DOI:10.13718/j.cnki.xsxb.2019.11.018

网络安全态势感知框架及随机森林评估模型①

钱真坤

四川文理学院 后勤服务处，四川 达州 635000

摘要：针对传统网络安全态势感知评估过多依赖专家经验的问题，提出一种基于随机森林的多层次网络安全态势感知(Cyber Security Situational Awareness，CSSA)框架评估模型. 首先将 CSSA 的过程与安全数据生命周期进行对齐，并分析 CSSA 的需求，提出 CSSA 多层次分析框架，然后采用随机森林算法，构建 CSSA 评估模型，该模型基于多个分类器组合的思想，由决策树构成，每棵树依赖于独立样本，以及森林中所有树的随机向量分布相同的值. 在进行分类时，每棵树投票并返回票数最多的类，这使得网络安全态势评估更为客观和准确. 实验表明，与贝叶斯网络相比，此模型可以更快速、更准确地评估当前的网络安全情况.

关　键　词：网络安全态势感知；多层次 CSSA；随机森林；决策树；评估模型

中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1000－5471(2019)11－0118－06

随着互联网技术的迅速发展，网络设备成为社会基础设施，给整个社会带来了方便，但随之增多的是网络攻击和破坏行为[1-3]，造成了较为恶劣的影响，因此对于网络安全态势感知的研究越来越受到关注[4-5].

网络安全态势是指受监控网络的全球安全状况，网络攻击在一定时间段内受到的影响，以及对网络安全总体目标的影响[6-7]. 一般来说，安全态势信息由时间维度和空间维度 2 个方面组成. 网络给人们的生活带来了方便，同时随着网络攻击和破坏行为增多，已经影响到了个人与社会的正常网络环境.

为了应对大规模网络中增加的信息安全威胁，已经使用了多种安全设备，这些设备会产生很多安全事件，就会造成过多的警告信息，很难准确获得整个网络的安全状态. 为了解决这个问题，态势感知的概念被引入网络安全系统[8]. Singh 等[9]构建了一种动态网络安全态势感知框架，该模型具有适应新实体的可扩展性，能够适应网络中的配置更改，同时处理异构数据以获得网络安全的整体视图. Wu 等[10]中提出了一个复杂网络的网络安全态势感知模型，引入复杂网络理论对 SSA 数据建模，通过分析网络的性质数据，可以更有效地挖掘当前数据的特征. Li 等[11]根据多方面信息的融合提出了一种改进的安全态势感知模型.

如何掌握整体安全态势，量化当前网络中不安全因素引起的网络损害程度，评估网络安全态势感知的准确成为一个重要方向. 张劭帅等[12]采用集对分析理论对无线传感器网络安全态势进行评估，该方法对中低强度的攻击灵敏度更高. Zhu 等[13]提出一种基于信息融合的网络安全态势感知模型评估方法，网络安全状况是融合脆弱性、威胁性和基本操作性 3 个方面评估的结果，该评估方法多维、准确、直观地描述了网络系统的整体安全态势感知. 李玺等[14]中提出了基于 Markov 决策过程和博弈论思想的网络安全态势评估方法，该方法能够准确给出装备保障信息网络的安全态势值.

基于国内外在网络安全态势感知领域的研究成果，本文提出了基于随机森林的多层次网络安全态势感

① 收稿日期：2018－06－08
基金项目：四川省教育厅资助科研项目(18ZB0511).
作者简介：钱真坤(1980－)，男，硕士，实验师，主要从事计算机应用技术研究.

知模型评估方法. 与贝叶斯网络、BP 神经网络相比，随机森林算法可更准确，更高效地工作. 另外，对于不平衡的数据，本文方法可以平衡错误，且不会导致过度拟合.

1　网络安全态势感知

文献[15]将情境意识定义为在一定时间和空间内对环境中的元素的感知，对其含义的理解及其在不远的将来的状态预测. 如图 1 所示，文献[15]在 3 个层次上描述了情境意识，即感知、理解和投影.

