

基于改进免疫遗传算法的网络流量分析算法

李超建, 李露璐, 谢 妙

(玉林师范学院 计算机科学与工程学院, 广西 玉林 53700)

摘 要: 文运用免疫遗传算法提出一种入侵检测流量关联规则提取方法, 同时运用优化后的算法建立的入侵检测模型对网络流量进行检测规则的提取, 提高了检测器对流量分析的效率. 实验结果表明, 该算法能高效地解决入侵检测中流量分析关联规则提取问题, 并且其效率优于传统算法.

关键词: 免疫遗传算法; 网络流量; 网络安全

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2013)12-0164-03

Algorithm on Analysis of Networks Traffic Based on Improving IGA

LI Chao-jian, LI Lu-lu, XIE Miao

(Yulin Normal College, Computer Science and Engineering Institute, Yulin 53700, China)

Abstract: The paper first present a solution on obtain association rule of networks traffic detection based on immune genetic Algorithm, and apply optimizing algorithm to build intrusion detection model improve the networks traffic analysis show more efficiency in Instruction detection. The experimental results prove that the algorithm has more performance in IDS analyses traffic association rule mining problem than traditional algorithm.

Key words: IGA; Network Traffic; Network Security

1 引言

入侵检测系统(Intrusion Detection System, IDS)的主要工作是监测计算机系统中是否有未经授权的用户操作计算机以及具有授权的用户滥用权限的操作^[1]. 入侵检测是从大量的非线性的流量数据中抽取出潜在的、线性的、可预测的有用信息、模式和趋势的. 当前网络中的流量数据的非常庞大, 网络流量分析及测定的算法常常成为入侵检测系统的瓶颈, 并且系统的性能也远远不能满足实际需要. 因此, 重新建立流量安全分析策略为解决这类问题提供了行之有效的途径, 它能够从繁杂流量信息中自动发现隐藏在数据中的模式信息, 从而做出预测性分析, 是当前入侵检测系统必须具备的能力.

入侵检测系统围绕着从海量的流量数据分析和

提取的关联规则并对其线性化, Wenke. lee^[2]把数据挖掘方法引入到入侵检测中, 通过数据挖掘程序处理搜集到的审计数据, 为各种入侵行为和正常操作建立精确的行为模式. 数据挖掘方法的关键点在于算法(分类算法、关联算法和序列分析)的选取和建立一个正确的体系结构. 常用的算法有 Apriori、FP-Growth^[3]等数据挖掘算法的缺点是计算复杂度和计算量非常大, 而通常入侵检测系统的服务器都是集中化的, 因此算法的效率问题成为其运行的瓶颈. 基于隐马尔可夫模型(HMM)^[4]入侵检测方法认为一个完全正常的系统调用, 也可以是危险的. 例如: 缓冲区溢出式攻击的程序, 拒绝访问攻击等. 它所产生的系统调用事件和在正常情况下产生的有着明显的不同. 因此, 可以通过构建正常情况下系统调用事件模型, 然后观察是否与此模型有明显的偏离, 以此来有效地检测入侵的产生. 隐马尔可夫模型

收稿日期: 2013-03-17; 修回日期: 2013-05-22

基金项目: 广西自然科学基金项目(2013GXNSFAA019337); 广西教育厅重点科研项目(2013ZD056); 广西教育厅研究项目(201010LX427); 广西教育厅科研项目(2013YB202); 玉林师范学院专项科研项目(2012YJZX04)

(HMM)是对观察到的符号序列构造模型的一种非常好的工具,在构造系统调用事件模型上有着比其它方法更好的性能,但它在构造正常行为模型时需要较长的时间.解决办法是提高计算机系统的性能,或者减少训练的数据.而基于人工免疫算法的入侵检测系统,在规则提取时具有很好的鲁棒性和隐含并行性^[5-10],不仅搜索精度高,而且具有收敛速度快,在解决大空间、非线性、全局寻优等复杂问题时,有明显的优势.基于人工免疫的异常入侵检测系统具有学习的能力,它可以紧密地模仿用户的行为并且根据最近的变化进行调整.它的另外一个特性就是允许模糊数据或噪音数据.此外,与统计理论相比,人工免疫更好地表达了变量之间的非线性关系.

