



基于隐马尔可夫模型的网络安全态势预测方法

詹雄, 郭昊, 张翎, 毛澍

(全球能源互联网研究院, 北京 102209)

摘要: 针对网络安全态势预测动态、实时等特性, 设计了基于隐马尔可夫模型网络安全态势预测模型。根据 t 时刻的网络安全态势计算出 $t+1$ 时刻的态势, 同时结合最大熵算法, 对态势进行一致性判定, 提高了预测准确率。分析及模拟实验表明该基于隐马尔可夫模型的网络安全态势预测方法具有自主学习及主动防御特征, 同时平衡了速度和准确度要求, 能够高效、准确地进行网络安全态势预测。

关键词: 网络安全态势预测; 隐马尔可夫模型; 规则化; 网络安全

中图分类号: TP393.08

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2015394

A network security situation prediction method based on hidden Markov model

ZHAN Xiong, GUO Hao, ZHANG Peng, MAO Shu

Global Energy Interconnection Research Institute, Beijing 102209, China

Abstract: Aiming at the dynamic and real-time characteristics of network security situation prediction, a network security situation prediction model based on hidden Markov model (HMM) was designed. According to the HMM, the situation in $t+1$ from t time can be calculated. Meanwhile, the maximum entropy algorithm was combined with HMM to improve the accuracy rate. The analysis and simulation experiments show that the method has the characteristics of autonomous learning and active defense, and it also balance the speed and accuracy.

Key words: network security situation prediction, hidden Markov model, regularization, network security

1 引言

随着各种通信技术尤其短距离通信技术(如 Wi-Fi、蓝牙等)的快速发展,使得人们对网络的接入更加广泛、更加便捷,因此也对人们的生活产生了巨大的影响。网络生活已经成为人们生活的重要组成部分,因此也带来了诸如隐私信息泄露、谣言误导、网络欺诈、网络钓鱼等各种安全问题。同时,高级持续威胁(APT)攻击等成为当今网络主流的攻击方式,并呈现愈演愈烈的趋势。云端恶意代码样本已从 2005 年的 40 万种增长至目前的 60 亿种,全球恶意

代码样本数目正以每天可获取 300 万个样本的速度增长,继“震网”和“棱镜门”事件之后,各种网络漏洞和网络攻击层出不穷,例如,“心脏出血”漏洞、Bash 漏洞严重威胁成千上万的网络服务器和网络基础设施,军事、工业、金融等重要领域的信息系统安全面临严峻挑战^[1]。攻击方式多样、隐蔽,常常通过各种伪装、隐藏服务等方式植入客户端,在用户不知觉的状态下实现对网络服务器、计算机终端、便携式移动设备等智能设备的资源获取和控制,进而通过各种非法手段获取用户信息、银行账户等,简单地通过传统手段(如防火墙、入侵检测、杀毒软件等方式)进行网络保



护已经不能满足用户的安全需求。

如何开展新的网络态势下的安全预防,开展网络安全态势预测,实现主动智能的安全防护成为目前的研究热点,并被广泛关注。国家计算机网络应急技术处理协调中心2015年4月30日发布的《2014年我国互联网网络安全态势报告》数据显示,中国互联网安全情况依然严峻。以2014年为例,CNERT通报的漏洞安全相关事件达9 068起,与2013年相比增长3倍^[2]。漏洞、恶意程序、仿冒站点等情况如图1所示。

因此,开展网络安全态势预测具有重要意义,网络安全态势预测主要是对网络信息安全相关内容(包括网络安全行为等)进行网络状态分析和评估,智能评判信息安全性和趋向性,给出安全形势预判。目前,基于免疫理论、神经网络理论、病毒传播动力学等理论和技术的网络态势预测研究已经取得了一定的成果,包括采用线性回归的方法、基于神经网络的态势预测方法、基于蚁群计算的预测方法等,但这些方法由于建模时间较长,实时效果不明显,不能实时反映网络的安全态势。

本文深入分析网络安全趋势预测的特性及运行机理,结合隐马尔可夫模型(hidden Markov model,HMM)和预测算法^[3,4],研究并设计了网络安全态势预测系统,实现了实时预测网络安全态势的方法。基于Markov模式的预测系统利用网络安全态势评估信息建立网络安全态势预测模型,在监测时通过实时数据不断地对网络安全影响参数进行动态修正,从而实现网络安全态势实时监控与预测。

