

高可用性网络入侵预警方法的改进研究

龚健虎

(广东培正学院 计算机科学与工程系, 广东 广州 510800)

摘 要: 高可用性网络通常会受到海量恶意入侵, 当前预警方法只能对历史与当前的网络入侵情况进行分析, 无法预测未来高可用性网络的入侵情况, 为了提高预警精度, 提出一种新的高可用性网络入侵预警改进方法, 介绍了隐马尔可夫模型, 在此基础上, 构建依据 HMM 模型的实时高可用性网络入侵预警模型, 依据危险理论对评估结果进行分析, 将高可用性网络状态划分成四个等级, 通过构建的 HMM-NSSP 模型实现网络入侵预警. 为了提高 HMM-NSSP 模型的预警精度, 引入最大熵法对高可用性网络安全态势进行一致性判定, 完成高可用性网络入侵预警方法的改进, 实验结果表明, 所提方法预警结果和实际结果基本一致, 具有很高的预警精度, 可用性很高.

关键词: 高可用性网络; 入侵; 预警

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)08-0134-05

High Availability Network Intrusion Early Warning Method Improvement Research

GONG Jian-hu

(Computer Science and Engineering Department, Guangdong Peizheng College, Guangzhou 510800, China)

Abstract: High availability network usually are a huge number of malicious invasion of current early warning method can only analyze the history and the current network intrusion status, cannot predict the future high availability network invasion situation, in order to improve the accuracy of early warning, put forward a new high availability network intrusion early warning method, this paper introduces the hidden markov model, on this basis, the building based on the HMM model of real-time high availability network intrusion early warning model, according to the analysis of risk theory to the evaluation result, the high availability network state was divided into four grades, by building the HMM - NSSP model to realize the network intrusion early warning. In order to improve the early warning of the HMM - NSSP model precision, the introduction of the maximum entropy method for high availability network security situational consistency judgement, to complete the high availability of network intrusion early warning method improvement, the experimental results show that the proposed method the early warning results and actual results are basically identical, has the very high accuracy of early warning, a high availability.

Key words: high availability network; Invasion; warning

1 引言

高可用性网络是一种可以提高网络带宽、提升网络性能的交流式新兴网络, 是高速局部网络发展的核心和关键. 因此保证高可用性网络的安全性非常关键^[1,2]. 众所周知, 防火墙是信息安全防护系统中的一种, 在计算机的发展史上起着举足轻重的作

用^[3,4]. 然而在高可用性网络技术飞速发展的今天, 随着高可用性网络攻击技术的不断革新, 防火墙渐渐无法防御攻击技术中的“高科技”^[5]. 于是, 高可用性网络入侵预警方法应运而生, 在网络被入侵之前采取防御措施, 使损失降到最低^[6].

文献[7]提出一种基于攻击树与贝叶斯网络的高可用性网络入侵预警方法, 该方法依据报警信息

对入侵者的入侵想法进行识别,通过贝叶斯网络对入侵进行预测,从而实现网络入侵预警.该方法实现过程简单,但需依据报警信息预先塑造贝叶斯网络,导致预警结果不准确.文献[8]依据对隐含着色 Petri 网(HCPN)的研究提出一种基于 HCPN 的高可用性网络入侵预警方法,该方法通过代理得到网络资源,依据观察值,通过报警信息实现高可用性网络入侵预警.该方法预警效率高,但观察值因易受到外界环境的干扰并不可靠,导致预警结果不可靠;文献[9]提出一种基于攻击响应事件相关性的高可用性网络入侵预警方法,根据攻击响应时间中、前驱后继关系预测未来一段时间内同一网络地址可能遭受的攻击.但实际上,前驱后继关系因为网络复杂性等原因往往不容易构造,很难得到准确的预警结果;文献[10]提出一种基于回归方程的高可用性网络入侵预警方法,构建回归模型,根据以往事件的发生规律预测中长期时段内网络遭受攻击的可能性.但该方法只适用时间较长的预测,实时性不高.

针对上述方法的弊端,提出一种新的高可用性网络入侵预警方法,介绍了隐马尔可夫模型,构建依据 HMM 模型的实时高可用性网络入侵预警模型,从而实现网络的入侵预警.为了提高 HMM-NSSP 模型的预警精度,引入最大熵法对网络安全态势进行一致性判定.实验结果表明,所提方法具有很高的预警精度,可用性很高.

