多态蠕虫特征码自动提取算法①

陈雪林

(绵阳师范学院 数学与计算机科学系, 绵阳 621000)

摘 要: 针对多态技术下变形蠕虫的特征与自动提取算法的问题, 提出一种多态蠕虫特征描述方法, 并给出特征码自动提取算法. 这种结合了 PADS 和 Polygraph 优点的 MS-PADS 特征提取方法, 能在强噪声下快速提取高质量的多态蠕虫特征, 具有低误报率、检测精度高和通用性好等特点.

关键词: 多态蠕虫; 特征提取; 内容过滤

Automatic Signature Generation Algorithm of Polymorphic Worm

CHEN Xue-Lin

(Department of Mathematics and Computer Science, Mianyang Normal University, Mianyang 621000, China)

Abstract: This paper researches the feature description and automatic signature generation algorithm of polymorphic worms. It proposes a new signature generation approach for polymorphic worms - MS-PADS (multiple separated string position-aware distribution signature), which integrate the advantage of PADS and Polygraph. It solved the problem of PADS signature width determination under strong background noise conditions. This method could generate high quality signatures of polymorphic worms with low false-positive rate and high detection precision rate.

Key words: polymorphic worms; signature generation; content-sifting

随着多态引擎技术[1]的出现, 使得变形蠕虫的行踪更加隐蔽, 检测更困难. 多态蠕虫特征码[2]的自动提取成为入侵检测系统的一个热点.

现有的特征自动提取研究中, 生成用于匹配的特征有基于模式特征、基于语义特征、基于漏洞特征等. 基于模式的特征, 即生成的特征是连续或非连续的子串或子序列. Honeycomb 是第一个自动提取特征的工具, 但不能提取 Polymorphic 蠕虫的特征. Polygraph 可以提取 Polymorphic 蠕虫的特征, 通过提取同类蠕虫的 “不变部分”并综合起来作为特征, 其提取的特征形式是最长公共子序列[3].

针对 PADS 和 Polygraph 的不足, 提出多态蠕虫特征片 MS-PADS(multiple separated string position-aware distribution signature)提取方法. 该方法利用 Polygraph 中的多序列特征码集合表征多态蠕虫, 避免单一 PADS 表示蠕虫特征特定性不足的缺点, 克服 Polygraph 中用固定多序列表示特征片段带来的容忍性不足缺点, 可广泛应用于其他病毒和攻击的特征提取以及检测.

1 PADS

PADS 继承了误用检测和异常检测的优点, 提出的蠕虫特征为二维特征矩阵[4]. 该特征相对于固定字符特征来说具有很高灵活性, 其每位的特征为可能出现字符的概率分布; 对比于 PAYL, 具有更高准确性.

设定 $\Theta=(f_0, f_1, f_2,.....f_w)$ 为关键词库 PADS 的特征形式, f_0 为正常流字节频率分布列向量, 在特征中只占一位, $(f_1, f_2,.....f_w)$ 为宽度为 w 的攻击特征频率分布列向量. 下面介绍 PADS 特征如何区分蠕虫流与正常流.

定义 1. 流 S_i 从 a_i 位置开始的匹配分数为:

$$\Lambda(\Theta, S_i, a_i, w)=\frac{M(\Theta, S_i, a_i)}{\overline{M}(\Theta, S_i, a_i)}=\prod_{p=1}^{w}\frac{f_p(S_{i,a_i+p-1})}{f_0(S_{i,a_i+p-1})} \quad (1)$$

① 收稿时间:2012-08-06;收到修改稿时间:2012-09-26

其中 $M(\Theta,S_i,a_i)=\prod_{p=1}^{w} f_p(S_{i,a_i+p-1})$，表示流从位置 a_i 开始与攻击特征样本间的匹配分数，其值为 w 个 $f_p(b)$ 值连乘，$\overline{M}(\Theta,S_i,a_i)=\prod_{p=1}^{w} f_0(S_{i,a_{i+p-1}})$，表示流从位置 a_i 开始与正常流的匹配分数. 定义 $\Lambda(\Theta,S_i,a_i)=M/\overline{M}$，则是用来反映该片段与正常流和攻击流的相似度.