2 网络安全态势预测系统

2.1 系统框架结构设计

网络安全态势预测主要包括两个方面:态势和预测。其中,网络安全态势是指网络中各个要素的运行预期状况,即当前网络设备状况、网站内容信息、用户网上行为等要素,以此态势为前提,研究其未来发展的状况。预测主要是基于历史状态,总结发展规律,推算未来发展情况。由此可见,态势预测可以看成属于人工智能范畴。态势预测是一个目标逐渐明晰的过程。首先,通过总结态

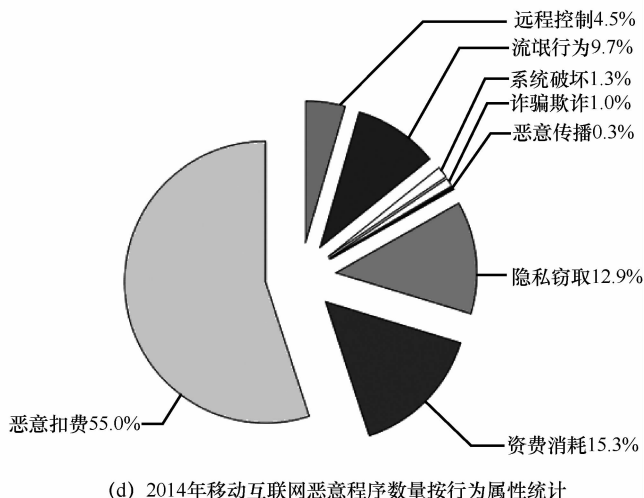
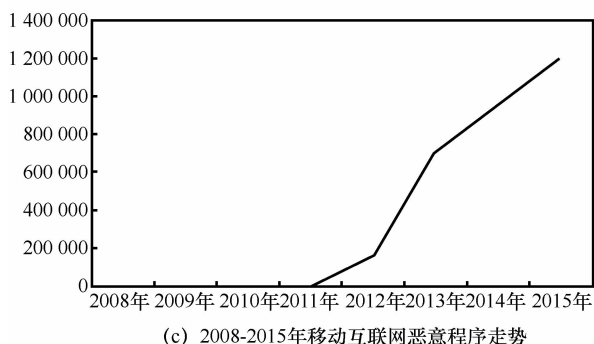
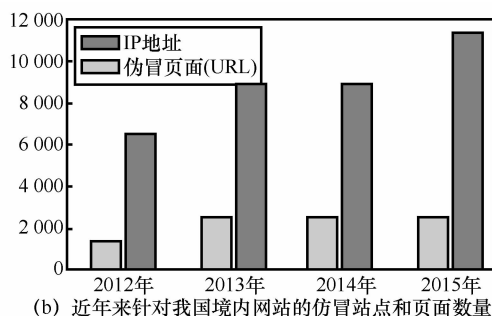
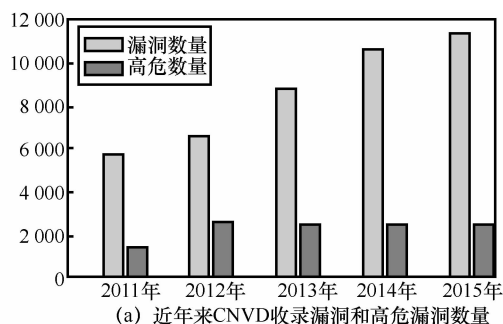


图1 2014年网络安全事件类型比例

势规律,建立态势感知模型;其次,根据态势基本特性和需求,定义并获取态势相关要素;最后,采集要素数据信息,进行数据梳理和清洗后,通过数据方法分析进行态势理解,进而对某个时间段内的态势发展进行预测。网络安全态势感知预测的主要目标是实现对网络安全趋势的短期预测。由于网络的互动性和随机性,这个系统是一个动态、实时的智能系统,通常包括数据采集、安全态势分析、安全态势感知、安全态势预测和判定等几个部分,如图2所示。

