量机的方法,这些方法具有较高的检测效率,但仅适用于训练数据较为充分的场合。Coull 等人[5]把生物信息学里的序列比对(sequence alignment)算法应用于用户行为异常检测。田新广等人[6]采用特殊的 HMM 建立合法用户的正常行为轮廓,采用运算量较小的序列匹配方法,训练时间大幅度降低,检测效率得到了提高。K. Wang 等人[15]提出两种单分类训练方法(one-class training),它们均达到了与多分类方法相同的检测效果。

2 问题描述

在一个实际的计算机网络系统中,一般会有多个合法用户,这些合法用户通常具有不同的操作权限,而且不同的合法用户具有不同的行为特点。在很多情况下,我们需要监视系统中一些合法用户的行为,检测其行为中的异常,以防止其他用户(包括非法用户)冒用这些合法用户的账号进行非法操作,或者防止这些合法用户进行非授权操作。用户行为异常检测方法在用户界面层中建立一个(或一组)合法用户的正常行为轮廓,通过比较该合法用户的当前行为和此正常行为轮廓来识别异常行为:如果该合法用户的当前行为较大程度地偏离了其历史上的正常行为轮廓,则认为发生了异常。这种异常可能是该合法用户本身进行了非授权操作,也可能是系统中其他合法用户或外部入侵者(非法用户)冒充该合法用户进行了非法操作[6]。在 UNIX 或 LINUX 平台上,shell 是终端用户与操作系统之间最主要的界面,很大比例的用户活动都是利用 shell 完成的;shell 命令在系统用户层比较容易获取,而且能够直接反映出用户的行为模式,所以现有的用户行为异常检测研究大都采用 UNIX 平台上的 shell 命令作为审计数据[4]。

文献[7]基于 Markov 链模型的方法以单个 shell 命令符号为基本数据处理单元,仅考虑了单个 shell 命令符号对应状态的相互转移关系。事实上,单个 shell 命令符号难以直接反映一个用户的实际行为模式,用户往往需要用长度大于 1 的 shell 命令序列才能完成一个具有明确意义的操作,因而用户的审计数据(shell 命令)具有短时相关性。自然的想法是把文献[7]中不同的单个 shell 命令符号对应不同状态推广为不同的 shell 命令序列(长度>1)对应不同状态,然而这样会使原来的问题更加严重:不同的 shell 命令序列对应的不同状态个数会随序列长度呈指数级增加,相应的计算复杂度也呈指数级增加等。本文在以 shell 命令序列为基本数据处理单元的基础上,根据其出现频率进行阶梯式数据归并来确定状态,以克服上述问题。

本文方法在训练阶段充分考虑了用户行为复杂多变的特点和审计数据的短时相关性,改进了对用户正常行为模式的表示方式,采用平稳的、时间齐次的离散时间 Markov 链(DTMC)模型对合法用户的正常行为进行建模,以不同的 shell 命令序列为基本数据处理单元,依据其出现频率通过阶梯式的数据归并来确定 Markov 链的状态,同现有的 Markov 链方法[7]相比,其提高了用户行为轮廓描述的准确性和对用户行为变化的适应性,并大幅度减少了状态个数,节约了存储成本。在检测阶段,考虑到实时性需求和准确度需求,通过对状态序列出现的概率进行加窗均值滤噪处理来计算判决值,并提供了两种不同的判决方案:单个固定长度窗口的方案和多个可变长度窗口的方案,降低了系统计算开销,提高了准确度。本文将该方法与现有的 4 种典型方法进行了性能比较。利用 UNIX 平台上的用户 shell 命令数据的实验表明,本文方法的存储成本小于文献[7]方法,计算成本低于文献[6,10,11]方法,检测准确度高于文献[6,7,10,11]方法,整体性能得到了提高。该方法具有很高的检测准确率,其可操作性也优于同类方法。

3 DTMC 的定义和定理

定义 1 一个随机过程$\{X_n, n\geqslant 1\}$称作具有状态空间$S=\{1,2,\cdots\}$的(一阶)离散时间 Markov 链(Discrete-Time Markov Chain,DTMC),如果对所有的$n\geqslant 1$,$j\in S$和$s_m\in S$$(1\leqslant m\leqslant n)$,有

$$\Pr(X_{n+1}=j\mid X_n=s_n, X_{n-1}=s_{n-1},\cdots,X_1=s_1)=\Pr(X_{n+1}=j\mid X_n=s_n) \quad (1)$$

定义 2 一个 DTMC$\{X_n, n\geqslant 1\}$称作时间齐次的,如果它满足不动性假设,即对所有的$j\in S$和$i\in S$,条件概率$\Pr(X_{n+1}=j\mid X_n=i)$与n无关。