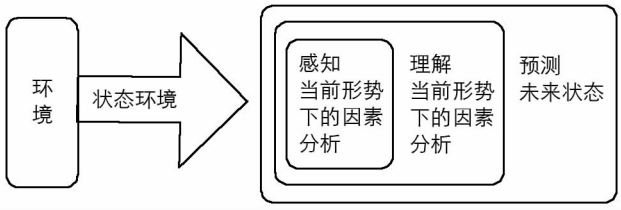

图 1　态势感知模型

第 1 层感知层，感知环境中对特定决策者非常重要的关键因素，感知包括根据从环境中不同来源收集的数据，对相关因素在不同时间和空间的状态、属性和动态进行识别和评估. 第 2 层理解层，理解第 1 层因素的意思. 理解涉及不相关元素的整合和关联，需要在决策者作出合理决策的角色中理解. 第 3 层预测层，将对未来情况的理解进行预测，以预测决策者在未来决策背景下对这些要素的影响. 首先要了解因素状态和动态，并要了解第 1 层和第 2 层特征元素，以预测一段时间内环境中会发生什么状况.

后来有研究者提出了一种态势感知模型，在动态环境中决策取决于对环境持续的最新分析，并且需要在相当狭窄的时间段内进行.

网络安全态势感知的主要思想是分析网络基础设施中的环境并创建特定事件和可视化，以便进行高效和快速的决策. 简而言之，CSSA 可以被描述为在网络基础设施中应用网络安全的态势感知. 网络安全态势感知具有以下层次.

感知涉及网络基础设施情况的证据收集. 感知是获得网络环境中元素的知识，例如入侵检测系统报告的警报、防火墙日志、扫描报告以及它们发生的时间，这为理解、预测和解决提供了基础保障.

理解包括分析证据来推断确切的威胁等级、攻击类型以及相关或相互依赖的风险. 理解利用一系列相关技术和程序来分析和汇总在网络基础设施中感知的证据数据.

预测涉及预测性评估，以缓解网络状况，进一步解决未来事件. 通过对安全态势量化评估，得到态势评估报告，然后进行安全态势动态预测，这个过程是基于从动态网络元素中提取的知识来进行预测的.

CSSA 涉及对攻击和攻击轨迹的理解，对攻击模式和相关性的理解，对基础设施和信息资产的影响，以及威胁级别在不久的将来会发生的事情预测.

2　基于随机森林的多层次 CSSA

2.1　多层次 CSSA 模型

CSSA 过程基本上对应于安全数据经历的生命周期，在生命周期中数据采取不同的形式，从原始传感器数据的开始，中间涉及清理数据、融合数据和感知事件，并以情景模式结束. 安全数据生命周期的上游主要涉及数据预处理、分布式数据存储、数据融合和事件处理，而安全数据生命周期的下游主要涉及情境评估和建模、顺序模式挖掘和模式分析、背景推理和管理以及情境可视化.

从安全信息获取到构建网络安全情况的模型，它必须是一个完整的过程. 为了从安全数据中挖掘更高层次的价值，CSSA 将进行多层次的分析过程. 图 2 说明了 CSSA 多层次分析框架中的信息流. 从安全传感器到情景模式的信息流形成了一个信息价值链，以实现 CSSA.

在最底层，传感器从包含系统基础设施和信息资产的设施中获取活动、配置和拓扑信息. 传感器数据在输入到分布式数据存储区时必须进行清理和标准化. 存储在分布式数据存储中，作为网络基础架构历史和当前状态的基本事实.

对于数据采集，将数据集成到各种传感器的不同格式中存在一个难点，从网络流量记录到使用统计数据和拓扑图，将不同格式的数据从异构数据源转换为合适的语法层面的共同表示格式是困难和昂贵的. 更实际的方式应该转向在语义或服务级别进行数据集成，例如数据联合、数据虚拟化、数据即服务. 来自该通用表示的元素在分布式数据存储中链接.