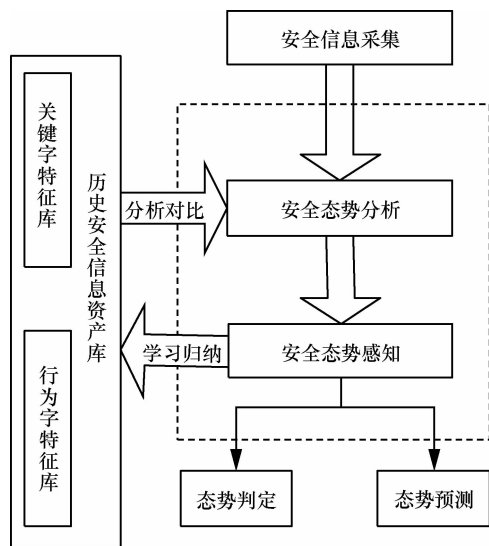


图2 网络安全态势预测系统框架

基于隐马尔可夫模型的网络安全态势系统主要由四大模块组成。

- 历史安全信息资产库。为了积累安全资料、加快查询速度,建立该资产库用于对已知安全信息特征码及安全行为特征的存储和查询,并进一步学习添加新的规则。
- 安全信息采集模块。该模块为信息入口,通过建立专用的采集方法,对安全信息进行挖掘,针对可能的内容和行为进行捕捉。
- 分析感知模块。分析模块主要针对采集模块信息进行历史判定,若为历史发生过的态势,则直接进入感知模块;如果在资产库里面不存在,移交安全态势感知模块对该信息进行抽象加工判定,加入资产库。
- 判定预测模块。该模块进一步对信息进行判断,依据相关规则,如果为非安全事件,依据挖掘及分类

技术算法进行判定,按正常信息事件处理;否则,通过隐马尔可夫模型进行态势预测,依据判定结果进行相关处理。

2.2 态势感知模型

态势感知模型是基于层次的感知模型,主要依据美国国防部联合指挥实验室(joint directors of laboratories, JDL)给出的数据融合模型结合实际应用建立的。态势感知是以态势分析为基础,依据态势分析结果,对已经发生过的和潜在要发生的安全事件进行感知。对于历史上已经发生过的安全事件,一方面有安全信息存留,因此产生安全态势;另一方面,该安全事件处于萎缩状态,发展速度和关切影响已经逐步衰落,其态势呈现衰减回落状态。该种态势判断只需记录即可。图3为典型的安全态势感知模型^[5]。

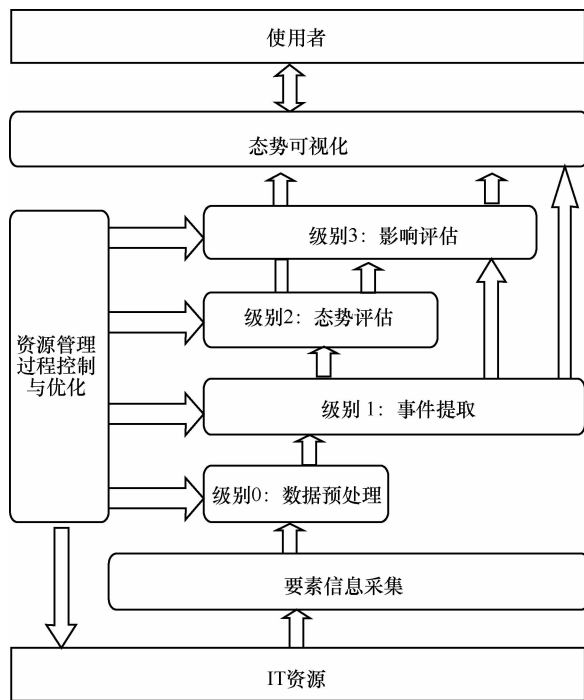


图3 典型安全态势感知模型

在该感知模型中,充分体现了人机交互状态。整个态势感知模型被分为了5个级别(阶段)^[4,5]。

- 数据预处理:该级别为可选级别,主要用于前置应用,对非规范化数据进行处理。在本文的模型中,对一些特殊非结构化数据需要进行数据预处理。
- 事件提取:要素信息采集后规范化和有序化,提取事件级别特性,抽象事件。