2 高可用性网络入侵预警方法改进研究

2.1 隐马尔可夫模型

隐马尔可夫模型(HMM)实质上是一种表示含有隐含未知参数的马尔可夫过程.该模型从 1980 年以来就在语音识别、计算机文字识别、移动通信核心技术、生物信息科学、故障诊断等领域被广泛应用.

HMM 模型包含双重随机过程,即隐马尔可夫链和显示随机函数集.其中隐马尔可夫链包含一定状态数. HMM 模型观察序列可以表示为 $O = (o_1, o_2, \dots, o_r)$, 则 HMM 模型可通过五元组 $\lambda = (N, M, \pi, A, B)$ 完成对系统的描述. 其中, N 代表隐状态数量, 记作 $S = (S_1, S_2, \dots, S_N)$; 每个隐状态 S_i 包含 M 个可观测状态 V_i , 记作 $v = (v_1, v_2, \dots, v_M)$; π 代表 $1 \times N$ 阶初始概率分布矩阵, 用来描述可观测序列 $O = (O_1, O_2, \dots, O_r)$ 在 $t = 1$ 时刻时, 隐状态 q_1 的初始状态概率分布. 假定初始概率分布矩阵用 $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ 来表示, 那么有:

$$(1) \pi_i = P(q_1 = S_i), 1 \leq i \leq N;$$

(2) $A = (a_{ij})_{N \times N}$ 是马尔可夫链的状态概率转移矩阵, 当 HMM 为一阶时, 有:

$$(a_{ij}) = P(q_{t+1} = S_j | q_t = S_i), \sum_{j=1}^N a_{ij} = 1; \\ 1 \leq i \leq N; 1 \leq j \leq N; \quad (1)$$

(3) $B = (b_{jm})_{N \times M}$ 表示符号概率观测矩阵, 公式描述如下:

$$b_{jm} = P(O_t = v_m | q_t = S_j); \\ 1 \leq j \leq N; 1 \leq m \leq M; \quad (2)$$

HMM 中矩阵 A 被称作随机矩阵, 是因为该矩阵每行之和均等于 1, 也就是从一个状态转移到任何一个隐状态的概率之和为 1, 随机矩阵 B 与矩阵 A 有着相似的特点.

2.2 基于 HMM 模型的实时高可用性网络入侵预警模型

HMM 模型依据观察序列计算概率, 从而实现高可用性网络状态预测. 本文利用 HMM 模型对实时高可用性网络入侵预警模型(HMM-NSSP) 是进行构建, 详细实现过程描述如下:

假设 APS 用来表述高可用性网络入侵强度等级序列; SE 代表高可用性网络安全状态集合, $SE = \{S_1, S_2, \dots, S_N\}$; APE 表示高可用性网络入侵强度等级集合, $APE = \{V_1, V_2, \dots, V_M\}$; λ 代表 HMM-NSSP 模型的转移参数, $\lambda = \{\text{Trans}, \text{Obs}, \text{Init}\}$.

HMM-NSSP 模型中高可用性网络安全状态初始分布概率记作 $\text{Init} = \{\pi_j\}$, 其中:

$$\pi_j = P(q_o = S_j), 1 \leq j \leq N, S_j \in SE \quad (3)$$

HMM-NSSP 模型中高可用性网络安全状态转移概率矩阵记作 $\text{Trans} = \{a_{ij}\}$; 高可用性网络入侵强度等级的概率分布记作 $\text{Obs} = \{b_j(k)\}$, $b_j(k)$ 用于描述模型 t 时刻处于状态 S_j 时观察到攻击强度等级 $v_t = V_k$ 的概率, 其中:

$$b_j(k) = P(v_t = V_k | q_t = S_j) \quad (4)$$

根据上述描述的参数, 对实时高可用性网络入侵预警进行建模, HMM-NSSP 模型可用一个六元组进行描述, 利用式(5) 进行表述:

$$\text{HMM-NSSP} = \langle \text{APS}, \text{SE}, \text{APE}, \lambda, \text{Init}, \text{Trans} \rangle \quad (5)$$

2.3 高可用性网络入侵预警方法研究

HMM-NSSP 模型利用高可用性网络安全态势评估结果实现入侵预警, 评估结果准确程度对预警

结果非常关键. 本节依据危险理论对评估结果进行分析, 将高可用性网络状态划分成四个等级, 通过构建 HMM-NSSP 模型实现网络入侵预警.

本节将高可用性网络入侵序列设为可见观察序列, 因为高可用性网络主机状态受到攻击序列影响不可直接被观察, 所以把它设置成隐藏状态. 先设定未来某时刻的入侵强度等级, 计算出该强度下概率最大的主机状态, 从而预测不同入侵强度概率下高可用性网络安全态势预测值, 若该值低于既定阈值, 则认为达到预警值, 需对高可用性网络进行检测.

