基于 shell 命令和 Markov 链模型的用户伪装攻击检测

肖喜[1]，田新广[2]，翟起滨[1]，叶润国[3]

(1. 中国科学院 研究生院 信息安全国家重点实验室，北京 100049;
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. 北京启明星辰信息安全技术有限公司，北京 100193)

摘 要：提出一种新的基于 shell 命令的用户伪装攻击检测方法。该方法在训练阶段充分考虑了用户行为的多变性和伪装攻击的特点，采用平稳的齐次 Markov 链对合法用户的正常行为进行建模，根据 shell 命令的出现频率进行阶梯式数据归并来划分状态，同现有的 Markov 链方法相比大幅度减少了状态个数和转移概率矩阵的存储量，提高了泛化能力。针对检测实时性需求和 shell 命令操作的短时相关性，采用了基于频率优先的状态匹配方法，并通过对状态短序列的出现概率进行加窗平滑滤噪处理来计算判决值，能够有效减少系统计算开销，降低误报率。实验表明，该方法具有很高的检测准确率和较强的可操作性，特别适用于在线检测。

关键词：网络安全；伪装攻击；入侵检测；shell 命令；异常检测；Markov 链

中图分类号：TP393 **文献标识码**：B **文章编号**：1000-436X(2011)03-0098-08

Masquerade detection based on shell commands and Markov chain models

XIAO Xi[1], TIAN Xin-guang[2], ZHAI Qi-bin[1], YE Run-guo[3]

(1.State Key Laboratory of Information Security, Graduate University of Chinese Academy of Sciences, Beijing 100049,China;
2.Key Laboratory of Network Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190,China;
3.Beijing Venustech Company Ltd, Beijing 100193,China)

Abstract: A novel method for masquerade attack detection based on shell commands was proposed. At the training stage, the variability of users' behavior and the feature of masquerade attack were thoroughly considered, and stationary homogeneous Markov chains were employed to profile the normal users' behavior. The shell commands were gradationally merged into multiple sets according to their frequencies and then states were constructed accordingly, which significantly reduced the number of states and the memory of the transition probability matrix and improved the generalization of the detection system, compared with existing Markov chain methods. Considering the real-time detection demand and the short-time relevance of shell commands, the states were matched with a high-frequency-first scheme at the detection stage, and the decision measure was computed by smoothing the probabilities of short state sequences. This decreased computational complexity and the false-alarm rate. Experimental results indicate that our method can achieve high detection accuracy and practicability, and is especially applicable for on-line detection.

Key words: network security; masquerade attack; intrusion detection; shell command; anomaly detection; Markov chain

---

收稿日期：2010-06-09；**修回日期**：2010-09-20

基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2006AA01Z452）；国家 242 信息安全计划基金资助项目（2005C39）

Foundation Items: The National High Technology Research and Development Program of China (863 Program) (2006AA01Z452); The National Information Security 242 Program of China (2005C39)

1　引言

伪装攻击（masquerade attack）是指非授权用户通过伪装成合法用户来获得访问关键数据或更高层访问权限的行为[1~3]。伪装攻击检测自 Smaha S E[1]1988 年首次提出以来，作为入侵检测的一个主要分支在近些年得到了越来越多的研究，并在网络信息安全领域中发挥着越来越大的作用[2~8]。入侵检测技术主要有异常检测和误用检测 2 种基本类型。目前的伪装攻击检测系统大多采用异常检测技术，这种技术对合法用户的正常行为进行建模，通过比较被监测用户的实际行为和合法用户的正常行为来检测入侵（攻击），其具有不需要过多有关入侵行为的先验知识，且能检测出未知攻击的优点，但在某些领域有虚警率高的缺点。Li M 在 DDOS 异常检测中采取有效的措施来降低虚警率，取得了一定成果[9~11]。