- 态势评估:是模型的重要组成部分,主要是对数据和事件进行态势分析并形成分析报告,进而形成网络综合态势图,包括态势分析和态势感知。
- 影响评估:是对未来结果的考量,通过将当前态势映射到未来,对行为的影响进行预测评估,实现具体的预测结果。
- 资源管理、过程控制与优化:根据要素标准,对数据和事件进行规则化,对整个态势发展过程进行实时监控与评价。

2.3 安全态势预测模型与算法

网络安全态势具有随机性和爆发性,常常与网络热点、网络操作、网络服务等相互绑定,通过欺诈、伪造、诱导等方式开展攻击,以取得对未授权的计算机或网络等资源的控制或窃取。

2.3.1 隐马尔可夫模型

隐马尔可夫模型是经典的分析统计模型,用来描述一个含有隐含未知参数的马尔可夫过程。隐马尔可夫模型不能直接对对象进行观察,但是能通过观测向量序列确定隐含参数。通过隐马尔可夫模型,可以观察到每个观测向量的概率密度分布表现,经过这个测量可以全面了解各个向量状态。针对网络安全态势,参照马尔可夫经典模型,依据态势特性,可以建立如下隐马尔可夫模型^[56]。

设 H 为隐马尔可夫模型下的网络安全态势,则有:

$$H=(S,O,\Gamma,A,B) \quad (1)$$

其中, S 为隐含状态,可定义为马尔可夫模型中实际所隐含的状态。即此状态为不可直接观测状态; O 为可直接观测状态,通常要与模型中的隐含状态建立关联。 Γ 为初始时刻 $t=1$ 的概率矩阵; A 为模型的隐含状态转移概率矩阵, $A_{ij}=P(S_j|S_i)$, ($1 \leq i, j \leq N$),表示在时间维度 t 时刻、网络的安全态势为 S_i 的条件下,在时间维度 $t+1$ 时刻网络安全态势状态是 S_j 的概率。其中, B 为可观测状态转移概率矩阵, A, B 两种状态相互关联,建立联系。 $B_{ij}=P(O_j|S_i)$, ($1 \leq i \leq M, 1 \leq j \leq N$), N 代表隐含状态数目, M 代表可观测状态数目,表示在时间维度为 t 的时刻、隐含网络安全态势状态是 S_j 的条件下,观察状态为 O_i 的概率。

2.3.2 安全态势算法

本文预测算法对隐马尔可夫模型前向算法进行改进,同时结合最大熵模型^[7,8],将信息安全内容设为可见观察序列,假定安全级别是不能够直接被观察的参数,进行网络安全态势预测。

假设隐马尔可夫模型的状态空间为: $S=\{\text{Normal}, \text{Noise}, \text{Attack}, \text{Danger}\}$ 。

可见观察序列符号集: $O=\{\text{high}, \text{normal}, \text{low}\}$ 。

在时间维度为 $t+1$ 时刻观察到态势值为 O_{t+1} 。则有:

$$H_{t+1}=H_t \times \sum_{i=1}^{101} P_i(O_{t+1}|O_t) \quad (2)$$

为了增加隐马尔可夫模型估算精度,针对在有限知识范围对未知状态预测的情况,可采用最大熵模型和原理。选取符合已知假设条件但熵值最大的概率分布,即在已经获取的资产库和采集的网络信息当前态势的知识库基础上,推断满足条件的合理分布,建立符合已知知识特性的不确定推断。结合最大熵理论,定义网络态势特性方式表达为二值特征 $f(x, y)$, 其中, x 为获取信息值, y 为期望值, E 表述为最大熵算法;若 f_i 对模型有用,则构建一个能生成训练样本 $\tilde{p}(x, y)$ 的约束模型。

$$P_E=\{p_e|E_p(f_i)=E_{\tilde{p}}(f_i), 1 \leq i \leq k\} \quad (3)$$

从而最大熵算法提出一致性最大熵值。通过信息采集分析网络信息安全态势特性;通过系统态势分析监护进程对网络内容和进程进行监控,抽取特征数据和特征进程动作;通过信息采集进行挖掘和聚类分析;通过关联算法进行识别及动作规则化。安全判定与预测过程如图4所示。