算法中涉及的参数主要有入侵强度等级序列 APS 和概率转移参数 λ . 首先假定模型的状态空间为 $S = \{\text{Safe}, \text{Probed}, \text{Attacked}, \text{Danger}\}$, 观察序列号为 $V = \{\text{high}, \text{medium}, \text{low}\}$, $T+1$ 时刻高可用性网络入侵强度等级为 O_{T+1} , 那么预测入侵强度序列可利用式(6)表示为

$$\text{APS}' = \text{APS} \cup \{O_{T+1}\} \quad (6)$$

找出 APS' 序列出现的概率最大时对应的网络安全状态. 定义 $a_t(i) = P(O_1, O_2, \dots, O_t = S_i | \lambda)$, 然后对式(7)进行求解:

$$\begin{cases} a_{t+1}(j) = [\sum_{i=1}^N a_t(i) a_{ij}] b_j(O_{t+1}), \\ a_1(i) = \pi_i b(O_1) \end{cases} \quad (7)$$

式中, $\pi_i \in \text{Init}$, $a_{ij} \in \text{Trans}$, $b_j(O_{t+1}) \in \text{Obs}$.

通过式(8)求出 $T+1$ 时刻高可用性网络处于 S_i 状态的概率; 通过式(9)标出概率最大的状态 q_k^* :

$$\begin{aligned} \varphi_{T+1}(i) &= P(\lambda q_{T+1} = S_i | \text{APS}', \lambda) \\ &= \alpha_{T+1}(i) / \sum_{i=1}^N \alpha_{T+1}(i) \end{aligned} \quad (8)$$

$$q_k^* = \arg \max_{1 \leq i \leq N} [\varphi_{T+1}(i)] \quad (9)$$

式中, $\alpha_{T+1}(i)$ 用于描述前向式; q_k^* 代表 $T+1$ 时刻下观察高可用性网络入侵等级 $O_{T+1} = V_k$ 时的主机安全状态, 此时得到的 $\varphi_{T+1}(i)$ 概率为最大值.

将式(8)和(9)带入, 根据式(10)求出主机安全态势预测值 se , 根据式(11)求出整个高可用性网络安全态势预测值 $L^*(t)$:

$$se = 1 - \frac{\varphi_{T+1}(i)}{1 + \ln(\sum_{k=1}^M C(q_k^*) \cdot b^*(k) + 1)} \quad (10)$$

$$L^*(t) = \sum_{h=1}^d \bar{\omega}_h \cdot se_h(t) \times q_k^* + \varphi_{T+1}(i) \quad (11)$$

式中, $C(\cdot)$ 用于描述不同安全状态下的危险代价矩阵; $b^*(k) = P(v_{T+1} = V_k | q_k^*)$ 用于描述 $T+1$ 时

刻出现入侵强度等级 V_k 的概率; $se_h(t)$ 用于描述节点 h 在 t 时刻的安全态势预测值.

通过 HMM-NSSP 模型实现高可用性网络入侵预警的详细过程如下所示.

输入: 高可用性网络入侵强度等级序列 APS 和概率转移参数 λ ;

输出: 高可用性网络安全态势预测值 $L^*(t)$;

(1) 观察 APS 和 APS' , 利用前向算法求出新的高可用性网络入侵强度序列出现的概率, 对各种时刻的 $\alpha_t(i)$ 进行计算;

(2) 求出新入侵强度下, 高可用性网络处于状态 S_i 的概率;

(3) 找出不同攻击强度等级序列概率最大的状态;

(4) 计算出所有节点安全态势预测值;

(5) 通过加权平均法对网络安全态势预测值进行计算, 将计算值与既定阈值进行比较, 若低于既定阈值, 则需报警.

2.4 高可用性网络入侵预警方法的改进

为了提高 HMM-NSSP 模型的预警精度, 可引入最大熵法对高可用性网络安全态势进行一致性判定, 从而实现对它的改进. 选择满足已知条件且熵值最高的概率分布, 也就是依据已知网络信息此刻态势, 对符合条件的合理分布进行预测, 得到满足已知知识库特征的推理结果, 详细实现过程如下所示.