伪装攻击检测面临的主要困难在于用户行为的复杂性和多变性, 即用户行为会随着工作内容、时间和其他不确定性因素的变化而改变[6,12,13]。国内外已经开展了机器学习、Markov 模型、贝叶斯分类器、支持向量机、数据挖掘等技术在用户伪装攻击检测中的应用研究。Lane T 等人[13,14]研究了基于实例学习的伪装攻击检测方法，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，该方法原理较为简单，有较强的适应能力，但在检测阶段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，因而准确率较低。田新广等人[8]在 Lane T 的基础上改进了对用户行为模式的表示方式，以 shell 命令序列为单位进行相似度计赋值，提高了检测准确率和检测性能的稳定性。Schonlau M 等人[15]研究了基于统计理论的伪装攻击检测方法，综合比较了 6 种不同的方法，并根据实验结果分析了各种方法的优势和局限性。Maxion R A 等人[12]引入了贝叶斯分类算法，对 Schonlau M 的检测方法进行了改进，提高了检测准确率。最近，Dash S K 等人[6]提出延迟检测概念（deferred detection concept），把适应性朴素贝叶斯方法应用到伪装攻击检测，检测性能得到进一步提高。Tian X G 等人[5]提出了基于 Markov 链模型的检测方法，有良好的检测性能；但是该方法把不同的 shell 命令(符号)当作不同的状态，存在状态数目过多，计算复杂度大，容错能力和泛化能力不强等缺点。Coull 等人[7]采用命令分组（command grouping）和二元计分（binary scoring）的方法模拟变异（mutation），把生物信息学里的序列比对(sequence alignment) 技术应用于伪装攻击检测。此外，Kim H S 等人[16]和 Shim C Y 等人[4]研究了基于支持向量机的方法，Szymanski B K 等人[17]和 Wu H C 等人[3]提出了数据挖掘的方法，这些方法检测效率较高，但对用户行为变化的适应性不强。

大部分伪装攻击过程都包含一系列前后相关的行为，这些相关的行为特征与正常行为过程有明显的差别。Markov 链模型是刻画前后相关行为之间转移关系的强有力的数学工具，它能够利用这种差别来检测用户行为的异常[18]。本文在以上工作的基础上，提出一种新的基于 Markov 链模型的用户伪装攻击检测方法。该方法在训练阶段充分考虑了用户行为的多变性和伪装攻击的特点，改进了对用户正常行为模式的表示方式，采用平稳的齐次 Markov 链对合法用户的正常行为进行建模，根据 shell 命令的出现频率进行阶梯式数据归并来划分状态，同现有的 Markov 链方法相比[5]，大幅减少了状态个数和转移概率矩阵的存储空间，提高了检测系统的泛化能力。在检测阶段，考虑到实时性需求和 shell 命令操作的短时相关性，采用了基于频率优先的状态匹配方法，通过对状态短序列的出现概率进行加窗平滑滤噪处理来进行判决值计算，能够有效减少系统计算开销，降低误报率。实验表明，同现有的 4 种典型检测方法相比，本文方法在减少存储成本和计算成本的同时，提高了检测准确率，改善了系统的整体性能，具有较强的实用性和可操作性，特别适用于在线检测。

2　相关知识

2.1　Markov 链

定义 1　考虑只取有限个值 1,2,…,N 的随机过程$\{X_n, n=1,2,\cdots\}$，它所取可能值的全体称为状态空间，记为 $\Omega=\{1,2,\cdots,N\}$。它在 i 时刻的状态记为 q_i，如果该随机过程在 $m+k$ 时刻所处的状态为 q_{m+k} 的概率，只与它在 m 时刻的状态 q_m 有关，而与 m 时刻之前它所处的状态无关，即

$$P(X_{m+k}=q_{m+k}/X_1=q_1,\cdots,X_{m-1}=q_{m-1},X_m=q_m)=P(X_{m+k}=q_{m+k}/X_m=q_m) \quad (1)$$

（$q_1,q_2,\cdots,q_m,q_{m+k}\in\Omega=\{1,2,\cdots,N\}$），则称该随机过程

$\{X_n\}$为 Markov 链。

定义 2 对于 N 个状态的 Markov 链$\{X_n\}$，称

$$P_{ij}(m,\ m+k)=P(X_{m+k}=j/X_m=i),\quad 1\leqslant i,j\leqslant N \tag{2}$$

为 k 步转移概率。如果 $P_{ij}(m, m+k)$与 m 无关，称这个 Markov 链为齐次 Markov 链。当 k=1 时，记 $a_{ij}=P_{ij}(m,\ m+1)$，称为（一步）转移概率。并记 $\pi_i=P(X_1=i)$为状态 i 的初始出现概率。

定义 3 对于 N 个状态的齐次 Markov 链$\{X_n\}$，称矩阵 $\boldsymbol{A}=(a_{ij})_{N\times N}$ 为状态转移概率矩阵；并称向量 $\boldsymbol{\pi}=(\pi_1,\ \pi_2,\cdots,\ \pi_N)$为初始状态概率向量（或初始概率分布）。

