

应用于网络态势感知的安全数据聚合分析

张 焱^{1,2}, 黄曙光¹, 朱俊茂², 任飞飞²

(1 电子工程学院, 安徽 合肥 230037; 2 北京邮电大学, 北京 100876)

摘 要: 为了处理异构网络安全设备产生的海量报警数据, 提出一种基于加权 DS 证据理论和 FCM 的安全数据聚合算法。首先利用网络安全设备的权重以及对不同攻击的置信度信息, 运用 DS 证据理论对报警数据进行初步融合, 然后借助 FCM 的模糊推理能力对数据进行全局融合, 克服了漏报率高的缺陷, 增强了系统处理复杂攻击的能力。

关键词: 数据融合; DS 证据理论; 模糊认知图; 态势感知

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)08-0123-03

Analysis of Security Data Fusion in Network Situation Awareness

ZHANG Yan^{1,2}, HUANG Shu-guang¹, ZHU Jun-mao², REN Fei-fei²

(1 Electronic Engineering Institute, Hefei 230037, China;

2 Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: To deal with the mass alert data generated by heterogeneous network security equipments, this paper proposes a security data fusion algorithm based on weighted DS evidence theory and Fuzzy Cognitive Map. According to different weight and confidence of network security equipments, it initially fuses the alert data by DS evidence theory. Then it takes overall alert fusion through fuzzy reasoning of FCM. The proposed method overcomes the problem of high false negative rate and enhances the ability to cope with complex attacks.

Key words: data fusion; DS evidence theory; fuzzy cognitive map; situation awareness

1 引言

大多网络安全态势感知(Network Security Situation Awareness, NSSA)系统都能为网络提供深度防御。但目前的 NSSA 技术在感知的可用性、灵活性以及准确性上还存在一些问题, 如大量重复报警、漏报误报严重、海量报警泛滥等, 影响了 NSSA 的充分性和高效性。其中的主要原因是没有有效利用攻击事件间的逻辑关系。文献[1]提出运用神经网络融合多重异构传感器信息, 并以时间和攻击严重性来衡量网络安全态势。文献[2]使用了简单加权以及灰色理论(Grey Theory)来实现网络安全态势感知, 并通过概念模型来表示。本文提出一种基于异构网络传感器的安全数据融合算法, 分别利用 DS 证据理论和模糊认知图的方法对不同的攻击形态进行

关联融合, 实时获取系统的安全态势信息, 从而更加有效地辅助决策者调整安全策略。

2 相关研究

2.1 DS 证据理论

DS 证据理论是由 Shafer 于 1976 年提出的, 用于描述不同等级的精确度, 由此引入对未知不确定性的描述, 通常应用于医疗诊断(diagnosics)、风险分析(risk analysis)以及决策分析(decision analysis)等领域。

使用 DS 融合规则前, 首先需要定义目标框架, 即欲理解的事件, 接着利用 BPA(Basic Probability Assignment)公式分配不同传感器对不同事件描述的置信程度。

假设有两项 IDS: o_1 及 o_2 , 一项攻击事件: H 。

o_1 中假设 H 为真的情况下置信度为 $m_1(H)$, 置信度 $m_1(H)$ 代表 o_1 支持 H 发生的概率; o_2 中假设 H 为真的情况下置信度为 $m_2(H)$, 置信度 $m_2(H)$ 代表 o_2 支持 H 发生的概率. 通过 DS 规则融合 o_1, o_2 的证据, 结果为

$$m_{12}(H) = \frac{\sum_{B \cap C = H} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)} \quad (1)$$

$m_{12}(H)$ 即为融合两个警讯后所得最后警讯可能发生的概率. 已有多项研究利用此方法改善警讯误报率(Rate of False Positive) 高的缺点^[3].

文献[4] 提出基于权重的 DS 证据理论, 使用不同传感器的不同可信程度及权重进行报警信度的融合, 评估报警正确性. 式(1)中假设 o_1 及 o_2 两项 IDS 的信任程度是一样的. 但这并不符合实际情况, 因此提出了权重式的 DS 证据理论, 其改进后的公式见式(2).

$$m_{12}(H) = \frac{\sum_{B \cap C = H} [m_1(B)]^{w_1} [m_2(C)]^{w_2}}{\sum_{B \cap C \neq \emptyset} [m_1(B)]^{w_1} [m_2(C)]^{w_2}} \quad (2)$$

式中, o_1 及 o_2 二项 IDS 的权重分别为 w_1 及 w_2 . 而权重的评估方式有许多种, 例如, 可基于过去所有评估的正确率统计或根据专家的经验进行调整.