根据式(1),设 $a_i \in A, b_i \in B, O_i \in O, \tau \in \Gamma, I$ 为所有长度

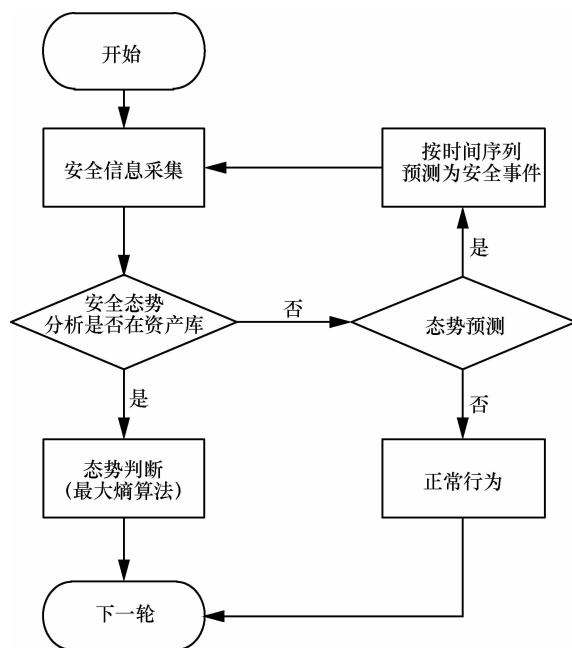


图4 网络安全态势判定与预测

为 T 的状态序列, $I=(i_1, i_2, \dots, i_T)$ 。则预测算法如下。

(1) 初始值 $t=1$, 观测为 $a_1(i)=\tau b_i(O_1)$, 其中 $i=1, 2, \dots, N$ 。由初始状态概率矩阵, 得出到达每个状态的概率 P_i , 然后再由 P_i 乘以 $P(O_1|I_i)$, 得到由各个状态产生观测 O_1 的概率。将计算出的结果放在一个 $N \times T$ 的二维数组 M 中, $a_1(i)$ 为第一列的元素。

(2) 递推 $t=2, 3, \dots, T$, 则有 $a_{t+1}(i)=\left[\sum_{j=1}^N a_t(j)a_{ji}\right]b_i(O_{t+1})$, $i=1, 2, \dots, N$ 。根据时刻 t 观测为 O_t 的概率, 计算时刻 t 到达状态 I_t 的概率 $P(I_t|t)$; 然后, 再乘以由状态 I_t 产生观测态势 O_t 的概率。由 $t=1$ 时刻达到 t 时刻的状态 I_t , 有 $N_1 \times N_2 \times \dots \times N_t$ 中可能的路径, 根据 $t-1$ 时刻二维数组 M 先验概率, 只需在 $t-1$ 列的基础上, 求出到达 I_t 的 N 种路径的概率之和, 即 $\sum_{j=1}^N a_t(j)a_{ji}$, 然后由第一步的结果乘以由状态 I_t 产生观测 O_t 的概率即可。

(3) $P(O|\tau)=\sum_{i=1}^N a_t(i)$, 将 M 数组的最后一列相加即为所求。

2.4 算法分析

利用隐马尔可夫模型, 对安全态势的概率进行计算分析, 依据时间调用序列, 在具体内容上结合最大熵的算法模型提取最优值, 进行安全态势的规则化和归一化量化处理。为了增加判定准确度和判断速度, 在信息采集阶段进行特征提取, 建立规则化数据, 在安全态势发生的初期对网络攻击进行捕捉, 并获得其特征, 对应调用态势分析模块和对比资产库, 并通过网络数据行为特征和时间特征序列, 开展特征性评判。随着系统的不断运行, 系统通过态势分析和感知, 不断提取新的安全态势特征, 逐步完善资产库, 提高模型自相似性、高概率、智能化。但是, 由于资产库规模及网络规模的双重影响, 同时根据判不准定理^[7], 基于网络的交互性及安全态势不断动态变化, 基于隐马尔可夫模型的判定概率会逐步提高, 在初始时间, 概率分析的条件不够充足, 需要不断学习和训练。为了提高准确性, 可以通过最大熵算法等, 对内容的一致性进行判定, 并不断优化判定方法和预测方法, 达到准确性比较高的监测预警。