定理 1 设$\{X_n,\ n=1,2,\cdots\}$为 N 个状态的齐次 Markov 链，则

$$P(q_i,q_{i+1},\cdots,q_j)=P(q_i)\prod_{k=i}^{j-1}P(q_{k+1}/q_k) \tag{3}$$

定义 4 设 N 个状态的 Markov 链$\{X_n\}$有状态转移概率矩阵 $\boldsymbol{A}=(a_{ij})_{N\times N}$，若存在一个概率分布 $\boldsymbol{\pi}=(\pi_1,\ \pi_2,\cdots,\ \pi_N)$满足：

$$\pi_i=\sum_{j=1}^{N}\pi_j a_{ji},\quad 1\leqslant i\leqslant N \tag{4}$$

则称 $\boldsymbol{\pi}$ 为该 Markov 链的平稳分布。

定理 2 如果 N 个状态的 Markov 链$\{X_n\}$的初始概率分布是平稳分布，即 $\pi_i=P(X_1=i)$满足式(4)，则状态 i 在 n 时刻出现的概率等于状态 i 的初始出现概率，即

$$P(X_n=i)=P(X_1=i)=\pi_i,\quad 1\leqslant i\leqslant N \tag{5}$$

注：从定理 1 和 2 可知，平稳的齐次 Markov 链由初始概率分布和状态转移概率矩阵完全刻画。

2.2 审计数据的分析及预处理

与文献[2~8,12~17]中的伪装攻击检测实验相同，本文的实验采用 UNIX 平台上的 shell 命令作为审计数据。主要是考虑到：1）shell 命令容易收集，形式简单，便于分析；2）在 UNIX 平台上，shell 是终端用户与操作系统之间最主要的界面，能反映用户的行为，且大部分用户活动都是利用 shell 完成的。

本文方法在训练和检测阶段，需要对用户的原始 shell 命令数据进行预处理。预处理的方式有 2 种。第 1 种方式如文献[5,8,13,14]所述，预处理时滤除 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了标识符号。第 2 种方式如文献[12,15]所述，预处理时只保留 shell 命令的名称，滤除命令参数和时间等信息。经过预处理后的原始 shell 命令数据表现形式都是 shell 命令有序字符串（按时序排列的若干个 shell 命令符号）。

2.3 序列流

定义 5 设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$为一个长度为 n 的有序字符串，称 $\overline{x_i}=(x_i,x_{i+1},\cdots,x_{i+l-1})$ 为在 $\boldsymbol{x}$ 上以 l 为窗长截取出来的第 i 个（短）序列（$1\leqslant i\leqslant n-l+1$），称由（短）序列 $\overline{x_i}$ 构成的序列 $\overline{\overline{x}}=(\overline{x_1},\overline{x_2},\cdots,\overline{x_{n+l-1}})$ 为由 $\boldsymbol{x}$ 以 l 为窗长生成的（短）序列流，简称为 $\boldsymbol{x}$ 的（短）序列流。

注：序列里的字符和序列流里的序列都是有序排列的，本文里的有序都是指时间先后次序。

3 新的基于 Markov 链模型的用户伪装攻击检测方法

3.1 训练

用户伪装攻击检测的具体实现过程分为训练和检测 2 个阶段。训练阶段的主要工作是根据训练数据对我们所关心的合法用户的正常行为进行建模。由于伪装攻击的用户行为的多变性，得到的训练数据往往不够充分[12]，这就要求训练和检测模型对用户行为具有一定的泛化能力和容错能力。非平稳非齐次的 Markov 链模型能够更加精细的描述用户行为。但模型描述能力过于精细，对用户复杂多变的行为适应性会较差，会使系统的容错能力和泛化能力较弱。反之，如果模型描述能力过于粗略，系统的泛化能力有可能增强，但同时会导致检测准确度的降低。兼顾到对用户行为的适应能力和实际检测中对检测准确度的要求，本文训练阶段中采用平稳的齐次 Markov 链模型对用户正常行为建模。

3.1.1 确定状态

本文方法预先设定 Markov 链的状态个数为 N，先根据 shell 命令的出现频率进行阶梯式数据归并，然后利用频率优先方法进行状态匹配。确定状态的具体步骤如下。

1) 获取该合法用户的正常行为的训练数据。设经 2.2 节预处理后的正常行为训练数据为 $\boldsymbol{s}=(s_1,s_2,\cdots,s_r)$，它是一个长度为 r 的 shell 命令有序字符串。

2) 提取出 $\boldsymbol{s}$ 中互不相同的 shell 命令符号并按其出现的频率从大到小排序。设 $\boldsymbol{s}$ 中互不相同的 shell 命令符号共有 W 个（$W\leqslant r$）。shell 命令符号在 $\boldsymbol{s}$ 中出现的频率等于它在 $\boldsymbol{s}$ 中出现的次数除以 r。将这 W 个不同的符号按其在 $\boldsymbol{s}$ 中出现的频率从大到小排序，排序后记为 $s_1^*, s_2^*, \cdots, s_W^*$，设 F_j^* 为 s_j^* 在 $\boldsymbol{s}$ 中出现的频率（$1\leqslant j\leqslant W$），则有 $F_1^*\geqslant F_2^*\geqslant\cdots\geqslant F_W^*$。