当 DTMC$\{X_n, n\geqslant 1\}$是时间齐次时,$p_{ij}=\Pr(X_{n+1}=j\mid X_n=i)$称作(一步)转移概率,由(一步)转移概率$p_{ij}$构成的矩阵$P=[p_{ij}]$称作(一步)转移概率矩阵。$a_i=\Pr(X_1=i)$称作状态$i$的初始出现概率,由状态的初始出现概率构成的行向

量 $A=(a_i)_{i\in S}$ 称作初始概率分布。

定理 1 时间齐次的 DTMC$\{X_n, n\geqslant 1\}$ 由初始概率分布 A 和转移概率矩阵 P 完全刻画，即

$$\Pr(X_1=s_1,\cdots,X_{n-1}=s_{n-1},X_n=s_n)=a_{s_1}p_{s_1,s_2}\cdots p_{s_{n-1},s_n} \tag{2}$$

定义 3 一个 DTMC$\{X_n, n\geqslant 1\}$ 称作平稳的，如果其初始概率分布 A 和转移概率矩阵 P 满足

$$A=A\times P \tag{3}$$

定理 2 对平稳的 DTMC$\{X_n, n\geqslant 1\}$，下式成立：

$$\Pr(X_n=i)=\Pr(X_1=i)=a_i \tag{4}$$

由定理 1 和定理 2 可得：

定理 3 对平稳的、时间齐次的 DTMC$\{X_n, n\geqslant 1\}$，下式成立：

$$\Pr(X_i=s_i,\cdots,X_{j-1}=s_{j-1},X_j=s_j)=a_{s_i}p_{s_i,s_{i+1}}\cdots p_{s_{j-1},s_j},1\leqslant i<j \tag{5}$$

以上定理的详细证明可参见文献[16]。

4 训练

本文方法利用平稳的时间齐次的 DTMC 模型对合法用户的正常行为进行建模：预先设定此 DTMC 的状态个数为 N，首先确定其状态，然后计算其初始概率分布 $A=(a_1,a_2,\cdots,a_N)$ 和转移概率矩阵 $P=[p_{ij}]_{N\times N}$。

4.1 DTMC 状态的确定

本文方法需要预先设定 DTMC 的状态空间 S 为$\{1,2,\cdots,N-1,N\}$，确定状态的具体步骤如下：

(1)获得合法用户正常行为的原始训练数据，生成 shell 命令序列串。首先将该用户正常行为的原始数据预处理成按时序排列的字符串 $o=(o_1,o_2,\cdots,o_t)$，再由 o 以 δ 为窗长生成 shell 命令序列串 $XLC(o,\delta)=(XL(o_1,\delta),XL(o_2,\delta),\cdots,XL(o_{l-\delta+1},\delta))$。其中，由 o 以 δ 为窗长生成序列串的操作是指先在 o 上以 δ 为窗长截取出 $l-\delta+1$ 个长度为 δ 的序列 $XL(o_i,\delta)=(o_i,o_{i+1},\cdots,o_{i+\delta-1})$，$i=1,2,\cdots,l-\delta+1$；然后将这些序列按其第 1 个字符的时间先后顺序排列，构成新的序列 $XLC(o,\delta)=(XL(o_1,\delta),XL(o_2,\delta),\cdots,XL(o_{l-\delta+1},\delta))$。$XLC(o,\delta)$ 简称为 o 的序列串。

以下以 shell 命令序列为基本数据处理单元，提高了用户行为轮廓描述的准确性和对用户行为变化的适应性。

(2)提取出 shell 命令序列串 $XLC(o,\delta)$ 中互不相同的 shell 命令序列(其个数记为 H)，并计算其在 $XLC(o,\delta)$ 中的出现频率。每个 shell 命令序列的出现频率＝其出现次数/$(l-\delta+1)$。

(3)将互不相同的 shell 命令序列按其在序列串 $XLC(o,\delta)$ 中的出现频率从大到小排序，然后根据出现频率进行阶梯式数据归并。设 b 是不大于 $H/(N-1)$ 的最大整数，$c=b+1$，$h=H-(N-1)b$，则 $H=hc+(N-1-h)b$。将互不相同的 shell 命令序列按其出现频率从大到小排序，排序后的前 hc 个序列按 c 个一组归并成 1 个集合，生成集合 $\Omega_1,\cdots,\Omega_h$；剩下的序列按 b 个一组归并成 1 个集合，生成集合 $\Omega_{h+1},\cdots,\Omega_{N-1}$。