基于隐马尔可夫模型的网络安全态势预测算法具有较好的性能, 以获取的网络安全态势样本为基础, 进行实验运行测试表明, 该预测算法可以有效判定和预测网络安

全态势。尤其针对网络相关内容规则化后, 可以大大提高检测的速度和准确度。

3 结束语

本文根据网络安全态势特性, 依据隐马尔可夫模型, 建立面向网络内容和网络行为的实时、智能的网络安全态势预测模型。同时优化判定和预测算法, 建立资产库。该模型在获取当前网络安全态势信息基础上, 开展基于隐马尔可夫模型的网络安全态势评估, 根据概率统计原理, 分析不同时间维度下不同攻击下系统的威胁值。根据 t 时刻的网络安全态势计算出 $t+1$ 时刻的态势, 同时结合最大熵算法, 对态势进行一致性判定, 提高了预测准确率, 对网络安全态势进行精准预测。通过实验样本的测试运行, 进一步表明该方法具有自主学习及主动防御特征, 有效地平衡了速度和准确度要求, 能够高效、实时地进行网络安全态势预测。

参考文献:

- [1] 中国互联网协会, 国家互联网应急中心. 中国互联网网站发展状况及其安全报告 (2015 年) [EB/OL]. [2015-03-20]. http://www.cac.gov.cn/2015-03/20/c_1114714197.htm.
Internet Society of China, CNCERT/CC. The development status of Chinese website and safety report (2015) [EB/OL]. [2015-03-20]. http://www.cac.gov.cn/2015-03/20/c_1114714197.htm.
- [2] 国家计算机网络应急技术处理协调中心. 2014 年我国互联网网络安全态势报告 [EB/OL]. [2015-05-12]. <http://www.cert.org.cn/publish/main/upload/File/2014%20security%20situation%20report.pdf>.
The National Computer Network Emergency Response Technical Team/Coordination Center of China. 2014 China internet network security situation report [EB/OL]. [2015-05-12]. <http://www.cert.org.cn/publish/main/upload/File/2014%20security%20situation%20report.pdf>.
- [3] 黄同庆, 庄毅. 一种实时网络安全态势预测方法[J]. 小型微型计算机系统, 2014, 35(2): 303-306.
HUANG T Q, ZHUANG Y. An approach to real-time network security situation prediction [J]. Journal of Chinese Computer Systems, 2014, 35(2): 303-306.
- [4] ARNES A, VALEUR F, VIGNA G, et al. Using hidden Markov models to evaluate the risk of intrusions [C]// The 9th International Symposium on Recent Advances in Intrusion Detection, September 20-22, 2006, Hamburg, Germany. [S.l.: s.n.], 2006: 145-164.
- [5] 叶蓬. 网络安全态势感知在安全管理平台中的应用研究[J].



信息网络安全, 2010(4): 51-54.

YE P. Network security situation awareness in the role of security operations center[J]. Network Information Security. 2010(4): 51-54.

[6] SNIDARO L, VISENTINI I, BRYAN K. Fusing uncertain knowledge and evidence for maritime situational awareness via Markov logic networks [J]. Information Fusion, 2015 (21): 159-172.

[7] 何鸿君, 罗莉, 董黎明, 等. 广义病毒的形式化定义及识别算法[J]. 计算机学报, 2010, 33(3): 562-568.

HE H J, LUO L, DONG L M, et al. Formal definition of generalized virus and its identifying algorithm [J]. Chinese Journal of Computers, 2010, 33(3): 562-568.

[作者简介]



詹雄(1978-),男,全球能源互联网研究院安全研究室主任,主要研究方向为电力信息安全技术。



郭昊(1981-),男,全球能源互联网研究院信息安全工程师,主要研究方向为网络与信息安全、数据库安全。



张翎(1981-),男,全球能源互联网研究院信息安全工程师,主要研究方向为信息安全评估与信息安全监测。



毛澍(1983-),男,全球能源互联网研究院信息安全工程师,主要研究方向为网络与信息安全、大数据监测分析。