参照最大熵理论, 将高可用性网络安全态势特性方式表达定义为二值特征 $f_i(x, y)$, x 代表获取信息值, y 代表期望值; 如果 $f_i(x, y)$ 能够对模型产生影响, 则利用式(12)建立一个可以产生训练样本 $\tilde{p} = (x, y)$ 的约束模型:

$$P_E = \{p_e | E_p(f_i) = E_{\tilde{p}}(f_i), 1 \leq i \leq k\} \quad (12)$$

式中, E 代表最大熵.

在约束模型的约束范围内, 利用式(13)对高可用性网络安全态势特征进行提取:

$$\omega_i = \frac{\text{rate}_i - \text{save}}{\sum_{i=1}^j (C \cdot \gamma)} \times \varphi_a \quad (13)$$

式中, save 代表提取网络安全系数的权重; j 代表提取安全态势特征的累积方差; C 代表特征选择参数; φ_a 代表样本数量; γ 代表归一化处理后特征矢量.

将提取特征和高可用性网络信息, 利用式(14)进行匹配, 实现安全态势的一致性评定, 大大增强了 HMM-NSSP 模型的预警精度:

$$p = \frac{X_j(n) - W^T(n)}{\omega_i \times X_j} \tag{14}$$

式中, X_j 代表提取网络安全态势特征的样本数量; n 代表特征信息的条件熵; W^T 代表潜在特征选择的阈值。

3 实例分析

3.1 高可用性网络安全态势数据

将某高可用性网络 2008 年 7 月 5 日—24 日监测数据作为实验样本数据,以验证本文方法的有效性。一天采样 2 次,共得到 40 个高可用性网络安全态势值,用图 1 进行描述。将前 30 个高可用性网络安全态势值当成训练样本,将后 10 个高可用性网络安全态势值当成测试样本,实验将 K-均值 RBF 方法和 SVM 作为对比,在 Matlab7. 0 环境下进行。本文安全态势阈值取 0. 6,低于 0. 6 则需进行预警。

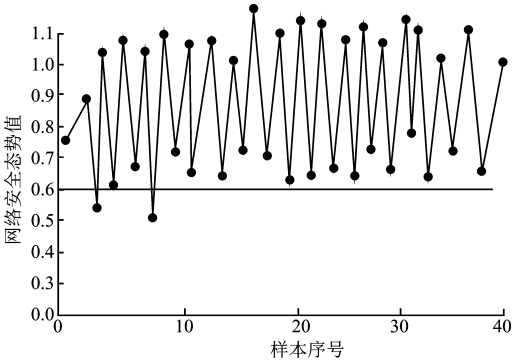


图 1 高可用性网络安全态势数据

3.2 预警结果

图 2 描述的是采用本文方法、K-均值 RBF 方法和 SVM 方法对 2008 年 7 月 24 日至 2008 年 8 月 6 日高可用性网络安全态势值的预测结果。

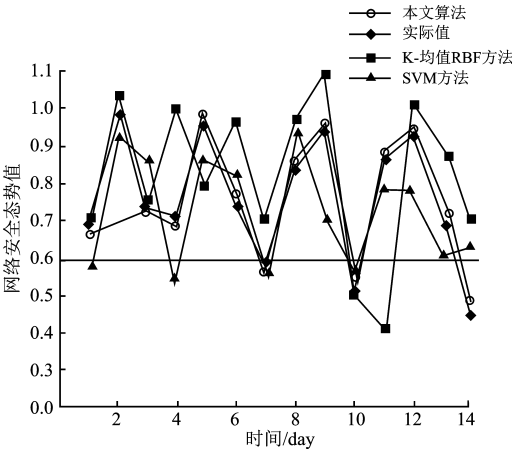


图 2 三种方法网络安全态势值预测结果

分析图 2 中曲线可知,实际上,在 2008 年 7 月

24 日至 2008 年 8 月 6 日,网络安全态势值有三次低于阈值 0. 6,也就是说需要进行三次预警,本文方法预测结果和实际结果相同,也进行了三次预警,且预警结果全部正确;K-均值 RBF 方法的预警次数是 2 次,但有一次误预警,两次没有预警,只有一次准确预警;SVM 方法的预警次数是 4 次,有两次误预警,一次没有预警,两次准确预警。整体看来,本文方法的预警结果优于 K-均值 RBF 方法和 SVM 方法,这是因为它能够遵循实际的高可用性网络环境自适应拟合高可用性网络安全数据之间的非线性映射关系,数据中心包含了更多的高可用性网络安全态势信息。

3.3 预警准确率分析

下面引入预警绝对误差来定量比较本文方法、K-均值 RBF 方法和 SVM 方法的准确性,绝对误差就是计算值和实际值的差值,绝对误差越低,预警准确性越高。三种方法预警绝对误差比较结果见图 3。