(4)将一个集合对应一个状态，并引入一个附加状态 N，对应没出现过的 shell 命令序列。这样状态对应的 shell 命令序列范围扩大，提高了系统的容错能力和泛化能力，并减少了存储成本。在实际操作中，可运用频率优先匹配方法按照频率从高到低的顺序依次与样本命令序列进行比较来确定状态，以节省 shell 命令序列的匹配时间。频率优先匹配方法确定 t 时刻出现的 shell 命令序列 $XL(o_t,\delta)$ 的状态 s_t 的伪代码(本文采用 C 语言伪代码)为：

```
for(j=1;j<=N-1;j++)
    if(XL(o_t,δ)∈Ω_j)
      { s_t=j;break;}
  if(j==N)
  s_t=N
```

4.2 DTMC 参数的计算

DTMC 参数的计算包括其初始概率分布 $A=(a_1,a_2,\cdots,a_N)$ 和转移概率矩阵 $P=[p_{ij}]_{N\times N}$ 的计算，具体过程的伪代码为(shell 命令序列串 $XLC(o,\delta)=(XL(o_1,\delta),XL(o_2,\delta),\cdots,XL(o_{l-\delta+1},\delta))$ 作为输入)：

```
for(i=1;i<=N;i++)
    { a_i=0;
      for(j=1;j<=N;j++)
          p_ij=0;}
p_NN=1;
for(t=1;t<=l-δ+1;t++)
{ 根据 4.1 节(4)确定 t 时刻出现的 shell 命令序列 XL(o_t,δ)的状态
  s_t;
  j=s_t;a_j=a_j+1;
  if(t>=2)
    { i=s_{t-1};p_ij=p_ij+1;} }
for(i=1;i<=N;i++)
    { sum_i=p_i1+p_i2+…+p_iN;
      a_i=a_i/(l-δ+1);
      for(j=1;j<=N;j++)
          p_ij=p_ij/sum_i;}
```

此 DTMC 是在不动性假设的前提下建立的，所以是时间齐次的。

5 检测

在检测阶段，针对检测实时性和准确度需求，通过计算状态序列的出现概率分析用户行为异常程度，并采用基于固定窗长度和可变窗长度的两种均值滤噪处理方法进行用户行为判决。具体步骤如下：

(1)获得被监测用户在被监测的时间内执行的 shell 命令原始数据，生成状态序列串。首先把这些原始数据预处理成长度为 r 的按时序排列的 shell 命令字符串，再以 δ 为窗长生成对应的 shell 命令序列串，然后根据 4.1 节(4)把这个 shell 命令序列串转化为状态字符串 $s=(s_1,s_2,\cdots,s_k)$，其中 $s_m(1\leqslant m\leqslant k)$ 是其 shell 命令序列串中的第 m 个 shell 命令序列对应的状态，$k=r-\delta+1$。最后由状态字符串 s 以 η 为窗长生成状态序列串 $XLC(s,\eta)=(XL(s_1,\eta),XL(s_2,\eta),\cdots,XL(s_{k-\eta+1},\eta))$。

(2)计算该用户的状态序列串 $XLC(s,\eta)$ 中每个状态序列 $XL(s_i,\eta)$ 出现的概率。记该用户的第 i 个状态序列 $XL(s_i,\eta)$ 出现的概率为 $\Pr(XL(s_i,\eta))$，由定理 3 可得 $\Pr(XL(s_i,\eta))$ 的计算方法：

$$\Pr(XL(s_i,\eta))=\Pr(X_i=s_i,X_{i+1}=s_{i+1},\cdots,X_{i+\eta-1}=s_{i+\eta-1})$$

$$=a_{s_i}\prod_{j=i}^{i+\eta-2}p_{s_j,s_{j+1}} \tag{6}$$

(3)在概率序列($\Pr(XL(s_1,\eta)),Pr(XL(s_2,\eta)),\cdots,\Pr(XL(s_{k-\eta+1},\eta))$)中通过加窗均值滤噪处理来定义判决值函数,对该用户的"当前行为"进行判决。具体判决方案有两种:1)单个固定长度窗口的方案;2)多个可变长度窗口的方案。

方案 1 设固定窗口长度为 w,定义两个不同的判决值函数:

① 直接定义判决值函数

$$D(n)=\frac{1}{w}\sum_{i=n-w+1}^{n}\mathrm{sgn}[\Pr(XL(s_i,\eta))-g] \tag{7}$$

式中,g 为概率门限,需预先设定。

② 先对 $\Pr(XL(s_i,\eta))$ 进行截幅处理:预先设定 $f>0$ 为截断常数,当 $\Pr(XL(s_i,\eta))>f$ 时,$\Pr(XL(s_i,\eta))$ 不变;当 $\Pr(XL(s_i,\eta))\leqslant f$ 时,$\Pr(XL(s_i,\eta))$ 统一取值为 f(目的是减少虚警概率)。然后在截幅处理后的 $\Pr(XL(s_i,\eta))$ 上定义判决值函数:

$$D(n)=\frac{1}{w}\sum_{i=n-w+1}^{n}\lg[\Pr(XL(s_i,\eta))] \tag{8}$$

判决时,设定一个判决门限 d。如果 $D(n)$ 大于判决门限 d,将该用户的"当前行为"判为正常,否则判为异常。

方案 2 设定 e 个可变长度窗口的窗长分别为 $w(1),w(2),\cdots,w(e)$,且 $w(1)<w(2)<\cdots<w(e)$;再设定 e 个判决上限分别为 $u(1),u(2),\cdots,u(e)$ 和 e 个判决下限分别为 $v(1),v(2),\cdots,v(e)$,其中 $u(m)$ 和 $v(m)$ 是第 m 个长度为 $w(m)$ 的窗口对应的判决上限和判决下限($1\leqslant m\leqslant e$),且 $u(1)>u(2)>\cdots>u(e-1)>u(e)=v(e)>v(e-1)>\cdots>v(2)>v(1)$。

定义与式(7)对应的可变窗长判决值函数:

$$D(n,m)=\frac{1}{w(m)}\sum_{i=n-w(m)+1}^{n}\mathrm{sgn}[\Pr(XL(s_i,\eta))-g] \tag{9}$$

或者定义与式(8)对应的可变窗长判决值函数:

$$D(n,m)=\frac{1}{w(m)}\sum_{i=n-w(m)+1}^{n}\lg[\Pr(XL(s_i,\eta))] \tag{10}$$

式(9)的参数说明及式(10)截幅处理与式(7)、式(8)一样。

对该用户的"当前行为"作出判决的伪代码为:

```
for(m=1;m<=e;m++)
    {
if(n<w(m))
  { 不对"当前行为"进行判决;break;}
if(D(n,m)>u(m))
  { 把"当前行为"判为正常;break;}
if(D(n,m)<=v(m))
  { 把"当前行为"判为异常;break;}
}
```

在以上方法中,当 $n<w(1)$ 时,不对该用户的"当前行为"进行判决;当 $w(1)\leqslant n<w(e)$ 时,不一定能够对该用户的"当前行为"作出判决;当 $w(e)\leqslant n\leqslant k-\eta+1$ 时,总可以对该用户的"当前行为"作出判决(因为 $u(e)=v(e)$)。实际中,最长的窗长度 $w(e)$ 远小于 $k-\eta+1$,因此其它情况可忽略不计。

方案 1 采用单个固定长度的窗口计算判决值,对系统资源的占用较少。固定的窗长度 w 是一个重要的参数,w 越小,检测的实时性就越强。但减小窗长度 w,检测准确度会降低。方案 2 采用多个可变长度的窗口并联合使用多个门限对被监测用户的行为进行判决,兼顾了检测时间和检测准确度,可提高检测的准确度,与方案一相比,其能够在保证同等检测准确度的前提下提高检测的实时性。

6 实验设计与分析

国际上常用的 shell 命令实验数据主要有两组:1)Purdue 大学实验数据[8,10];2)SEA 实验数据——AT&T Shanon 实验室数据[12]。作者分别在上述两组实验数据上进行了实验,检测了本文方法的性能。

6.1 在 Purdue 大学实验数据上的实验

本组实验采用 Purdue 大学实验数据中 4 个用户 user1、user2、user3、user4 的数据。将 user1 设为合法用户,user2、user3、user4 设为非法用户。原始数据的预处理方式如下(详见文献[7,8,10]):滤除 shell 命令中的主机名、网址等信息,保留 shell 命令的名称及参数;各命令符号按照在 shell 会话中的出现次序进行排列,不同的 shell 会话按照时间顺序进行连接,每个会话开始和结束的时间点上插入标识符号。预处理后,每个用户的数据中有 15000 个 shell 命令。其中 user1 的前 10000 个命令作为正常行为训练数据,用于确定 DTMC 状态和计算参数;后 5000 个命令作为正常行为测试数据,用于测试虚警概率(False positive rate);其他 3 个用户数据的后 5000 个 shell 命令均作为异常行为测试数据,用于测试检测概率(True positive rate)。实验中的检测阶段分别采用式(7)和式(8)计算判决值,对用户行为进行判决,参数设置为 $N=3,\delta=4,\eta=2,w=301,g=5\times10^{-4},f=10^{-20}$。

实验时,正常行为训练数据中互不相同的长度为 4 的 shell 命令序列共有 2975 个,DTMC 的状态个数 N 为 3,初始概率分布 $A=(0.85134,0.14866,0)$,转移概率矩阵:

$$P=\begin{pmatrix}0.92338 & 0.07662 & 0\\ 0.43876 & 0.56124 & 0\\ 0 & 0 & 1\end{pmatrix}$$