数据经过处理后，CSSA 的核心过程是情景评估和预测，它将产生对现状的理解和表示，并在不久的将来对情况的发展趋势进行预测.

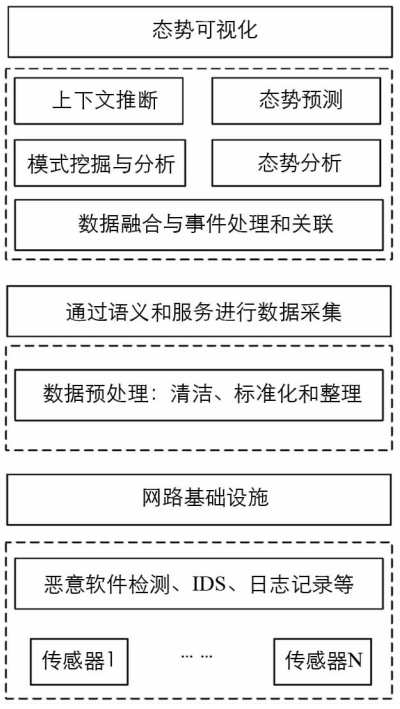

图 2 多层次 CSSA 分析框架

2.2 随机森林评估模型

随机森林是一种统计学习理论，由美国加利福尼亚大学伯克利分校创建. 随机森林的基本单位是决策树. 决策树模型往往不准确，容易出现过拟合，所以可以通过组合多个模型来提高预测的准确性. “Bagging”方法用于对决策树进行分组，其基本思想是通过使用名为 Bootstrap 的重采样方法并通过每个 Bootstrap 样本建模来从原始样本中提取多个样本. 然后，可以组合多个决策树来预测并通过投票得到最终的预测结果. 图 3 是随机森林流程图.

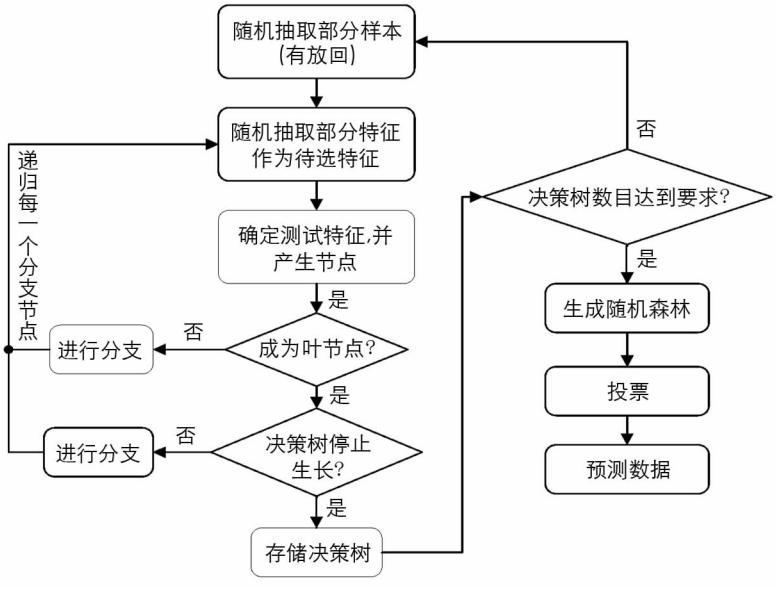

图 3 随机森林流程图

随机森林是本文提出的一种新的网络安全态势评估方法，其建模原则为：训练样本生成的每个决策树定义为 X，样本大小为 K，随机向量定义为 θ_k，随机向量的序列定义为 θ_k，$k=1, 2\cdots K$ 是独立同分布的，所有决策树分类器被定义为 $h(X, \theta_k)$，每个决策三元模型通过投票选择输入变量 x 的分类结果定义为

$$H(x)=\max_{Y}\sum_{i=1}^{k} I(h_i(x)=Y) \tag{1}$$

其中，$H(x)$ 是随机森林的分类结果，$h_i(x)$ 是单个决策树的分类结果，Y 是分类目标，$I(\cdot)$ 是指示函数.