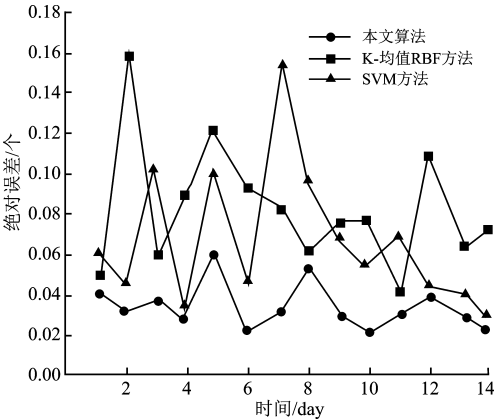


图 3 三种方法预警绝对误差比较结果

从图 3 中曲线分析得知,相比于其他两种预警方法,本文方法的预警绝对误差较低,绝对误差值都能保证在 0. 06 以内,说明本文方法预警准确率更高。

为了进一步验证本文方法的预警准确性,将平均相对误差和均方根误差作为衡量标准,平均相对误差、均方根误差的计算公式分别用式(15)、(16)进行描述:

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - y'_i|}{y_i} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2} \tag{16}$$

式中, y_i, y'_i 分别用于描述高可用性网络入侵实际值预警数量和计算预警数量; N 用于描述数据量。

由表 1 数据可知,本文方法的平均相对误差、均方根误差与 K-均值 RBF 方法及 SVM 方法相比较明显降低,说明本文方法的预警准确性很高.

表 1 预警准确性比较			
	本文方法	K-均值 RBF 方法	SVM 方法
MAPE/个	0.98	1.62	2.31
RSME/个	0.48	2.59	4.85

4 结束语

本文提出一种新的高可用性网络入侵预警方法,介绍了隐马尔可夫模型,在此基础上,构建依据 HMM 模型的实时高可用性网络入侵预警模型,依据危险理论对评估结果进行分析,将高可用性网络状态划分成四个等级,通过构建的 HMM-NSSP 模型实现高可用性网络入侵预警. 为了提高 HMM-NSSP 模型的预警精度,引入最大熵法对高可用性网络安全态势进行一致性判定. 实验结果表明,所提方法预警结果和实际结果基本一致,具有很高的预警精度,可用性很高.

参考文献:

[1] 宋杰,刘雪冰,朱志良,等. 一种能效优化的 MapReduce 资源比模型[J]. 计算机学报,2014,38(1):61-78.

[2] 吕海燕. 网络安全监控数据流的查询优化方法[J]. 电子设计工程,2014,22(13):29-31.

[3] 李绍晖,刘纪伟. 一种面向工业控制系统的改进 CUSUM 入侵检测方法[J]. 电子技术应用,2014,41(9):109-115.

[4] 黄伟建,周鸣爱. MapReduce 高可用性的研究与优化[J]. 计算机工程与设计,2015,35(12):3069-3098.

[5] 左毅,陈勇,游华. 网络入侵应急模型快速响应方法研究与仿真[J]. 计算机仿真,2015,31(3):268-302.

[6] 姜海洋,谢高岗. 众核处理器上的高可用性网络入侵检测系统[J]. 高技术通讯,2015,24(3):859-871.

[7] 黄亮,姜帆,荀浩,等. 面向软件定义网络的入侵容忍控制器架构及实现[J]. 计算机应用,2014,35(11):2898-3024.

[8] 赵夫群. 基于混合核函数的 LSSVM 网络入侵检测方法[J]. 现代电子技术,2014,38(21):87-88.

[9] 杨飞,朱志祥,梁小江. 基于 Ceph 对象存储集群的高可用设计与实现[J]. 微电子学与计算机,2015,32(12):59-63.

[10] 汪波,聂晓伟. 基于多目标数学规划的网络入侵检测方法[J]. 计算机研究与发展,2014,52(10):2340-2367.

作者简介:

龚健虎 男,(1967-),博士,讲师. 研究方向为数据库与知识库和数据挖掘. E-mail:gongtiger@21cn.com

(上接第 133 页)

作者简介:

张雪萍 女,(1968-),博士,教授. 研究方向为空间数据挖掘,智能信息处理技术等.

李围成(通讯作者) 男,(1991-),硕士研究生. 研究方向为空间数据挖掘. E-mail:13598591208@163.com.

祝玉华 女,(1965-),博士,教授. 研究方向为空间数据挖掘,粮食信息处理等.