# 确定性退火算法在/ 伪装0入侵行为检测中的应用

赵俊忠[1], 黄厚宽[2], 田盛丰[2]
(11 北京航空航天大学理学院, 北京 100083; 21 北方交通大学计算机与信息技术学院, 北京 100044)

摘 要: 本文提出了一种基于确定性退火算法的检测/ 伪装0 入侵行为的方法. 在该方法中, 每一个用户被看作是一个离散变长记忆的平稳信源, 被/ 伪装0的入侵者利用的账户所产生的命令行字符序列可以被看作是由该账户的相应用户和/ 伪装0的入侵者两个不同信源在不同时段活动的混合结果. 我们通过对命令行字符序列的分析来重构原信源模型以判断是否存在入侵行为. 实验结果表明该模型是可行的.

关键词: 网络安全; 入侵检测系统; 信息率失真理论; 确定性退火
中图分类号: TP393108 文献标识码: A 文章编号: 037222112 (2004) 022030320 3

# Detecting Masquerades in Intrusion Detection Based on Deterministic Annealing

ZHAO Jun2zhong[1], HUANG Hou2kuan[2], TIAN Sheng2feng[2]
(11 School of Science, Beijing University of Aeronautics and astronautics, Beijing 100083, China;
21 School of Computer and Information Technology, Northern Jiaotong University, Beijing 100044, China)

Abstract: A new model based on deterministic annealing for detecting intruders/ users masquerading as other users is presented. In our model, each user is viewed as a discrete stationary source with variable memory. A sequence of characters composed of command lines from a user. s account is regarded as the result that is potentially generated by the user and the intruder in different period. We determine masquerades by finding the source( s) in the sequence. Our experiment shows that the model is feasible.

Key words: network security; intrusion detection system; rate distortion theory; deterministic annealing

## 1 引言

/ 伪装0 是指为了隐藏身份和躲避检测而有意地模仿其他计算机用户的入侵行为. 不同的用户由于所从事的工作、所具有的兴趣、所受的教育和所成长的环境不同, 其行为和行为方式通常总是具有显著的可统计量化的差异. 同一个用户的行为尽管也会随时间而发生变化, 但这种行为变化通常具有一种平稳的前后连续性, 通常不会造成统计特征突然和显著的变化, 而其长期形成的习惯性行为方式就更难以出现这种变化了. 迄今为止的绝大多数异常检测算法通常是针对正常行为和异常行为的差异或异常行为的共性, 而并没有强调不同用户行为的个体差异, 这是造成异常检测算法误报率和检测率不理想的主要原因之一. 相应地, 利用个体的统计描述进行入侵检测的模型表现出了较高的检测率和准确性[1,2]. 另外, 当前的入侵检测系统都不检测命令行参数, 而只对用户所使用的命令进行分析, 这不可避免地影响到了检测的可靠性和准确性. 在本文中, 我们将每个用户看作是一个具有某种统计特性的符号发生器, 即信源, 并通过对每个用户所键入的命令行字符序列的分析来重构该用户的信源模型, 以把握不同用户的不同行为特征.

## 2 离散有记忆平稳信源的上文函数和层次树表示

设集合 $E = \{c_1, c_2, \cdots, c_r\}$ 为有限字符集, 随机变量序列 $X = X_1X_2\cdots$ 表示一个离散有记忆平稳信源, D 为定义在 X 上的一个映射函数. 对 X 的一个样本序列 $x_1x_2\cdots$ 的任意子序列 $x_1x_2\cdots x_i$ ($i = 1, 2, \cdots$), 定义

$$\begin{aligned} &D(x_1x_2\cdots x_i) = x_{i-m+1}x_{i-m+2}\cdots x_i \\ &m = \min\{k \mid P(x_{i+1} \mid x_1x_2\cdots x_i) \\ &\quad = P(x_{i+1} \mid x_{i-k+1}x_{i-k+2}\cdots x_i)\} \end{aligned} \tag{1}$$

特别地, 当 m= 0 时, 定义 $D(x_1x_2\cdots x_{i-1}x_i) = <$. 在此, 函数 D 称为 X 的上文函数, $D(x_1x_2\cdots x_i)$ 称为变量 $x_{i+1}$ 的上文序列[3]. 一个离散有记忆平稳信源的所有上文序列形成一个集合, 记为上文集合 C.