随机森林模型的优点是收敛和广义误差上界. 对于一组分类器，余量函数用于度量平均正确分类数超过平均错误分类数的程度，余量值越大，分类预测越可靠，余量函数义为

$$mg(X, Y)=av_k I(h(X, \theta_k)=Y)-\max_{j\neq Y} av_k I(h(X, \theta_k)=j) \tag{2}$$

其中，$av_k(\cdot)$ 为求平均的函数，$I(\cdot)$ 是指示函数，$h(X, \theta_k)$ 为分类模型的序列.

根据简单投票的特点，定义泛化误差.

$$PE^* = P_{X,Y}(mg(X, Y) < 0) \tag{3}$$

随着随机森林中决策树数量的增加，所有序列 θ_1，…θ_k，PE^* 几乎处处收敛，泛化误差的收敛性为

$$\lim_{k\to\infty} PE^* = P_{X,Y}\begin{pmatrix} P_\theta(h(X, \theta_k) = Y) \\ -\max_{j\neq Y} P_\theta(h(X, \theta_k) = j) < 0 \end{pmatrix} \tag{4}$$

这表明该模型不会随着决策树的增加而产生过度拟合问题.

在广义误差的上界中，切比雪夫不平等为

$$PE^* \leqslant \frac{\mathrm{var}_{X,Y}(mg(X, Y))}{E_{X,Y} mg\ (X, Y)^2} \tag{5}$$

根据式(5)，我们定义了单个决策树的分类强度.

$$s = E_{X,Y} mg(X, Y) \tag{6}$$

则泛化误差的上界函数为

$$PE^* \leqslant \frac{\bar{\rho}(1 - s^2)}{s^2} \tag{7}$$

其中，$\bar{\rho}$ 表示决策树之间的相关性越小，强度越大，模型越精确.

3　CSSA 评估结果与分析

在本文中，选取了中国国家互联网应急中心(CNIEC)的数据，包括 2012 －2015 年间的 140 周报告形成的实验数据，其中包括 5 个病毒特征：受感染病毒的主机数量、被绑架篡改站点数量、被绑架在后边的跨境站点数量、跨境网站中的钓鱼页面数量以及新的信息安全漏洞数量.

通过参考网络威胁和漏洞等要素，结合国家突发公共事件总体应急预案，将网络安全态势安全等级分成 5 级：安全、轻度危险、一般危险、中度危险和高度危险，具体内容见表 1.

表 1　网络安全态势感知等级表

级别	威胁性	脆弱性	容灾性	稳定性
安全	低	低	高	高
轻度危险	中	低/中	高/中	高
一般危险	高/中	高/中	中	中
中度危险	高	中/高	中/低	低
高度危险	高	高	低	低

表 1 中安全等级的权重用区间[0，1]数值进行定量描述，如表 2 所示，方便对模型进行评估与分析.

表 2　网络安全态势对应表

级别	威胁指数	脆弱指数	容灾指数	稳定指数
低	[0，0.4)	[0，0.3)	[0，0.4)	[0，0.3)
中	[0.4，0.7)	[0.3，0.6)	[0.5，0.8)	[0.3，0.7)
高	[0.7，1.0]	[0.6，1.0]	[0.8，1.0]	[0.7，1.0]

对于安全态势指标评估，本实验中测试样本与训练样本比为 110∶40，实验中部分测试样本的网络输出与期望输出的比较结果见表 3.