本文借鉴免疫遗传算法的思想,并对遗传算法进行改进后运用到入侵检测流量分析的关联规则提取中,采用免疫克隆算法进行提取规则,同时对规则提取量的度量机制直行优化,加快了算法对规则的收敛速度,以此来提高入侵检测对流量分析的效率.

2 改进的流量分析算法

2.1 数据的定义

本文建立的入侵检测系统主要是对流量分析.通常系统日志中记录着包括 TCP 包、UDP 包、IP 包、IMCP 包等协议的数据.系统将提取出日志中大部分信息特征值如表 1 所示.

表 1 流量信息表

包(Packets)	特征值(Values)
IP Id Number	0-65535
IP TTL	0-255
IP Address Flag	0-65535
IP Source	Valid IP address
IP Destination	Valid IP address
UDP Source Port	0-65535
UDP Destination Port	0-65535
UDP Sequence Num	0-4294967295
Individual UDP Flags	0/1
.....	

流量信息包含了大量的网络包特征数据,如 IP 地址、端口、访问的信息大小,UDP 源端口,UDP 目标端口等.流量存放于日志文件中,通常以通用日志格式 CLF(Common Log Format)和扩展通用日志格式 ECLF(Extended Common Log Format)进行存放.因此,进行规则提取之前必须进行流量日志内容的预处理.

定义 1 本文主要监测网络环境中对流量日志数据扫描结果表示如下,其中将产生的流量的会话集表示为事件向量 $\alpha: (\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n)$,其中 α_1 为一个三元组,表示为 $\{T, 0, \omega\}$:

- (1) T 为时间属性,记为 $T(\alpha_1)$,表示事件 α_1 发生的时间;
- (2) O 表示为宿主属性,记为 $O(\alpha_1)$,表示事件 α_1 的发起宿主;
- (3) ω 存放发生的会话内容.

规定所有会话都是按时间排序的,即对所有的 $i \geq 1, T(\alpha_i) < T(\alpha_{i+1})$.

定义 2 定义 1 流量日志中的每一个事件中 $w(p_i, u)$ 为 $\alpha_i (i = 1, 2, \cdots, n)$ 事件上的权重.

定义 3 定义事件 α_i 的支持度:设第 $(0 \leq j \leq N)$ 个属性最易导致第 $(0 \leq k \leq M)$ 类事件的加权事件为 O_{ij} ,则记:

$$k | \sum_{i=0}^2 W_i O_{ik} = \max(\sum_{i=0}^2 W_i O_{ij})$$

定义 4 定义规则的可信度为 S :
 S 在事件 T 中的信息增益 G 为

$$G(T, S) = O(T) - \sum_{Support(s_i)} O(T_i) \log \frac{T_i}{T}$$

式中, $Support(s_i)$ 是 S 属性对事件 T 的所有支持序列集合, T_i 是当属性 S 在 T 的子序列 T_j .

2.2 流量分析算法

2.2.1 克隆算法

首先利用克隆算法对预处理的数据进行变异和交叉等操作实现日志中事件的信息交换,而且还要充分利用事件在变化过程中已经获得的规则匹配的特性,进一步增加规则提取的多样性^[11-12].

算法 1:克隆算法

$$A(k) \xrightarrow{\text{clone}} \bar{A}(k) \xrightarrow{\text{mutation}} \bar{A}^T(k) \xrightarrow{\text{select}} A(k+1);$$

Clone:
 $T_c^c(\bar{A}(k)) = [T_c^c(A_1(k)), T_c^c(A_2(k)), \cdots, T_c^c(A_n(k))]$

2.2.2 亲和度算法的改进

危险模式算法的重要模块是亲和度函数,亲和度的评价模型会直接影响到整个算法的收敛速度,如果模型建立不当甚至导致算法失败.由算法 1 式得:

- (1) 抗体 A_i 的 q_i 代克隆是
 $T_c^c(A_i(k)) = I_i \times A_i(k), i = 1, 2, \cdots, n;$
- (2) 迭代代数函数:
 $q_i(k) = g(N_c, f(A_i(k)));$

(3) 亲和力表示为: $f(A_i(k))$;

(4) 克隆样本变异操作 T_s^c ;

(5) $\forall i = 1, 2, \dots, n$, 存在变异后的抗体

$$B = \{A'_{ij}(k) \mid \max f(A'_{ij}), j = 1, 2, \dots, q_i - 1\}$$

则 B 取代 $A_i(k) \in \bar{A}(k)$ 的概率为

$$p_s = \begin{cases} 1 \\ \exp \left[-\frac{f(A_i(k)) - f(B)}{\alpha} \right] \\ 0 \end{cases}$$