定义 2. 重要区域段 ρ 是流中使匹配分数最大的 a_i. 重要区域段也就是特征序列在攻击流中起始位置.

定义 3. 流 S_i 匹配分数:

$$\Omega(\Theta,S_i,w)=\max_{a_{i=1}}^{lx-w+1}\sum_{p=1}^{w}\frac{1}{w}\log\frac{f_p(S_{i,a_{i+p-1}})}{f_0(S_{i,a_i+p-1})} \quad (2)$$

流匹配分数 Ω 作为正常流或怀疑流的判断根据: 若 Ω 大于 0, 即检测的流频率分布更接近于攻击流而偏离于正常流, 若其大于一个已设的阀值, 我们有理由怀疑其为蠕虫; 反之如果其小于 0, 则为正常流.

从对 PADS 特征描述中可知, 已知 PADS 的特征 Θ 可以求出流的重要区域段 ρ；已知所有流的重要区域 ρ，也就可求出 PADS 特征 Θ. 这可通过迭代算法 EM 或者 Gibbs 来解决. 由于 EM 算法可能收敛到局部最优而得不到最优解,使用模拟退火算法 Gibbs Sampling[5].

对于任意序列 Sx, 长度用 lx 表示, 则任意位置 ax $\in$[1…lx-W+1]都有可能是其特征重要区域段的起始位置 ρ. 根据定义 2, ax 作为 Sx 的特征区域段的起始位置的概率将与 $\Lambda(\Theta,Sx,ax)$成正比. 即有:

$$\Pr(a_x)=\frac{\Lambda(\Theta,S_x,a_x)}{\Sigma_{a_x=1}^{l_x-W+1}\Lambda(\Theta,S_x,a_x)}$$

因此, 其期望可描述如下:

$$E(R_x)=\sum_{a=1}^{l_x-W+1}R_x\times\Pr(a_x)=\sum_{a=1}^{l_x-W+1}R_x\times\frac{\Lambda(\Theta,S_x,a_x)}{\Sigma_{a_x=1}^{l_x-W+1}\Lambda(\Theta,S_x,a_x)}$$

算法执行: 对 Ri 分配一个随机的开始位置. 从 S 中随机选出 Sk, 并将它从 S 中减去, 在剩下的序列中进行计算求 Θ. 更新 Sk 的 Rk 的起始位置(依据不同初始位置的匹配分数得到的分配概率各不相同).

初始: 随机分配各个序列 S1、S2…到 Sn 的起始位置 a1、a2、到 an. 更新: 随机选择 S 中的 Sk, 基于集合 S-Sk 求签名表 Θ. 若平均的匹配分值在(1+ ε)时, 算法终止.

2 MS-PADS

2.1 单个特征序列长度判定方法

PADS 特征相对于固定字符串的特征其长度可有一定的灵活性, 但当应用于大型 IDS, 则为提高系统判决速度和精度, PADS 特征长度准确判断至为重要.

通过流的最小匹配分数与最大匹配分数的差值来粗略判断 PADS 特征宽度, 但未准确提出判定 PADS 特征宽度的方法. 若因流分类技术的误差而引入正常流噪声或其他类型的蠕虫样本, 怀疑流样本空间受到干扰时, 用 Gibbs 算法迭代出来的最小匹配值与最大匹配值不会随宽度减少趋于一致, 也不能作为粗略判定特征宽度的依据. 故引入平均匹配分数, 利用平均匹配分数的变化来判断特征宽度, 其定义为:

$$\overline{\Omega}(w)=\frac{1}{N}\sum_{i=1}^{N}\Omega(\Theta,S_i,w) \quad (3)$$

其中, N 为蠕虫怀疑流的个数. 利用平均匹配分数 $\overline{\Omega}(w)$ 来判定流的重要区域 ρ 后, 该估计量的准确性可以用怀疑流迭代后的 $E(R_i)$ 与实际重要区域 ρ 的近似程度来描述, 我们用逼近度 τ 来表示:

$$\tau=\frac{1}{N}\sum_{i=1}^{N}\frac{E(R_i)}{a_i} \quad (4)$$

当初始迭代选择特征宽度超过特征最大子串长度时, 由定义 3 流的匹配分数可看出, 迭代后的 $E(R_i)$ 越逼近实际重要区域 ρ 时候, 无论是向左还是向右逼近, 其 w 特征宽度内涵括的有效特征子串长度越长, 匹配值越大, τ 应与 $\overline{\Omega}(w)$ 成正比关系. τ 接近 1, 迭代计算产生的 $E(R_i)$ 即是该特征的重要区域.

图 1 反映的是在正常流中插入单一特征, 对其运用 Gibbs 算法迭代分析, 不同随机参数配置初始位置迭代后的 $\overline{\Omega}(w)$ 匹配分数与逼近度 ρ 的关系, 可见两者的变化趋势是一致的, 与我们的理论分析相吻合.

在判定了流的重要区域 ρ 后, 再确定 PADS 特征宽度 w. 当怀疑流库中带有正常流噪声后, 对于正常

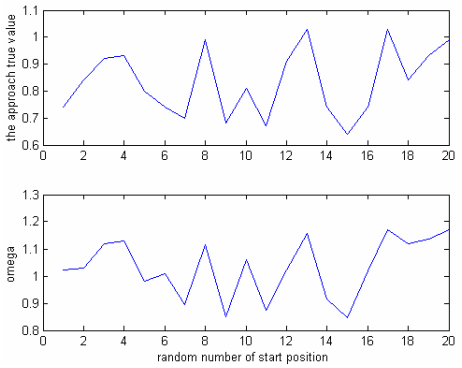

图 1　匹配值与逼近度关系

流，无论算法收敛到任何重要区域，其匹配分数都为一个较小的负值，与流库中最大匹配值相差很大，因此无法作为判定特征宽度的依据. 我们引入最佳特征宽度 w^*，利用 $\overline{\Omega}(w)$ 的变化情况来判定特征宽度 w. 若

$$\overline{\Omega}(w^*,\Theta)=(\frac{1}{n}\sum_{w=w^*}^{w^*-n}\overline{\Omega}(w,\Theta))(1\pm\varepsilon) \qquad (5)$$

则 w^* 为特征序列的最佳特征宽度，n 视 w^* 而定，一般取 w^*/3. ε 为一个很小的值，加入 ε 参数是为了让匹配分数有一定的波动. PADS 位特征 $f_p(b)$ 是基于统计概率的，利用公式(2)求匹配分数时会有一定的偏差.

在已知重要区域 ρ 情况下，若初始迭代宽度 w 超过实际蠕虫特征宽度，由定义 3 可知，PADS 特征中涵括的正常流部分对匹配分数值为负. 减少宽度时，其匹配值 Ω 应增大. 当迭代特征宽度小于或者等于最佳特征宽度 w^*，匹配值应近似保持不变，因为其是对 w 取均值. Ω 与随 w 减少的关系应先增大到趋于平稳.

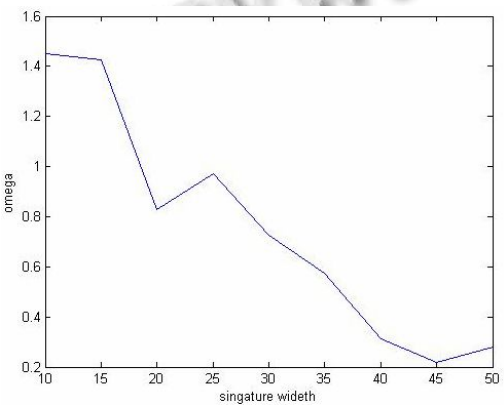

图 2　匹配值与特征宽度关系

图 2 所示为在正常流中插入特征宽度为 15 的特征时匹配分数与宽度关系，可看出理论分析与实验结果相吻合. 故当 $\overline{\Omega}(w)$ 满足(5)时，宽度为最佳特征宽度.