2.2 模糊认知图

模糊认知图^[5](fuzzy cognitive map, FCM) 是一个将模糊反馈动力系统因果事件、目标与趋势通过各概念间关系连接起来的带有反馈的有向网络, 如图 1 所示.

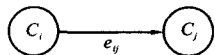


图 1 二项 FCM 示意图

在概率 C_i 与 C_j 之间的连接值为 e_{ij} , 其值介于 -1 与 1 之间. 当该值等于 0 时, 代表 C_i 与 C_j 之间没有任何关系; 当该值大于 0 时, 代表 C_i 与 C_j 之间有正向的因果关系, 当 C_i 增加时 C_j 则以 e_{ij} 倍数增加. 当 e_{ij} 的绝对值越高时, 其因果关系的影响越强.

针对来自不同安全事件 I_k (影响力为 e_{ij}) 的证据, 资源 R_i 所处的安全状态可用式(3)表示:

$$IAS(R_i)(t_{n+1}) = \left[\frac{\sum_{k=1}^n (I_k)(t_n) \times e_{ki}(t_n)}{\sum_{k=1}^n e_{ki}(t_n)} \right] \quad (3)$$

式中, I 代表安全事件, 将各安全事件量化后乘上其所具有的影响力, 再除以所有影响力之和, 得到 IAS 的数值. 按照 FCM 模型计算出的 IAS 数值越

高代表所处的环境越不安全. 因此使用 FCM 结合不同安全事件所代表的证据力, 即可以评估整体网络的安全态势.

3 安全数据聚合算法

本文提出一种异构网络传感器数据处理算法(图 2), 利用不同的数据融合方法, 和管理者设定各报警系统的信度及权重, 以及回馈机制, 解决在异构网络传感器中产生的报警冲突等问题.

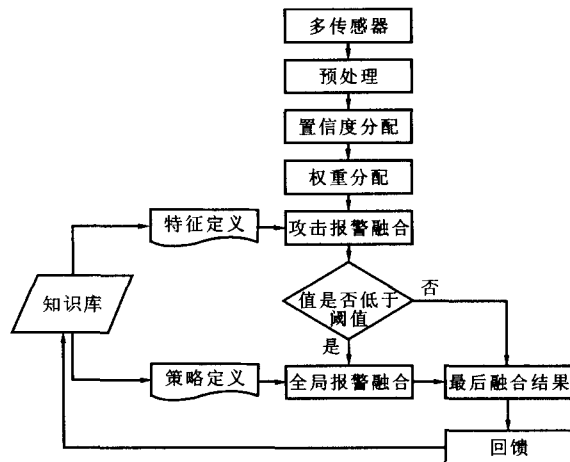


图 2 异构网络传感器报警融合流程

从各种不同的网络传感器送回的攻击报警数据经过预处理转化为统一格式, 经过可信度及权重分配后进行基于 DS 证据理论的攻击报警融合. 如果是紧急报警, 则直接上报网管人员; 否则继续进行安全策略的确认, 即进行全局的报警融合. 最终得到的报警信息经过后端管理人员判别后, 回馈到知识库. 知识库中主要包含攻击特征定义及安全策略定义, 前者为某传感器对于某一攻击之可信程度以及权重值, 后者定义了辨识某一攻击型态所需使用的辅助信息.

3.1 基于加权 DS 理论的攻击报警融合

对于各安全设备除了通过可信度来判定检测的精确性外, 同时应该对传感器赋予不同的权重信息. 我们采用最大熵(Maximum Entropy, MaxEnt)原则^[6]来测量权重. 由最大熵原则定义可知, 在已知的条件下, 熵(Entropy)最大的事物, 最可能接近它的真实状态.

最大熵理论在给定一群已知的事件集合中, 选择符合所有事件的最优模型并且尽可能的达到均匀(Uniform)分配. 即在给定多个已知的训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, 其中 x 为输入, y 为输出. $\tilde{p}(x, y)$ 是在 x 出现的情况下, y 的经验概

率,也就是 y 在样本集中出现的概率,见式(4).

$$\hat{p}(x, y) = \frac{1}{N} \times \text{样本集中}(x, y) \text{ 出现的次数} \quad (4)$$

而 $p(x, y)$ 则为在 x 出现的情况下, y 的实际概率. 因此随机事件的不确定性可以用以下条件熵来衡量,见式(5).

$$H(p) = \sum_{x, y} \hat{p}(x) p(y | x) \log p(y | x) \quad (5)$$

我们将最大熵原理运用于权重设定上,根据前文提到的融合公式(2),将其套用后转变成式(6).

$$m_{12} = \frac{1}{N_k} \exp(w_1^k f_1(H, t_i) + w_2^k f_2(H, t_i)) \quad (6)$$

式中, t_i 为时间, $f_i(H, t_i) = \log(H | t_i)$, N_k 的定义见式(7).

$$N_k = \sum_{H \in \{A_k, \neg A_k\}} \exp(w_1^k f_1(H, t_i) + w_2^k f_2(H, t_i)) \quad (7)$$

对于最大熵的参数估算采用 D. Pietra 等人所提出的 IIS(Improved Iterative Scaling Algorithm)演算法^[7], IIS 演算法是目前最大熵参数求解中的最常用算法. 最大熵理论配合 IIS 演算法求出参数最优值,通过这种方式可以合理地设定权重.