表 3　模型评估的实际输出与期望输出

样本编号	实际输出	期望输出	实际威胁级别	期望威胁级别
86	0.87	0.85	高	高
87	0.92	0.89	高	高
88	0.78	0.83	中	中
89	0.65	0.68	中	中
90	0.59	0.56	中	中
96	0.25	0.29	低	低
97	0.47	0.41	低	低

以上实验表明，用随机森林评估模型进行网络安全态势评估，110个训练集和40个测试集都与实际输出结果较为接近.

另外采用了时间、kappa统计、平均绝对误差、均方根误差和TPR(真阳性率)来评估本文模型.

$$TRP = \frac{TP}{TP + FN} \tag{8}$$

其中，TP表示把正类预测为正类，FN表示把正类预测为负类.

从表4中数据可以看出，本文基于随机森林的多层次CSSA模型评估方法，在准确性、绝对误差和预测速度等性能方面都优于贝叶斯网络与BP神经网络方法，说明本文方法的有效性.

表4 不同算法下CSSA的评估结果对比

性能指标	贝叶斯网络	BP神经网络	随机森林
TPR/%	75.00	87.52	97.50
Kappa统计平均值	0.230 8	0.387	0.931
绝对误差	0.150 1	0.095	0.057 8
均方根误差	0.301 9	0.210 3	0.138 2
所用时间/s	0.05	0.05	0.03

采用本文与其他方法对网络安全态势进行预测，具体结果见图4.

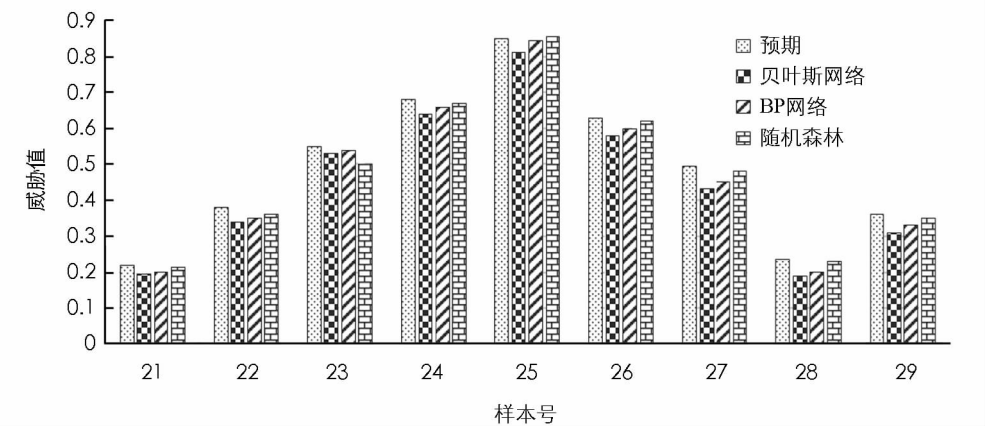

图4 网络安全态势感知预测结果

可以从图中看出，本文方法在网络安全态势感知的预测性能更接近真实值，优于其他两种方法的预测，因此采用随机森林进行多层次安全态势感知预测是有效的.

4 结 论

本文构建了一个多层次网络安全态势感知模型，并提出了基于随机森林的态势感知模型评估方法，能够实现快速准确的态势预测. 本文构建的多层次CSSA模型将CSSA过程与安全数据生命周期对齐，在多级分析框架中，数据采集和存储遵循分布式结构，每一种监控网络基础设施获取数据都要进行相应处理. 采用随机森林算法对多层次CSSA模型进行评估，将多个分类器组合，并依赖于独立样本，这使得网络安全态势评估更为客观和准确. 实验表明，本文模型和评估方法可行，在预测准确性、时间等性能上优于其他现有方法，能够提供选项来评估网络安全. 未来研究方向是对本文方法进行优化与改进，以便本文方法能够适合大规模复杂网络的模型评估.

参考文献：

[1] HOQUE N, BHUYAN M H, BAISHYA R C, et al. Network Attacks: Taxonomy, Tools and Systems [J]. Journal of Network and Computer Applications, 2014, 40: 307-324.