2.1 多态蠕虫 PADS 片段提取算法

多态蠕虫中作为其特征码的序列可分布在网络流的任意位置且长度不一. 对多态蠕虫而言，如何准确、高效、自适应地从怀疑流库中分别提取出长短不一致的特征码片段，是对多态蠕虫特征提取的挑战.

假设在多态蠕虫特征分布的最坏情况下，Gibbs 算法可能收敛到局部最优. 在初始化选取迭代开始位置时以多重随机参数配置开始，其不同的随机迭代初始位置可能收敛于流中不同特征重要区域 ρ，进而在得到的结果中寻找最优解. 由分析可知，在固定迭代宽度 w 时，$\overline{\Omega}(w,\Theta)$ 值大的随机参数配置迭代收敛时其包括的特征长度要长. 如迭代可能收敛于重要区域处，但收敛于其一选取匹配值 $\overline{\Omega}(w,\Theta)$ 要大，故其包括的有效特征长度要长. 基于以上分析，我们提出的分段求解 PADS 特征片段算法如下:

输入: 怀疑流库 M={S1,S2 ,S3….}; 输出: 一系列 PADS 特征库; 初始化: 设置初始的迭代宽度，选取 K 组多重随机参数配置作为初始迭代开始位置.

1) 应用 Gibbs 算法，求出每组匹配分数 $\overline{\Omega}_w(i)$，对于匹配分数最大的那组，保留其迭代后的重要区域段集 ρ_i 和匹配分数 $\overline{\Omega}_w(i)$;

2) 减少宽度 w，ρ_i 作为迭代的新初始输入，求出收敛时匹配分数 $\overline{\Omega}_{w-1}$，若 w <2，退出;

3) 若 $\overline{\Omega}_{w-1}$ 满足定义 5，保留 $w^*=w$ -1，在流中裁剪掉从迭代收敛的重要区域段开始的长度为 w^* 的那部分流，提取出的 PADS 特征加入特征库，进入 1)，其开始迭代的宽度为 w^*，反之重复 2).

算法中先求特征宽度的最大特征子串，下次迭代时初始宽度从该宽度开始. 由于将前次迭代求得的重要区域宽度作为下次求解的初始值，压缩了搜索空间，减少了迭代次数，可极大地提高算法效率. 由于迭代后的重要区域段更接近于实际特征位置，对于 “缺失数据问题”的求解可避免 Gibbs 算法陷于局部最优.

3　实验分析

选取 Code-red II 蠕虫作为测试用例，用病毒变形引擎 Clet 和 ADMmutate 进行变形处理.

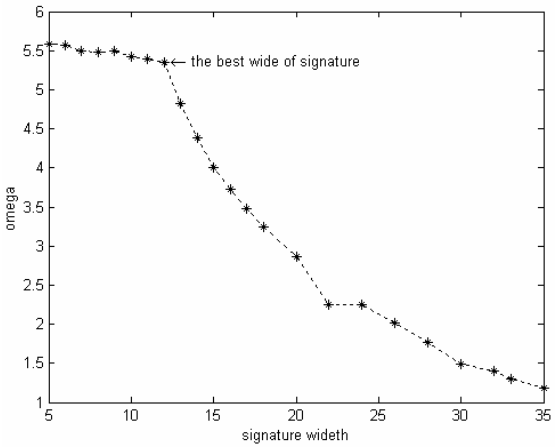

图 3 宽度为 12 的特征提取过程

按照 MS-PADS 提取算法, 先选择 30 组多重随机参数作为迭代开始的初始位置, 宽度初始设置为 35byte, 在 30 组初始选择的随机开始位置中选取匹配分数最高的第 17 组作为最优的迭代结果, 再不断减少特征宽度. 从图 3 可以看出, 实验提取了 code-red II 中最大部分的序列, 提取每个位置出现概率最大值还原为 http/1.0\r\n; 当在流中把该部分去掉后继续运行该算法, 得到图 4 所示提取过程, 其提取的特征部分还原出来为.ida?, 流中余下部分的特征提取过程类似.