按以上公式迭代条件循环运行, 当到达满足条件时, 得到问题的解。

3 仿真实验

实验所用数据来自于某高校的网络流量日志 (www.lzzy.net), 其中包含约 100 万条日志记录。日志经过本系统的预处理, 匹配得到 2 315 个向量事件。实验所用的硬件平台基本配置为 Intel Xeon E5620、16 GB 内存, 软件平台为 Windows server 2008。算法运行了 100 次, 得到的关联规则提取的数量与 Apriori 算法相比较的结果, 见图 1。图 1 显示的是入侵检测系统运行本文的算法, 算法运行了 20 次, 得到的平均结果。从图 1 中可以看出, IGA 算法能有效地提取出流量的关联规则, 并且有较好地收敛速度。

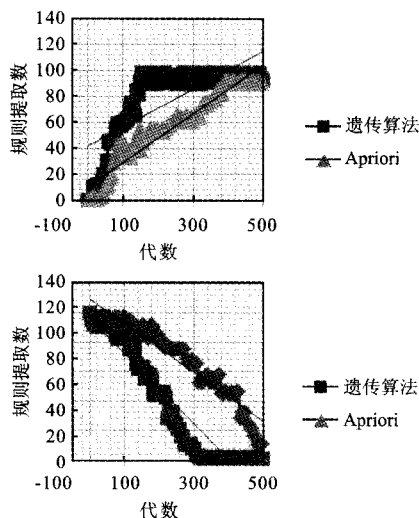


图 1 两种算法的比较

4 结束语

本文采用免疫遗传算法对页面进行了入侵检测流量分析, 并就提取出的规则的效率和传统的

Apriori 算法进行了对比, 实验结果表明该算法用于入侵检测流量数据规则提取所得出结果的正确率和效率都优于传统方法。

参考文献:

- [1] Biswanath Mukherjee, Todd L Heberlein, Karl N Levitt. Network intrusion detection[J]. IEEE Network, 1994, 8(3): 26-41.
- [2] W Lee, S J Stolfo. Data mining approaches for intrusion detection[C]//Proceedings of the 7th USENIX Security Symposium, Oakland, California; IEEE, 1998.
- [3] Fox K L, Henning R R, Reed J H, et al. A neural network approach toward intrusion detection[C]//Proceeding of the 10th National Computer Security Conference. [s. l.]: IEEE, 1990: 116-174.
- [4] L R Rabiner and B H Juang. An introduction to hidden Markov models[J]. IEEE ASSP Magazine, 1986: 4-16.
- [5] 鲁云平. 基于免疫原理的网络入侵检测技术研究[D]. 重庆: 重庆大学, 2003.
- [6] Matzinger P. Tolerance, danger, and the extended family[J]. Ann Rev Immunol, 1994(12): 991-1046.
- [7] Matzinger P. Essay on the danger model in its historical context[J]. Scand J Immunol, 2001(4): 4-9.
- [8] Yingfeng Chen, Lianying Zhou. An innovative IDS immune system model proceedings of the international conference on systems and cybernetics. Hague: IEEE, 2004: 4810-4814.
- [9] Xue-Yuan Yang, Wan-Wu Zhou, PingWei. Hybrid NIDS based on biological immunology[C]// Proceedings of 2004 International Conference on Machine Learning and Cybernetics. [s. l.]: IEEE, 2004: 2889-2892.
- [10] 王慧. 危险模式入侵检测报警算法优化[J]. 煤炭技术, 2010(10): 175-179.
- [11] 许国艳, 史宇清. 遗传算法在关联规则挖掘中的应用[J]. 计算机工程, 2002, 23(7): 122-124.
- [12] 贾兆红, 倪志伟, 赵鹏. 改进型遗传算法及其在数据挖掘中的应用[J]. 计算机应用, 2002, 22(9): 31-33.

作者简介:

李超建 男, (1979—), 硕士, 讲师. 研究方向为计算机网络、嵌入式技术.

李露璐 女, (1981—), 硕士, 讲师. 研究方向为计算机网络、网络信息安全.

谢妙女, (1976—), 硕士, 副教授. 研究方向为无线传感器网络、数据隐私保护.