3) 根据 shell 命令的出现频率进行阶梯式数据归并。把排序后的 shell 命令符号 $s_1^*, s_2^*, \cdots, s_W^*$ 按频率从大到小进行数据归并，聚合为 $N-1$ 个集合。设 $a=\lceil W/(N-1)\rceil$（$\lceil y\rceil$ 表示不小于 y 的最小整数），第 1 个集合为 $Q_1=\{s_1^*, s_2^*, \cdots, s_a^*\}$，第 2 个集合为 $Q_2=\{s_{a+1}^*, s_{a+2}^*, \cdots, s_{2a}^*\}$，…，第 $N-1$ 个集合为 $Q_{N-1}=\{s_{(N-2)a+1}^*, s_{(N-2)a+2}^*, \cdots, s_W^*\}$。

4) 利用频率优先匹配方法确定 m 时刻出现的 shell 命令符号 s_m 的状态 q_m：依次在第 1、2、3、… 个集合 Q_1、Q_2、Q_3、…中查找 s_m，如果在第 $j(1\leqslant j\leqslant N-1)$ 个集合 Q_j 中查找到 s_m，则 $q_m=j$；如果在所有 $N-1$ 个集合中都查找不到 s_m，则 $q_m=N$。此 Markov 链的状态空间为 $\Omega=\{1,2,\cdots,N-1,N\}$。

在以上步骤中，步骤 3）根据 shell 命令的出现频率进行阶梯式数据归并以划分状态是对文献[5]中用户行为模式表示方式的改进，状态对应的 shell 命令范围扩大，提高了检测系统的泛化能力和容错能力。步骤 4）在检测阶段频繁使用，此处利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符号进行比较，这样可以节省 shell 命令符号的匹配时间，减少了系统的计算开销。

3.1.2 计算参数

参数计算是 Markov 链在用户行为异常检测中应用的关键问题。设描述该用户正常行为的 Markov 链的初始状态概率向量为 $\boldsymbol{\pi}=(\pi_1, \pi_2, \cdots, \pi_N)$，其具体计算方法为

$$\pi_j=\begin{cases}\sum\limits_{i=(j-1)a+1}^{ja} F_i^*, & 1\leqslant j\leqslant N-2\\ \sum\limits_{i=(j-1)a+1}^{W} F_i^*, & j=N-1\\ 0\ , & j=N\end{cases}\tag{6}$$

设描述该用户正常行为的 Markov 链的状态转移概率矩阵为 $\boldsymbol{A}=(a_{ij})_{N\times N}$。为了减少检测系统的复杂度，不考虑具体时间对转移概率的影响，分前后 2 个步骤来计算 $\boldsymbol{A}$（此时的训练数据已经在 3.1.1 节的步骤 1）中被预处理成 shell 命令有序字符串 $\boldsymbol{s}=(s_1, s_2, \cdots, s_r)$）。

第 1 步：利用频率优先匹配法把 shell 命令有序字符串转化为状态有序字符串，见以下①～③；

第 2 步：统计转移次数，然后归一化为转移概率，见以下④～⑦。

计算 $\boldsymbol{A}$ 的伪代码如下：

① $m:=1$；

② 根据 3.1.1 节步骤 4）确定 shell 命令符号 s_m 对应的状态 q_m；

③ $m:=m+1$；

if $m\leqslant r$，返回②；else 继续；

④ $\boldsymbol{A}=(a_{ij})_{N\times N}:=0$，$a_{NN}:=1$；$k:=1$；

⑤ $i:=q_k$，$j:=q_{k+1}$；$a_{ij}:=a_{ij}+1$；

⑥ $k:=k+1$；

if $k\leqslant r-1$，返回⑤；else 继续；

⑦ 对 $1\leqslant i,j\leqslant N$，if $\sum\limits_{j=1}^{N} a_{ij}>0$，$a_{ij}:=a_{ij}\Big/\sum\limits_{j=1}^{N} a_{ij}$。

3.2 检测

检测阶段的工作是根据训练阶段所建立的合法用户正常行为轮廓，利用特定的检测模型来识别被监测用户当前行为中的异常。因为 shell 命令操作有短时相关性，利用以上计算出的参数，基于 shell 命令短序列对应的状态短序列的出现概率对该用户的当前行为进行判决。主要步骤如下。

1) 获取被监测用户在被监测的时间内执行的 shell 命令行，把这些命令行数据预处理成一个长度为 t 的 shell 命令有序字符串，然后利用 3.1.2 节第一步的频率优先匹配方法快速把这个 shell 命令有序字符串转化为状态有序字符串 $\boldsymbol{q}=(q_1,q_2,\cdots,q_t)$。