经过实验评估, 采用 MS-PADS 提取出的 PADS 特征码具有高检测精度及适用范围广等特点.

4 结语

目前在多态蠕虫的特征码提取研究中, PADS 在容忍性和检测新攻击方面优势突出. 本文针对单一 PADS 特征表征多态蠕虫特定性不足, 探讨了多态蠕虫特征码的自动提取算法, 分析了 PADS 特征片段长度判定方法. 该算法能在高噪声情况下产生良好的特征, 提取速度快, 具有很强的容错能力, 可应用于病毒或攻击特征的提取过程, 具有广泛应用前景.

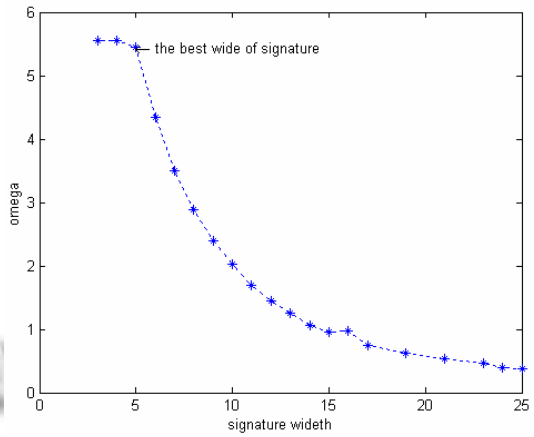

图 4 宽度为 5 的特征提取过程

参考文献

1 盛津芳,谭云桥,王斌.网络攻击中多态变形技术分析及其对抗策略.计算机安全,2007,1:11–13.

2 祝仰金,秦拯.Zero_day 多态蠕虫特征自动提取技术研究.微计算机信息,2011,27(1):198–200.

3 赵旭,何聚厚.Polymorphic 蠕虫特征自动提取算法及检测技术研究.计算机工程与应用,2008,44(36):106–108.

4 Tang Y, Chen SG. An Automated Signature-Based Approach against Polymorphic Internet Worms. IEEE Trans. on Parallel & Distributed Systems, 2007,18(7): 879 –892.

5 Hideo K, Genya K. Gibbs sampling methods for Bayesian quantile regression. Journal of Statistical Computation & Simulation, 2011,81(11):1565–1578.

(上接第 132 页)

12 王雯娟,黄振杰,郝艳华.一个高效的基于证书数字签名方案.计算机工程与应用,2011,47(6):89–92.

13 Li J, Huang X, Zhang Y, et al. An Efficient Short Certificate-Based Signature Scheme. The Journal of Systems and Software, 2012,85:314–322.

14 Liu J, Baek J, Susilo W, et al. Short and Efficient Certificate-Based Signature. Networking 2011 Workshops, LNCS6827, Berlin: Springer-Verlag, 2011: 167–168.

15 Cheng L, Xiao Y, Wang G. Cryptanalysis of a Certificate-Based on Signature Scheme. Procedia Engineering, 2012, 29:2821–2825.

16 李志敏,徐馨,李存华.高效的基于证书数字签名设计方案.计算机应用研究,2012,29(4):1430–1433,1444.

17 Barreto P, Kim H, Lynn B, et al. Efficient Algorithms for Pairing Based Cryptosystems. Crypto 2002. LNCS 2442, Berlin: Springer-Verlag, 2002: 354–368.