上述的每个离散有记忆平稳信源都用一棵层次树来表示[4]. 在层次树中, m= 0 的第 0 层只有一个根节点, 对应空上文序列 <; 其余各层( m> 0) 的每个节点对应上文集合 C 中的一个非空的上文序列. 每个节点的描述由五部分构成. 前两个部分是二进制位向量 $B = (b_1, b_2, \cdots, b_{|E|})$ 和条件概率向量 $P = (p_1, p_2, \cdots, p_{|E|})$. 向量 B 中的每个分量 $b_i$ 用一位二进制数来表示, 根据该节点的各子节点所表示的上文序列最前导

收稿日期: 2002212226; 修回日期: 2003206228

命令行字符将向量 B 中与该字符相对应的位置 1; 向量 P 中的每个分量$p_i$是该节点所表示的序列为上文的条件下, 下一个命令行字符为 $c_i$ I E 的概率.

在层次树中, 每个节点并不存储完整的上文序列, 而只是存储上文序列中最前导的命令行字符, 这就是第三部分; 每个节点所表示的完整的上文序列由从该节点到根节点的路径形成. 描述的第四和第五部分为两个指针, 分别指向当前节点的下一个兄弟节点(同长度的另一个上文序列) 和子节点链的第一个子节点.

## 3 混合信源模型识别算法

设 E 是由组成 UNIX 命令及参数的所有合法字符构成的有限字符集合, 而 x= $x_1x_2$, $x_l$ 是某用户账户下最近使用过的命令行(包括命令及其参数) 依序连接而形成的一条定义在 E 上的长度为 l 的字符序列.

设 T= $\{T_j\}_{j=1}^{k}$为当前的 k 个离散有记忆平稳信源的层次树组成的集合. 对于任意 $T_j$ I T, 用 $T_j$ 编码以 $x_i$ 为中心的字符序列$x_{i-M}$, $x_{i-1}x_ix_{i+1}$, $x_{i+M}$(窗口长度为 2M+ 1) 的局部失真函数 d($T_j$, $x_i$)定义为:

$$d(T_j, x_i) = -\sum_{t=i-M}^{i+M} \log P_{T_j}(x_t | x_1x_2, \ x_{t-1}) \tag{2}$$

相应地, 用 $T_j$ 编码 x 时的整体失真函数定义为 d($T_j$, $x_i$)的期望, 即平均失真函数.

$$d(T_j, x) = E(d(T_j, x_i)) = \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{k} d(T_j, x_i) P(T_j | x_i) \tag{3}$$

这样, 满足失真限度 D 的信息率失真函数 R(D) 就是:

$$R(D) = \min_{\{P(T_j|x_i)|\, d(T_j,x)\ FD,\ \sum_{j=1}^{k} P(T_j|x_i)=1\}} I(T, x) \tag{4}$$

$$I(T, x) = \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{k} P(T_j | x_i) \log \frac{P(T_j | x_i)}{\frac{1}{l}\sum_{i=1}^{l} P(T_j | x_i)} \tag{5}$$

记 d($T_j$, $x_i$)、P($T_j$| $x_i$)和 P($T_j$)为 $d_{ji}$、$P_{ji}$和$P_j$, 可用下面的迭代公式得到在失真限度 D 的约束下最优的指定概率 P($T_j$| $x_i$). 其中, K作为与失真限度D 相对应的解参数.

$$P_j(n) = \frac{1}{l}\sum_{i=1}^{l} P_{ji}(n) \tag{6}$$

$$P_{ji}(n+1) = \frac{P_j(n)\, e^{Kd_{ji}}}{\sum_{A=1}^{k} P_A(n)\, e^{Kd_{jA}}},\ j = 1, 2, \ , \ , k,\ i = 1, 2, \ , \ , l \tag{7}$$

迭代启动时可以任选 $P_{ji}$, 但一般对所有 i= 1, 2, , , l 的选择$P_{ji}(1) = 1/k$.