简单计算容易验证其满足平稳 DTMC 的定义。

图 1 示出了由式(8)计算出的判决值曲线。图中上方的实线是合法用户 user1 的测试数据对应的判决值曲线,下方的 3 条虚线分别是非法用户 user2,user3,user4 的测试数据对应的判决值曲线。可见,判决值曲线具有良好的可分性。

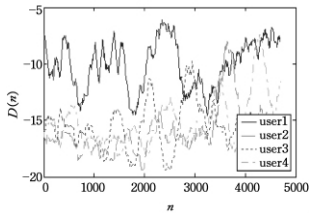
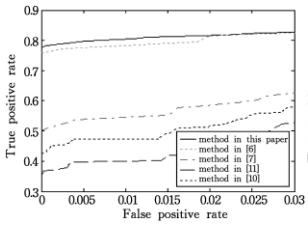
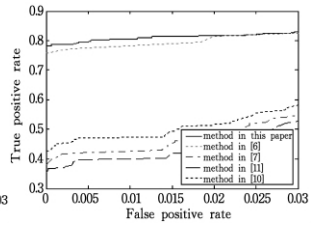


图 1 式(8)对应的判决值曲线

图 2 和图 3 分别示出了由式(7)和式(8)作为判决值的 ROC 曲线和文献[6,7,10,11]中 4 种方法的 ROC 曲线。图中不同方法的参数设置保证了平均检测时间基本相同[17]。从图 2、图 3 可看出,式(7)、式(8)计算的判决值均获得了较高的检测准确率;而且,两者对应的检测准确率比较接近,这说明基于状态序列出现概率的判决值计算方法是一种性能稳健的方法。由图可见,本文方法比文献[6,7,10,11]中 4 种方法的检测准确率均有明显的提高。



本文方法 $N=3$，式(7)作为判决值



本文方法 $N=3$，式(8)作为判决值

图 2　5 种不同方法的 ROC 曲线　图 3　5 种不同方法的 ROC 曲线

6.1.1　序列长度 δ 对检测性能的影响

图 4 示出了本文方法 3 个状态由式(8)作为判决值序列长度 δ 不同时的 ROC 曲线。由图可见，当序列长度 δ 取 1～4 时，检测准确率随序列长度的增加而提高；当序列长度 δ 取 4～6 时，检测准确率随序列长度的增加而降低；序列长度 $\delta=4$ 的检测准确率最高。考虑到存储量和计算量，我们在试验中选择序列长度 $\delta=4$。

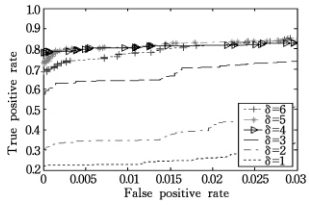
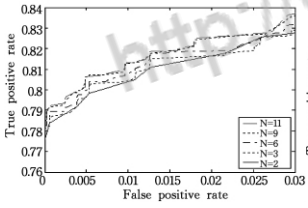
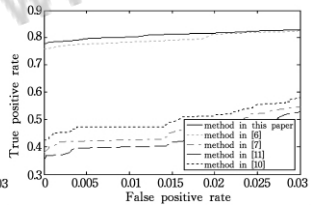


图 4　式(8)作为判决值时序列长度 δ 不同的 ROC 曲线

6.1.2　状态个数 N 对检测性能的影响

图 5 示出了本文方法 $\delta=4$ 由式(8)作为判决值时不同状态个数 N 下的 ROC 曲线。从图 5 可看出 ROC 曲线的准确率随状态个数的增加而提高，$N=11$ 的 ROC 曲线稍好，但与 $N=3$ 差别不大。考虑到存储量和计算量，我们在试验中选择状态个数 $N=3$。





本文方法 $N=2$，式(8)作为判决值

图 5　式(8)作为判决值时状态个数 N 不同的 ROC 曲线　图 6　5 种不同方法的 ROC 曲线

实际上，利用本文方法状态个数 $N=2$ 时也比其它 4 种方法好，图 6 示出了本文方法 $N=2$ 时以式(8)作为判决值的 ROC 曲线和其它 4 种方法的 ROC 曲线。图中本文 DTMC 方法状态个数 $N=2$，其它与图 2、图 3 一样。

6.1.3　与现有典型方法的性能比较

文献[7]提出了基于 Markov 链模型的方法，文献[6]提出了基于 HMM 的方法，文献[10,11]分别提出了两种不同的机器学习方法。下面是本文方法和以上 4 种方法在性能上的比较。