3.2 基于 FCM 的全局融合

并非所有攻击都能通过单一的攻击特征发现,通常需要使用安全策略来辅助检测. 在定义安全策略时,我们发现 FCM 模型能有效地截取事件的因果关系同时支持因果推论.

以口令试探攻击为例,入侵者渗透特定主机,并尝试使用多个用户及口令来获取主机的控制权. 然而,当产生“登陆失败”报警时,并不能说明系统发生了异常. 对于这种攻击,仅靠单一攻击特征是无法检测的,其中涉及的特征包括“登录失败的次数”、“间隔时间”、“用户名”以及“主机”. 图 3 描述了使用 FCM 模型表达的口令试探攻击. 此外,每一种特征对事件的影响是不同的,需要为它们设定不同的权重,如“间隔时间”权重设定为 -0.5,负数代表间隔时间和判断是否发生异常呈反比关系,也就是说间隔时间越长代表发生异常的现象越不可能发生,若为正数则代表呈正向关系. 其数值部分则表示正向或负向的影响程度有多大.

对于由多个攻击特征才能检测的情况,我们定义了一系列规则来检查该现象是否发生. 针对图 3 所描述的情况,可将规则设定如下:

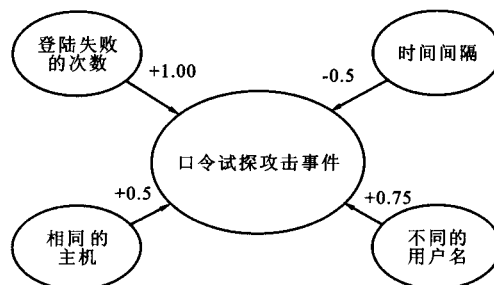


图 3 口令试探攻击的 FCM 模型表示

If number of login failures is moderate and interval is short and this happened for same machine and different users then, Login_Failure_Same_Machine_Diff_User is activated highly.

当符合上述模糊规则时,就可判定系统或网路正处于异常状态,并上报后端管理人员作进一步处理.

4 结束语

本文通过对复杂攻击场景的研究,采用加权的 DS 证据理论对攻击报警进行融合,并采用模糊认知图对具有时序逻辑特征的复杂攻击进行了融合. 该方法克服了传统方法的缺点,能够有效降低检测漏报率,增强系统处理复杂安全事件的能力.

参考文献:

- [1] Wang Huiqiang, Lai Jibao, Ying Liang. Network security situation awareness based on heterogeneous multi-sensor data fusion and neural network[C]// Second International Multisymposium on Computer and Computational Sciences. USA: Iowa City, 2007.
- [2] 赖积保, 王颖, 王慧强, 等. 基于多源异构传感器的网络安全态势感知系统结构研究[J]. 计算机科学, 2011, 38(3): 144-149.
- [3] Liu Mixi, Yu Dongmei, Zhang Qiuyu, et al. Network security situation assessment based on data fusion[C]// 2008 Workshop on Knowledge Discovery and Data Mining. Australia: Adelaide, 2008.
- [4] Dong Yu, Deborah Frincke. Alert confidence fusion in intrusion detection systems with extended dempster-shafer theory[C]// 43rd ACM Southeast Conference. Kennesaw, GA, USA, 2005.
- [5] 刘萍萍, 罗钧旻, 乔奎贤. 基于神经网络的模糊认知图的演化研究[J]. 微电子学与计算机, 2007, 24(7): 69-71.