对于一个给定的 K 值, 当 x 中的所有字符$x_i$( i= 1, 2, , , l) 都根据上面求得的指定概率 P($T_j$| $x_i$) 指定给相应的层次树 $T_j$ 后, 接下来要根据指定重新生成层次树. 这相当于聚类算法中的重新计算聚类中心. 在生成层次树的过程中, 离散有记忆平稳信源模型的记忆长度通过编码技术中的最小描述长度原理来限制, 以生成更符合用户行为描述的模型[5].

在层次树生成过程中, 指定概率 P($T_j$| $x_i$)作为 $x_i$ 指定给 $T_j$ 的可信度度量, 记为权 $w_{ji}$. 用层次树对命令行字符序列进行编码的总描述长度由层次树的描述长度和数据的编码长度两部分组成. 设前面描述的层次树中单个节点的描述长度为 $L_{node}$, 则以上文序列为 s 根节点的子树 $T_j^s$ 的描述长度就是

$$Len(T_j^s) = L_{node} + \sum_{\{cs|\, c\, I\, E,\ cs\, I\, T_j\}} Len(T_j^{cs}) \tag{8}$$

其中, cs 是层次树 $T_j$ 的节点. 因而, 层次树 $T_j$ 的描述长度也就是 Len($T_j^<$). 用 $T_j$ 对 x 编码的总描述长度为:

$$TotalLen = Len(T_j^<) + \sum_{\{x_i|\, x_i\, I\, x\}} w_{ji} \# H(T_j^<) \tag{9}$$

$$H(T_j^s) = \sum_{\{c|\, c\, I\, 2,\ cs\, I\, T_j\}} P(c|s) \# H(T_j^{cs}) + \sum_{\{\not c|\, \not c\, I\, 2,\ \not cs|\, T_j\}} P(\not c|s) \# H(s) \tag{10}$$

$$P(c|s) = \frac{w(cs)}{\sum_{\{c_r|\, c_r\, I\, 2\}} w(c_r, s)},\ w(cs) = \sum_{\{|\, x_i|\, x_i = c,\ cs\, I\, x\}} w_{ji} \tag{11}$$

$$P(\not c|s) = \frac{w(\not cs)}{\sum_{\{c_r|\, c_r\, I\, 2\}} w(\not c, s)},\ w(\not cs) = \sum_{\{|\, x_i|\, x_i = \not c,\ \not cs\, I\, x\}} w_{ji} \tag{12}$$

$$H(s) = \sum_{\{c|\, c\, I\, 2\},\ sc\, I\, x} P(c|s) \log P(c|s) \tag{13}$$

$$P(c|s) = \frac{w(sc)}{\sum_{\{c_r|\, c_r\, I\, 2\}} w(sc_r)},\ w(sc) = \sum_{\{x_i|\, x_i = c,\ sc\, I\, x\}} w_{ji\,1} \tag{14}$$

为了避免使上面的序列划分和模型重建的求解结果陷入局部极小点, 我们以K作为解参数, 通过确定性退火算法[6] 来获得最优解. 算法首先从一棵层次树开始, 在需要时将其不断地分裂成两个不同的层次树( 对每个层次树先生成两个拷贝, 然后对两个拷贝中的每个节点的条件概率向量进行小的随机扰动而得到两个新的不同的层次树), 来寻找实际的层次树模型个数. 当层次树模型过多时, 有的层次树模型会由于不能获得足够大的指定概率而被删除. 因此, 算法具有自调节以获取实际混合信源模型的能力.

算法: 混合信源模型识别算法 IDM(Identification of Differ2 ent Models).

输入: 命令行字符序列 x= $x_1x_2$, $x_l$.