1)检测准确率：由图 3 和图 4 可以看出，本文方法的检测准确率优于文献[6]基于 HMM 的方法，在虚警概率相同的情况下检测概率比文献[7]基于 Markov 链模型方法和文献[10,11]机器学习方法提高了将近 1 倍。

2)存储空间：本文方法需要 11909 个存储单元，文献[7]基于 Markov 链模型的方法需要 33305 个存储单元。本文方法是文献[7]存储空间的 11909/33305＝35.76%，减少了 2/3。

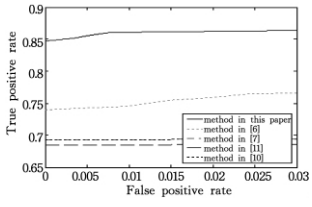
3)实验时间：实验时间是指实验中进行训练和检测所需要的时间，它与检测方法的计算成本成正比，并在一定程度上反映了检测的实时性。文献[10]、文献[11]、文献[6]方法和本文方法的试验时间分别为(单位：秒) 1777.1，29.563，30.656，19.797。本文方法实验时间比文献[6]基于 HMM 的方法和文献[10,11]机器学习方法短很多。本文方法实验时间仅是文献[10]T. Lane 的方法的 19.797/1777.1＝1.11%，减少了两个数量级。

本文方法的检测准确度高于文献[6,7,10,11]的方法，存储成本小于文献[7]方法，计算成本低于文献[6,10,11]方法。综合这 3 方面，本文方法的整体性能优于已有的 4 种方法。

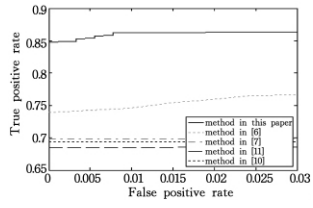
6.2　在 SEA 实验数据上的试验

本组实验采用 AT&T Shannon 实验室实验数据中前 4 个用户 user1、user2、user3、user4 的数据。将 user4 设为合法用户，user1、user2、user3 设为非法用户。原始数据的预处理方式较为简洁(详见文献[9,12])：只保留 shell 命令的名称，滤除命令参数和时间等信息。预处理后每个用户有 5000 个 shell 命令。user4 的前 4000 个命令作为训练数据，用于正常行为建模；后 1000 个命令作为测试数据，用于测试虚警概率；其他 3 个用户的 5000 个 shell 命令均作为测试数据用于测试检测概率。实验中的检测阶段分别采用式(7)和式(8)计算判决值对用户行为进行判决，参数设置为 $N=3$，$\delta=4$，$\eta=2$，$w=100$，$g=5\times10^{-4}$，$f=10^{-20}$。

图 7 和图 8 分别示出了由式(7)和式(8)作为判决值的 ROC 曲线和其它 4 种方法的 ROC 曲线。可见，本文方法比文献[6,7,10,11]中的 4 种方法的检测准确率均有明显的提高。



本文方法 $N=3$，式(7)作为判决值。



本文方法 $N=3$，式(8)作为判决值。

图 7　在 AT&T Shannon 实验室数据上 5 种不同方法的 ROC 曲线　图 8　在 AT&T Shannon 实验室数据上 5 种不同方法的 ROC 曲线

结束语　本文提出一种基于 shell 命令和 DTMC 模型的用户行为异常检测新方法，该方法提高了检测系统的综合性能，具有很高的检测准确率和较强的可操作性。在实际应用中，根据具体需求和具体用户的行为特点选择合适的方案和参数还可以得到更好的检测性能。本文的时间齐次 DTMC 模型只是一阶的，多阶的时间齐次 DTMC 模型在用户行为异常检测中的应用有待研究。

参 考 文 献

[1]　胡艳维，秦拯，张忠志. 基于模拟退火与 K 均值聚类的入侵检测算法[J]. 计算机科学，2010，37(6)：122-124

[2]　Dash S K，Reddy K S，Pujari A K. Adaptive Naive Bayes Method for Masquerade Detection[J]. Security and Communication Networks，2010：DOI：10.1002/sec.168

(下转第 82 页)

但其识别率接近于原始的 20 个特征的识别率。原始 20 个特征经过 KPCA 降维之后，得到了 17 个特征，虽然识别率相对于原始 20 个特征的识别率有所提高，但提高的程度不大，说明这 20 个原始特征的非线性不是很强。正交试验方法的识别率均高于 PCA 方法和 KPCA 方法的识别率。由图 2 可知，正交试验在多径环境下仍能保持较高的识别率。