(下转第 128 页)

$\{\text{Kad ID}, \text{userID}\}$, $\{\text{Kad ID}, \text{IP}, \text{userID}\}$ 作为候选节点标识, 使用第 2 节所述的方法进行验证, 得到 $\{\text{userID}\}$ 为最佳节点标识。

表 2 采集数据

采集时间	7 day
随机采集次数	280
采集节点总数	3×10^6
可建立 TCP 连接节点信息记录数	1.1×10^6

$\{\text{userID}\}$ 做为最佳节点标识可以在如下场景中应用: Kad 网络的用户跟踪; Kad 网络用户行为分析等。本节为证明这点, 采用 Kcrawler 对 Kad 网络用户的共享文件信息进行了为期 7 天的探测, 共获得用户拥有的文件条数约 530 万条, 然后分别采用 $\{\text{KadID}\}$, $\{\text{IP}\}$, $\{\text{userID}\}$, $\{\text{Kad ID}, \text{IP}\}$, $\{\text{IP}, \text{userID}\}$, $\{\text{Kad ID}, \text{userID}\}$, $\{\text{Kad ID}, \text{IP}, \text{userID}\}$ 作为标识对用户所拥有的文件进行了统计, 实验结果如图 1 的“箱线图”所示。由图可见, $\{\text{userID}\}$ 在短期 (只有 7 天) 的测量中也能对应的最多的文件数量, 故采用 $\{\text{userID}\}$ 作为节点标识可以更准确地探测用户所拥有的文件情况。从而证明了 $\{\text{userID}\}$ 是集合 $\{\text{KadID}, \text{IP}, \text{userID}\}$ 中最佳的节点标识。

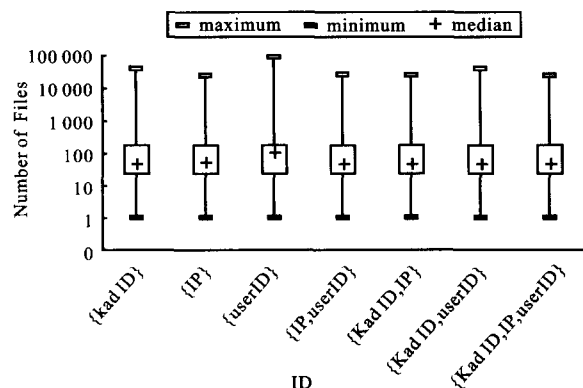


图 1 标识对应文件数量箱线图

5 结束语

本文设计并实现了采用随机采集方式的 Kad

网络节点信息爬虫: Kcrawler; 然后采集了实际 Kad 网络的候选标识: Kad ID, IP, userID 之间对应关系的节点信息, 通过计算和实验验证得出 $\{\text{userID}\}$ 是 Kad 网络的最佳节点标识, 为 Kad 网络的研究奠定了基础。Kad 网络的研究受到越来越多的关注, 更进一步的研究工作包括: Kad 网络共享内容的获取及其分类, Kad 网络完全拓扑图的采集等。

参考文献:

- [1] Merkur. eMule[EB/OL]. [2011-05-13]. <http://www.eMule-project.net>.
- [2] 彭凯, 邓天平. 基于市场机制的 P2P 流媒体网络节点选择算法研究[J]. 微电子学与计算机, 2009, 26(3): 97-101.
- [3] Steiner M, En-Najjary T, Biersack E W. Analyzing peer behavior in kad[R]. London: Institut Eurecom, 2007.
- [4] Subhabrata Sen, Oliver Spatscheck, Dongmei Wang. Accurate, scalable in-network identification of P2P traffic using application signatures[C]// WWW 2004. New York: ACM Press, 2004: 512-521.
- [5] Yoram Kulbak, Danny Bickson. The eMule protocol specification[R]. Israel: The Hebrew University of Jerusalem, 2005.
- [6] Saroiu S, Gummadi P K, Gribble S D. A measurement study of peer-to-peer file sharing systems[C]// Proc. of the Multimedia Computing and Networking (MMCN). London: IEEE Press, 2002.
- [7] Stutzbach D, Rejaie R, Sen S. Characterizing unstructured overlay topologies in modern P2P file-sharing systems[C]// Proc. of the 5th ACM SIGCOMM Conf. on Internet Measurement. Berkeley: ACM Press, 2005.
- [8] Stutzbach D, Rejaie R. Improving lookup performance over a widely-deployed DHT[C]// Proc. INFOCOM 2006. Barcelona: IEEE Press, 2006: 185-192.

作者简介:

孙宁青 男, (1963-), 副教授. 研究方向为计算机应用技术。

(上接第 125 页)

- [6] Stephen G Batsell, Nageswara S Rao, Mallikarjun Shankar. Distributed intrusion detection and attack containment for organizational cyber security[EB/OL]. [2011-05-22]. <http://www.ioc.ornl.gov/projects/documents/containment.pdf>.
- [7] Adam Berger, Stephen Della Pietra, Vincent Della Pi-

etra. A maximum entropy approach to natural language processing[C]// Computational Linguistics. University of Amsterdam, Netherlands, 1996.

作者简介:

张焱 男, (1983-), 博士研究生. 研究方向为计算机网络、信息安全。