2) 由状态有序字符串 $\boldsymbol{q}=(q_1,q_2,\cdots,q_t)$ 以 u 为窗长生成状态短序列流 $\overline{\overline{q}}=(\overline{q_1},\overline{q_2},\cdots,\overline{q_{t-u+1}})$，其中，第 i 个状态短序列 $\overline{q_i}=(q_i,q_{i+1},\cdots,q_{i+u-1})$，$1\leqslant i\leqslant t-u+1$。

3) 根据 $\boldsymbol{\pi}$ 和 $\boldsymbol{A}$，计算该用户的状态短序列流 $\overline{\overline{q}}$ 中每个状态短序列 $\overline{q_i}$ 的出现概率。记该用户的第 i 个状态短序列 $\overline{q_i}$ 的出现概率为 $P(\overline{q_i})$，i 时刻出现状态 q_i 的概率为 $P(q_i)$，从状态 q_i 到 q_{i+1} 的转移概率为 $P(q_{i+1}/q_i)$。设 $q_i=m$，$q_{i+1}=n$，则 $P(q_{i+1}/q_i)=a_{mn}$，且由定理 2 式(5)可知 $P(q_i)=\pi_m(1\leqslant m,n\leqslant N)$。由定理 1 中式（3）可得 $P(\overline{q_i})$ 的计算公式为

$$P(\overline{q_i}) = P(q_i) \prod_{h=i}^{i+u-2} P(q_{h+1}/q_h) \tag{7}$$

特别当 u=2 时，$P(\overline{q_i}) = \pi_m \times a_{mn}$。

4) 在概率序列 $(P(\overline{q_1}), P(\overline{q_2}), \cdots, P(\overline{q_{t-u+1}}))$ 中定义判决值函数 $D(n)$。因为用户在短时间内的行为可能会偏离其历史行为，所以检测时并不直接利用 $P(\overline{q_i})$ 对用户行为进行判决，而是对概率序列进行加窗平滑滤噪处理来计算判决值。设窗口长度为 w，采用以下 2 个判决值：

第 1 个：定义判决值为

$$D(n) = \frac{1}{w} \sum_{i=n-w+1}^{n} \mathrm{sgn}[P(\overline{q_i}) - a] \tag{8}$$

式中 a 为概率门限需预先设定。

第 2 个：定义判决值为

$$D(n) = \frac{1}{w} \sum_{i=n-w+1}^{n} \lg[\hat{P}(\overline{q_i})] \tag{9}$$

其中

$$\hat{P}(\overline{q_i}) = \begin{cases} P(\overline{q_i}), & P(\overline{q_i}) > e \\ e, & P(\overline{q_i}) \leqslant e \end{cases}, \quad 1 \leqslant i \leqslant t-u+1 \tag{10}$$

式中 $e \geqslant 0$ 为截断常数需预先设定。此处对 $P(\overline{q_i})$ 进行截幅处理使对数运算恒有意义，且能防止判决值过小，减小虚警概率。

5) 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”判为正常行为；否则，判为异常行为。根据文献[18]，序列概率越大，其代表正常行为的可能性越大，而代表异常行为的序列概率较小。这里，“当前行为”是相对于状态短序列 $\overline{q_n}$ 而言的，它是指以 $\overline{q_n}$ 为终点的 w 个状态短序列 $\overline{q_{n-w+1}}, \overline{q_{n-w+2}}, \cdots, \overline{q_n}$ 对应的行为，也即以 q_{n+u-1} 为终点的 $w+u-1$ 个状态 q_{n-w+1}，q_{n-w+2}，…，q_{n+u-1} 对应的行为。检测中，判决门限 d 的选择应综合考虑检测要求、判决值计算公式、状态短序列长度 u、窗长度 w 等多种因素。

4 特点分析

本文基于平稳齐次 Markov 链模型的伪装检测方法有以下几个特点。

1) 同文献[8,13,14]中的机器学习方法相比，本文方法考虑了用户行为模式的频率分布和行为模式之间的相关性，更加精确地描述了用户正常行为轮廓，具有更高的检测准确率。同文献[5]基于 Markov 链模型的方法相比，本文方法充分考虑了用户行为模式的多变性和伪装攻击的特点，没有把不同 shell 命令符号当作不同状态，而是把不同 shell 命令符号根据其出现的频率阶梯式归并成为数不多的集合，每个集合对应 1 个状态，检测系统的泛化能力增强，检测准确率得到提高。