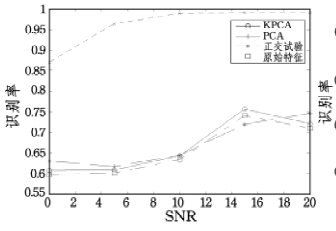


图 1 高斯信道下识别性能对比图

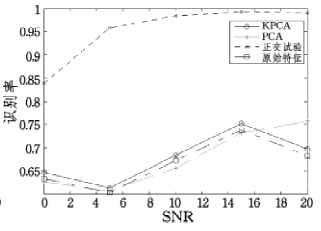


图 2 多径信道下识别性能对比图

这 3 种优化特征参数方法的共同点在于减少无用特征，保留有用特征，用较少的特征达到更高的识别率。3 种方法的不同点在于：PCA 和 KPCA 方法都是利用方差贡献率来选取较优特征，PCA 方法的较优特征是由原始 20 个特征经过线性变换得到的新特征；而 KPCA 方法的较优特征是由原始的 20 个特征先经过非线性变换升维之后，再经过线性变换降维得到的新特征；正交试验法是根据构造的正交表中的特征组合进行试验分析得到较优特征，较优特征只是从原始特征中选取出来的特征，不经过任何变换，不是新特征。选用特征参数的相关性太强，非线性不强，导致 PCA 以及 KPCA 方法的优点没有凸显出来，尽管达到了一定的优化特征参数，而相对于正交试验方法识别率明显较低；在正交试验方法的分析中，没有考虑各个特征参数的相关性，即其组合特征的识别率，故其分析之后的较优特征有可能不是最优特征组合。

结束语 本文用正交实验方法、PCA 方法和 KPCA 方法对 9 种数字调制信号的 20 个特征参数进行了优化选择，然后分别在高斯和多径信道下利用优化特征对数字调制信号进行了识别。结果表明，基于正交实验的数字调制信号特征参数优化的效果均优于 PCA 和 KPAC 方法。

参 考 文 献

[1] 吕铁军，王河，肖先赐. 新特征选择方法下的信号调制识别[J]. 电子与信息学报，2002，24(5)：661-666

[2] 韩志艳，王健，王旭. 基于正交实验设计的语音识别特征参数优化[J]. 计算机科学，2010，37(1)：214-216

[3] 杨大利，徐明星，吴文虎. 语音识别中一种新的特征参数选择方法[J]. 清华大学学报：自然科学版，2003，43(1)：79-82

[4] 杨大利，徐明星，吴文虎. 语音识别特征参数选择方法研究[J]. 计算机研究与发展，2003，40(7)：963-969

[5] 彭策，熊屹，陈文西. 病态噪声识别特征参数的优化选择[J]. 中国生物医学工程学报，2007，26(5)：675-679

[6] 李侃，孙进平，成功，等. 一种基于支持向量机的数字调制识别方法[J]. 电路与系统学报，2010，15(3)：7-12

[7] 张弛，吴瑛，周欣. 基于高阶累积量的数字调制信号识别[J]. 数字采集与处理，2010，25(5)：575-579

[8] 李楠，曲长文，平殿发，等. 基于分形理论的辐射源识别算法[J]. 航天电子对抗，2010，26(2)：62-64

[9] Zhang Jing-jing，Li Bing-bing. A new modulation identification scheme for OFDM in multipath rayleigh fading channel[C]// International Symposium on Computer Science and Computational Technology. Shanghai：China，IEEE，2008：793-796

[10] Subasi A，Gursoy M I. EEG signal classification using PCA，ICA，LDA and support vector machines[J]. Expert Systems with Applications，2010，37(12)：8659-8666

[11] 任若恩，王惠文. 多元统计数据分析[M]. 北京：国防工业出版社，1997：92-109

[12] Xu Y，Lin C，Zhao W. Producing computationally efficient KPCA-based feature extraction for classification problems[J]. Electronics Letters，2010，46(6)：452-453

(上接第 58 页)

[3] Wu H C，Huang S H S. Masquerade Detection Using Command Prediction and Association Rules Mining[C]// 2009 International Conference on Advanced Information Networking and Applications. Aina：IEEE，2009：552-559

[4] 田新广，段洣毅，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[J]. 计算机学报，2010，33(4)：697-705

[5] Coull S E，Branch J W，Szymanski B K，et al. Sequence Alignment for Masquerade Detection[J]. Computational Statistics & Data Analysis，2008，52(8)：4116-4131

[6] 田新广，段洣毅，孙春来，等. 采用 shell 命令和隐 markov 模型进行网络用户行为异常检测[J]. 应用科学学报，2008，26(02)：175-181

[7] Tian Xin-guang，Duan Mi-yi，Li Wen-fa，et al. Anomaly Detection of User Behavior Based on Shell Commands and Homogeneous Markov Chains[J]. Chinese Journal of Electronics，2008，17(2)：231-236

[8] Lane T. Machine Learning Techniques for the Computer Security Domain of Anomaly Detection[D]. West Lafayette，Indiana：Purdue University，2000