输出: 一组信源模型 T 和命令行字符序列的一个相应的划分.

```
(1) l= Len(x);
    /* 获得命令行字符序列的长度* /
(2) for (i= 0; i< l; i+ + )  w[0][i]= 1;
    /* 所有字符都指定给层次树模型 T0* /
(3) LearnTree(x, w, T);
    /* 根据上述指定生成层次树模型 T0* /
(4) N= 1; Lambda= K0; M= m;
    /* 设置初始模型数、解参数初值和停机模型数* /
(5) while (N[ M)  {
(6) SplitTree(T);
    /* 将每个层次树模型分为两个* /
(7) Delta= Assignment(x, T, w, Lambda);
```

[1] 对于一个子序列 s, 如果 $sx_i$( $x_is$) 是序列 x 的结束(起始) 于位置 i 的子序列, 则称 $sx_i$ I x($x_is$ I x).

```
    /* 序列划分, Delta = max{$P_j}_{j=1}^{N} = */
(8) while (Delta \ E)  {
(9) LearnTree(x, w, T);
    /* 层次树生成*/
(10) Delta = Assignment(x, T, w, Lambda);
(11)  }
(12) Trim(T, P)
    /* 删除指定概率 P(T_j)小于阈值的信源模型*/
(13) if (N \ |T|)  Lambda = Lambda - $K;
    /* 若模型数未增加则减小 K*/
(14) N = |T|;
    /* 重新设置信源模型*/
(15)  }
```

## 4　实验和结论

我们的实验平台为 PentiumIII 800MHz 微机, 操作系统为 RadHat Linux6l2, 历史命令文件为/home/user/. bash_history. 实验数据来自 4 个不同用户的历史命令文件. 每次随机选择 2 个不同用户的历史命令文件构成一条长度约为 10000 个字符的命令行字符序列(包括命令参数), 比较算法生成的层次树模型对各混合序列中自己子序列和非己子序列预测的平均失真程度 d. 实验结果如表 1 所示.

表 1　各混合序列生成的层次树模型对序列中自己子序列和非己子序列预测的失真程度 d

| 模　型 | 1- 2 | 1- 3 | 1- 4 | 2- 3 | 2- 4 | 3- 4 |
|---|---|---|---|---|---|---|
| 自己子序列 | 41- 38 | 45- 48 | 43- 39 | 43- 56 | 41- 41 | 51- 38 |
| 非己子序列 | 3172261 | 3612395 | 3472292 | 3372431 | 3282307 | 4172283 |

实验表明, 确定性退火算法可以对不同统计特性的字符子序列形成的混合序列进行划分, 并生成相应的层次树预测模型. 本文提出的利用确定性退火算法进行/ 伪装0 入侵行为检测的方法经实验测试是可行的. 该方法具有捕捉能代表不同用户行为统计特性的频繁短序列的能力; 由于该方法中存在着信源模型之间的竞争过程, 从而能够自动调节信源模型的个数. 对于算法的一些更深入的机理研究还需要作更细致的实验和分析工作.

参考文献:


[1] Matthias Schonlau, Martin Theus. Detecting masquerades in intrusion detection based on unpopular commands[J]. Information Processing Letters, 2000, 76: 33- 38.

[2] Jake Ryan, Meng2Jang Lin. Intrusion detection with neural networks [A]. In Proceedings of the 10th Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press, 1998.

[3] Buhlmann Wyner. Variable length markov chains[J]. Annual Statistics, 1999, 27: 480- 513.

[4] Bejerano Yona. Variations on probabilistic suffix tree: Statistical modeling and prediction of protein family[J]. Bioinformatics, 2001, 17: 23- 43.

[5] Barron Rissanen Yu. The minimum description length principle in coding and modeling[J]. IEEE Trans Information Theory, 1998, 44: 2743 - 2760.

[6] Rose. Deterministic annealing for clustering, compression, classification, regression and related optimization problems[J]. IEEE Trans Information Theory, 1998, 80: 2210- 2239.


作者简介:

赵俊忠　男, 1968 年生于河北省, 博士后, 研究方向为智能化网络安全, 研究内容为基于免疫机制的入侵检测系统, 感兴趣的领域还有进化计算、非线性理论和复杂适应系统.

黄厚宽　男, 1940 年生于四川省遂宁, 教授, 博士生导师, 研究方向为数据库知识发现、计算机免疫、多 Agent 系统和人工智能.