2) 本文方法中的状态个数可综合具体被监测用户的行为特点、存储成本、计算成本和检测效率来灵活设定，可以取到最小值 2。已有文献[5]基于 Markov 链模型的方法中状态个数是固定不变的，本文方法的灵活性和可操作性增强。

3) 与文献[13,14]中基于实例学习的方法中存储等长 shell 命令序列和文献[8]中机器学习方法中存储多个不同长度的 shell 命令序列相比，本文只存储不同的 shell 命令符号，存储空间减少。与文献[5]基于 Markov 链模型的方法相比，存储的 shell 命令符号一样多，但状态个数和转移概率矩阵的存储空间得到了大幅度的减少。

4) 文献[13,14]中 Lane T 提出的序列相似度计算方法要与样本序列库里的所有序列的所有字符比对，运算量很大，而本文状态序列概率的计算只与初始概率和转移概率矩阵有关，本文状态个数可以取得很少，计算量能显著减少。文献[8]需要处理多个长度不等的 shell 命令序列，而本文仅需处理单个的 shell 命令符号（即一个固定长度为 1 的序列），计算成本降低。文献[5]参与计算的转移概率矩阵维数远大于本文方法，相应的计算量大于本文方法。同时，频率优先匹配法的使用进一步降低了系统的计算成本。

5 实验设计与结果分析

作者利用国际上公用的伪装检测数据——Purdue 大学的 shell 命令实验数据对上述方法的性能进行了实验。Purdue 大学的数据包含 8 个 UNIX 用户在 2 年时间内的活动记录（实验数据的详细说明见文献[13,14]）。实验中采用了其中的 4 个用户 user1、user2、user3、user4 的数据，并且将 user3 设为合法用户，将 user1、user2、user4 设为伪装用户。user3 的实验数据（shell 命令有序字符串）中

有 15 000 个 shell 命令，其中，前 10 000 个命令作为正常行为训练数据用于 Markov 链的状态确定和参数计算，后 5 000 个命令作为正常行为测试数据用于检测性能（主要是虚警概率）的测试。user1，user2，user4 的实验数据各包含 5 000 个 shell 命令，这些命令均作为异常行为测试数据用于检测概率的测试。参数设置为 N=3，u=2，w=91，a=10^{-4}，e=10^{-20}。

实验时，正常行为训练数据中互不相同的 shell 命令符号共有 200 个，Markov 链的状态个数为 3。Markov 链的初始状态概率向量 $\boldsymbol{\pi}$=(0.987 8, 0.012 2, 0)，状态转移概率矩阵

$$\boldsymbol{A}=\begin{pmatrix} 0.988\,66 & 0.011\,34 & 0 \\ 0.918\,03 & 0.081\,97 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

计算可知 $\boldsymbol{\pi}\times\boldsymbol{A}$=(0.987 8, 0.012 2, 0)=$\boldsymbol{\pi}$，满足平稳分布条件式(4)，故此链是平稳的。又因为计算 $\boldsymbol{A}$ 时没有考虑具体时间对转移概率的影响，所以它是平稳的齐次 Markov 链。

5.1　检测准确率分析

图 1 给出了由式(8)计算出的判决值曲线。由图可看出，判决值曲线具有很好的可分性。

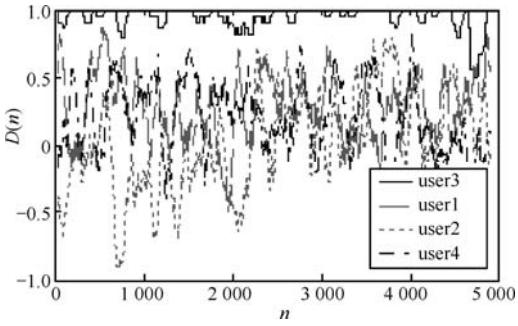


图 1　式（8）对应的判决值曲线

图 2 示出了本文方法由式(8)和式(9)作为判决值反映虚警概率与检测概率对应关系的 ROC 曲线和文献[5,7,8,14]中方法的 ROC 曲线。实验结果表明采用式(8)、式(9)计算判决值均获得了很高的检测准确率；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，并且可获得很高的检测准确率。由图可见，本文方法比已有文献[5,7,8,14]中的 4 种方法的检测准确率均有明显的提高。（图中各种方法的参数是在保证平均检测时间基本相同的前提下设置的[19]。）