[9] Maxion R A，Townsend T N. Masquerade Detection Using Truncated Command Lines[C]// International Conference on Dependable Systems & Networks. Washington，DC，USA，2002：219-228

[10] Lane T，Brodley C E. An Empirical Study of Two Approaches to Sequence Learning for Anomaly Detection[J]. Machine learning，2003，51(1)：73-107

[11] 孙宏伟，田新广，李学春，等. 一种改进的 IDS 异常检测模型[J]. 计算机学报，2003，26(11)：1450-1455

[12] Schonlau M，DuMouchel W，Ju W H，et al. Computer Intrusion：Detecting Masquerades[J]. Statistical Science，2001，16(1)：58-74

[13] Szymanski B K，Zhang Y Q. Recursive Data Mining for Masquerade Detection and Author Identification[C]// Proceedings of the 5th IEEE System，Man and Cybernetics Information Assurance Workshop. West Point，NY，USA，2004：424-431

[14] Kim H S，Cha S D. Empirical Evaluation of SVM-based Masquerade Detection Using Unix Commands[J]. Computers & Security，2005，24(2)：160-168

[15] Wang K，Stolfo S J. One Class Training for Masquerade Detection[C]// ICDM Workshop on Data Mining for Computer Security(DMSEC 03). Citeseer，2003

[16] Karlin S，Taylor H M. A First Course in Stochastic Processes [M]. Second Edition. Beijing：Post&Telecom Press，2007

[17] 田新广. 基于主机的入侵检测方法研究[D]. 长沙：国防科学技术大学，2005

阅读此文的还阅读了：

1. 一种基于选择性协同学习的网络用户异常行为检测方法
2. VBS的妙用
3. 基于Shell命令和共生矩阵的用户行为异常检测方法
4. 基于频谱细化的列车轮对轴承故障在线检测
5. 一种新的基于Markov链模型的用户行为异常检测方法
6. 留意你的DNS:从一些DNS/AD故障案例中吸取教训
7. 计算机使用技巧:doskey宏的使用
8. 移动视频监控及异常场景检测实现
9. 点极性法在检测PT二次回路异常中的应用
10. 德国研究人员开发出眼控电脑软件系统
11. 宁镇地区铜矿床地质-地球物理-地球化学模型及找矿标志
12. 动词第二人称命令式的用法小结
13. 基于shen命令和Markov链模型的用户行为异常检测
14. Biliary reflux detection in anomalous union of the pancreatico- biliary duct patients
15. 基于支持向量机的Web用户行为异常检测方法
16. A data driven approach for detection and isolation of anomalies in a group of UAVs

17. 基于进程行为的异常检测模型

18. 基于用户击键数据的异常入侵检测模型

19. 兰成渝输油管道阴极保护系统异常段的检测与修复

20. 基于shell命令和多重行为模式挖掘的用户伪装攻击检测

21. 一种基于隐马尔可夫模型的IDS异常检测新方法

22. 基于隐马尔可夫模型的IDS程序行为异常检测

23. 基于模式挖掘的用户行为异常检测算法

24. 一种基于隐马尔可夫模型的IDS异常检测新方法

25. 基于异常的入侵检测方法分析

26. 弹性波检测技术在水利工程施工中的应用

27. 基于shell命令和Markov链模型的用户行为异常检测

28. 基于隐马尔科夫模型的用户行为异常检测方法

29. 基于行为模型的IP Forwarding异常检测方法

30. 改变自己和改变世界

31. 在Windows Vista中，如何能够找到某个用户具有访问权限的所有文件？

32. 奶牛的发情鉴定及异常发情表现

33. 甘三酯分析新方法:RP—HPLC/APCI—MS

34. 英国牛津大学检测异常系统行为项目（2013—105-英国-064）

35. 动物疫病检测新方法研究进展

36. 基于异常的入侵检测方法分析

37. 用户行为异常检测模型

38. 质监部门执法维权小集锦

39. 广西百色盆地油气综合化探方法研究及效果

40. 一种新的有效检测纸和纸板Z向强度的方法

41. 孩子喝饮料过多易患多动症（1）

42. 一种基于HMM异常检测的新方法

43. 基于命令的程控交换机仿真训练系统的设计与实现

44. 新形势下高校思想政治工作新方法模型研究

45. 摊铺机:专家之眼用户之见:采纳专家之高见——中国摊铺机发展现状

46. 儿童异常行为与心理治疗的探讨

47. 采用新技术、新方法检测连杆螺栓

48. 一个准确高效的基于程序行为的异常检测模型

49. Homeland Security Research:截至2016年,生物识别监控及行为异常探测市场总值将达到32亿美元

50. Vizrt：意即图形实时