[2] 赵新杰，刘 渊，孙 剑，等. 基于迁移学习和D-S理论的网络异常检测 [J]. 计算机应用研究，2016，33(4)：1137-1140.

[3] 李 响. 基于经验模态分解的局域网络入侵检测算法 [J]. 西南师范大学学报(自然科学版)，2016，41(8)：132-137.

[4] EVESTI A, KANSTRÉN T, FRANTTI T. Cybersecurity Situational Awareness Taxonomy [C]//Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), 2017 International Conference On. London: IEEE, 2017.
[5] PARK H K, KIM M S, PARK M, et al. Cyber Situational Awareness Enhancement with Regular Expressions and an Evaluation Methodology [C]//Military Communications Conference (MILCOM), MILCOM 2017-2017 IEEE. Baltimore: IEEE, 2017.
[6] WEBB J, AHMAD A, MAYNARD S B, et al. A Situation Awareness Model for Information Security Risk Management [J]. Computers & security, 2014, 44: 1-15.
[7] 王　坤，邱　辉，杨豪璞. 基于攻击模式识别的网络安全态势评估方法 [J]. 计算机应用，2016，36(1)：194-198，226.
[8] KANSTRÉN T, EVESTI A. A Study on the State of Practice in Security Situational Awareness [C]// IEEE International Conference on Software Quality, Reliability and Security Companion. Vienna: IEEE, 2016.
[9] SINGH M, BHANDARI P. Building a Framework for Network Security Situation Awareness [C]//Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. New Delhi: IEEE, 2016.
[10] WU Z T, LIU J, XU S Y. A Cyberspace Security Situation Awareness Model Based on Complex Network [C]//Reliability, Maintainability and Safety (ICRMS), 2016 11th International Conference on. Hangzhou: IEEE, 2016.
[11] LI F W, ZHANG X Y, ZHU J, et al. A improved Network Security Situation Awareness Model [J]. EAI Endorsed Transactions on Security and Safety, 2015, 2(5): 1-5.
[12] 张勃帅，袁津生. 基于集对分析的 WSN 安全态势感知模型的研究 [J]. 计算机工程与科学，2017，39(3)：505-511.
[13] ZHU L, XIA G, ZHANG Z, et al. Multi-dimensional Network Security Situation Assessment [J]. International Journal of Security and Its Applications, 2016, 10(11): 153-164.
[14] 李　玺，卢　昱，刘　森，等. 基于 Markov game 模型的装备保障信息网络安全态势感知方法研究 [J]. 计算机应用研究，2017，34(11)：3441-3445.
[15] RASMUSSEN H B, LÜTZEN M, JENSEN S. Energy Efficiency at Sea: Knowledge, Communication, and Situational Awareness at Offshore Oil Supply and Wind Turbine Vessels [J]. Energy Research & Social Science, 2018, 44: 50-60.

Network Security Situation Awareness Framework and Random Forest Assessment Model

QIAN Zhen-kun

Logistics Service of Sichuan University of Arts and Science, Dazhou Sichuan 635000, China

Abstract: In view of the fact that traditional network security situational awareness assessment relies too much on expert experience, this paper proposes a multi-layer cyber security situational awareness (CSSA) framework and a network security situation assessment model based on random forest. In this method, the CSSA process has first been aligned with the security data lifecycle, the CSSA requirements analyzed, a CSSA multi-level analysis framework proposed, and then the random forest algorithm used to build the CSSA assessment model. This model is based on multiple classifiers. The idea of composition consists of a decision tree, each tree relies on independent samples, and the random vectors of all trees in the forest distribute the same value. When classifying, every tree voted and returned the class with the most votes, which made the network security situation assessment more objective and accurate. Experiments show that compared with Bayesian networks, this model can assess the current network security situation more quickly and accurately.

Key words: network security situational awareness; multi-level CSSA; random forest; decision tree; evaluation model

责任编辑　夏　娟