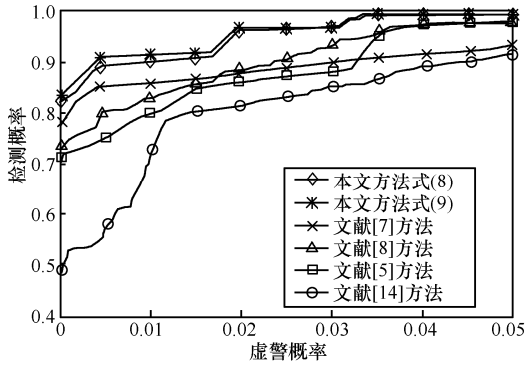


图 2　5 种不同方法的 ROC 曲线

5.2　状态个数对检测性能的影响

图 3 给出了由式（8）作为判决值时不同状态的 ROC 曲线。从图 3 可看出式（8）作判决值时，2、3、6、11 个状态的 ROC 曲线很接近，11 个状态的检测准确率稍好一点。考虑到方法的一般性，在实验中选择了 3 个状态。实际上由图 2 和图 3 可知，利用本文的 2 个状态的 Markov 链方法也比其他 4 种方法好（因为 2 个状态和 3 个状态的 ROC 曲线基本重合了）。

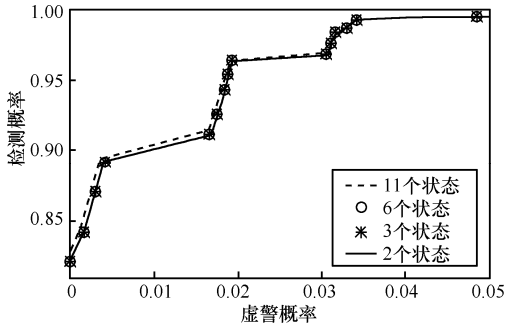


图 3　式(8)作为判决值时不同状态的 ROC 曲线

5.3　存储空间和实验时间分析

现有文献[13,14]和文献[8]分别提出了 2 种不同的机器学习方法，文献[5]提出了基于 Markov 链模型的方法，文献[7]提出了生物信息学的方法。表 1 列出了本文方法和以上 4 种不同方法的存储空间和实验时间（相同实验条件下测量）。

表 1　　5 种方法的存储空间和实验时间

指标	文献[14]方法	文献[7]方法	文献[8]方法	文献[5]方法	本文方法
存储单元/个	2 544	10 000	3 512	40 601	212
实验时间/s	1 818.6	1 267.4	28.318	19.208	17.068

从表 1 可看出本文方法存储空间是所有方法中最少的，仅是文献[5]的 0.52%，减少了 3 个数量级。

文献[5]中绝大部分存储空间用于转移概率矩阵，其 Markov 链的状态 201 个，而本文实验方法只取 3 个状态，减少了 2 个数量级，转移概率矩阵的存储空间仅为文献[5]的 0.02%，减少了 4 个数量级。同时，从表 1 可看出本文方法实验时间是所有方法中最少的。本文方法实验时间仅是文献[14]实验时间的 0.94%，减少了 2 个数量级；同时它也是文献[5]实验时间的 88.86%(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)。

综合考虑检测准确率、存储空间和实验时间，本文方法的整体性能优于已有的 4 种方法。

5.4 在其他数据上的实验

同时，作者也在 AT&T Shannon 实验室的 shell 命令实验数据中（详细说明见文献[12,15,17]）前 4 个用户 user1、user2、user3、user4 的数据上进行了实验。其中，每个用户有 5 000 个 shell 命令，实验时将 user3 设为合法用户，该用户的前 4 000 个命令作为训练数据用于正常行为建模，后 1 000 个命令作为测试数据用于测试虚警概率；其他 3 个用户设为伪装用户，它们的 5 000 个 shell 命令均作为测试数据用于测试检测概率。参数设置为 N=3，u=2，w=100，a=0.03，e=10^{-20}。图 4 示出了本文方法由式(8)和式(9)作为判决值的 ROC 曲线和文献[5,7,8,14]中方法的 ROC 曲线。可见，本文方法比已有文献[5,7,8,14]中的 4 种方法的检测准确率均有明显的提高。

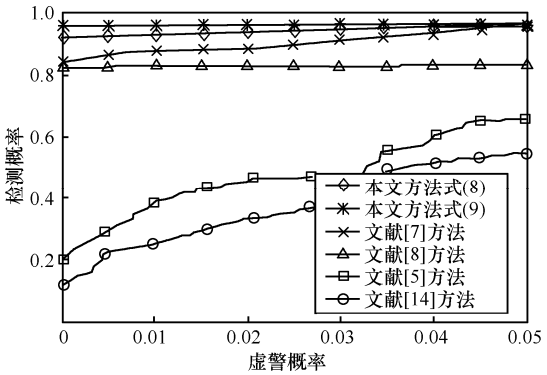


图 4 在 AT&T Shannon 实验室数据上 5 种不同方法的 ROC 曲线

6 结束语

本文提出一种高效的基于 shell 命令和 Markov 链模型的伪装攻击检测方法。该方法在训练阶段，改进了对用户正常行为模式的表示方式，同现有的 Markov 链方法[5]相比大幅度减少了状态个数和转移概率矩阵的存储量，提高了检测系统的泛化能力；在检测阶段，利用频率优先匹配状态，通过对状态短序列出现的概率序列进行加窗平滑滤噪处理以计算判决值，有效减少系统计算开销，降低误报率。实验表明，同已有的 4 种典型检测方法相比，本文方法在减少存储成本和计算成本的同时，提高了检测准确率，改善了系统的整体性能，具有很强的实用性和可操作性，特别适用于在线检测。而且，在实际应用中还可以通过优化参数设置进一步提高性能。

参考文献：

[1] SMAHA S E. Haystack: an intrusion detection system[A]. The Fourth IEEE Aerospace Computer Security Applications Conference[C]. Orlando, Florida, 1988.

[2] 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.
TIAN X G, DUAN M Y, CHENG X Q. Masquerade detection based on shell commands and multiple behavior pattern mining[J]. Chinese Journal of Computers, 2010, 33(4): 697-705.

[3] WU H C, HUANG S H S. Masquerade detection using command prediction and association rules mining[A]. 2009 International Conference on Advanced Information Networking and Applications[C]. Aina,2009. 552-559.

[4] SHIM C Y, KIM J Y, GANTENBEIN R E. Practical user identification for masquerade detection[A]. Advances in Electrical and Electronics Engineering-IAENG Special Edition of the World Congress on Engineering and Computer Science 2008[C]. San Francisco, California, USA, 2008. 47-51.

[5] TIAN X G, DUAN M Y, LI W F, *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.

[6] 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.

[7] 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.

[8] 田新广, 高立志, 张尔扬. 新的基于机器学习的入侵检测方法[J]. 通信学报, 2006, 27(6):108-114.
TIAN X G, GAO L Z, ZHANG E Y. Intrusion detection method based on machine learning[J]. Journal on Communications, 2006, 27(6): 108-114.

[9] LI M. An approach to reliably identifying signs of DDOS flood attacks based on LRD traffic pattern recognition[J]. Computers & Security, 2004, 23(7): 549-558.

[10] LI M. Change trend of averaged Hurst parameter of traffic under

DDOS flood attacks[J]. Computers & Security, 2006, 25(3): 213-220.

[11] LI M, WANG S, ZHAO W. A real-time and reliable approach to detecting traffic variations at abnormally high and low rates[J]. Lecture Notes in Computer Science, 2006, 4158: 541-550.

[12] MAXION R A, TOWNSEND T N. Masquerade detection using truncated command lines[A]. Proceedings of the International Conference on Dependable Systems and Networks[C]. Washington, DC, USA, 2002. 219-228.

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

[14] LANE T, CARLA E B. An empirical study of two approaches to sequence learning for anomaly detection[J]. Machine Learning, 2003, 51(1): 73-107.

[15] SCHONLAU M, MOUCHEL W. Computer intrusion: detecting masquerades[J]. Statistical Science, 2001,16(1): 58-74.

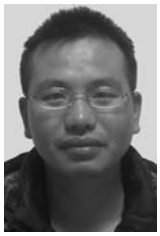
[16] KIM H S, CHA S D. Empirical evaluation of SVM-based masquerade detection using UNIX commands[J]. Computers & Security, 2005, 24(2):160 -168.

[17] SZYMANSKI B K, ZHANG Y Q. Recursive data mining for masquerade detection and author identification[A]. Proceedings of the 5th IEEE System, Man and Cybernetics Information Assurance Workshop[C]. West Point, NY, USA, 2004. 424-431.

[18] PATCHA A, PARK J M. An overview of anomaly detection techniques: Existing solutions and latest technological trends[J]. Computer Networks, 2007, 51(12): 3448-3470.

[19] 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.
TIAN X G. Anomaly Detection Methods for Host-based Intrusion Detection Systems[D]. Changsha, China: National University of Defense Technology, 2005.

作者简介：



肖喜（1979-），男，湖南宜章人，中国科学院研究生院博士生，主要研究方向为入侵检测、信息安全和密码应用技术。

田新广（1976-），男，河北吴桥人，中国科学院博士后，主要研究方向为网络安全、入侵检测和智能信息处理。

翟起滨（1947-），男，黑龙江哈尔滨人，中国科学院研究生院教授、博士生导师，主要研究方向为密码学、信息安全和入侵检测。

叶润国（1976-），男，江西萍乡人，博士后，北京启明星辰信息安全技术有限公司资深安全工程师，主要研究方向为数据挖掘和网络安全。