Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1368-1374
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116140

Hans汉斯

基于深度森林的网络安全态势评估方法

马嘉辉，汪　洋，黄玉妍

哈尔滨师范大学计算机科学与信息工程学院，黑龙江 哈尔滨

收稿日期：2022年11月29日；录用日期：2022年12月12日；发布日期：2022年12月28日

摘　要

本文针对当前网络安全态势评估方法中存在的态势要素样本数据少、评估准确性不足和模型训练耗时长等问题，提出了一种基于深度森林的网络安全态势评估方法。首先，在数据预处理阶段融合并量化多源获取到的态势要素数据，将原始态势要素样本信息转换为更适合深度森林中级联森林的有效特征。然后，将特征输入多粒度扫描模块，进行强表征特征提取。最后将特征向量输入特征优化后的级联森林模块逐层训练，完成网络安全态势评估。仿真结果表明，与其他传统网络安全态势评估模型相比，所提模型具有更高的精确率和召回率。

关键词

网络安全态势评估，深度森林，集成学习

Network Security Situation Assessment Based on Deep Forest

Jiahua Ma, Yang Wang, Yuyan Huang

College of Computer Science and Information Engineering, Harbin Normal University, Harbin Heilongjiang

Received: Nov. 29th, 2022; accepted: Dec. 12th, 2022; published: Dec. 28th, 2022

Abstract

In order to solve the problems existing in current network security situation assessment methods, such as fewer sample data of situation elements, insufficient assessment accuracy, and long training time of model, a network security situation assessment method based on deep forest is proposed. First, in the data preprocessing stage, the situation element data obtained from multiple sources are fused and quantified, and the original situation element sample information is converted into the effective features of the intermediate forest that are more suitable for the deep forest. Then, the features are input into the multi-granularity scanning module to extract strong

文章引用：马嘉辉，汪洋，黄玉妍．基于深度森林的网络安全态势评估方法[J]．软件工程与应用，2022，11(6): 1368-1374. DOI: 10.12677/sea.2022.116140

characterization features. Finally, the feature vector is input into the cascaded forest module after feature optimization for layer by layer training to complete the network security situation assessment. The simulation results show that the proposed model has higher accuracy and recall than other traditional network security situation assessment models.

Keywords

Network Security Situation Assessment, Deep Forest, Integrated Learning

1. 引言

随着互联网规模的不断扩大与应用，网络技术的不断发展，信息化的不断推进，互联网不断为各企业、政府和个人带来了便利，成为各个国家对推动不同领域高质量发展的重要基础设施。然而，各种网络攻击手段、安全漏洞以及安全事件也日益增多，且呈现越来越复杂和多样的趋势，不断威胁互联网安全，阻碍各领域信息化建设的高速发展，对保障网络安全提出了更为严峻的挑战。传统的网络安全技术，例如数据加密、数字签名、入侵检测、防火墙、访问控制和漏洞扫描等能够对互联网各领域起到一定的防护作用。但是，传统网络安全技术局限于某一特定网络领域，无法在宏观层面整体全面地反映当前的网络威胁状况。网络安全态势感知在宏观层面对当前网络的安全状态进行分析以及评估，能够改善传统网络安全技术的局限性以及被动性。网络安全态势评估作为网络安全态势感知的重要步骤，使用整体网络的全部安全要素进行捕获、整理、理解、分析、量化，最终得到当前网络的总体威胁情况以及安全态势[1]。

目前，国内外已有许多相关研究。Jin 等[2]提出了一种基于随机森林(Random Forest, RF)的网络安全态势评估模型。该模型将多个决策树组合成多分类器，每个决策树都依赖于独立的样本，并且森林中的所有决策树都具有相同的随机向量分布值。在进行网络安全态势要素的分类时，每棵决策树都要进行投票并返回投票最多的类，提高了网络安全态势评估的速度和准确性。张然[3]等提出了一种基于 SAA-SSA-BPNN 的网络安全态势评估模型。该模型利用模拟退火算法(Simulated Annealing Algorithm, SAA)优化麻雀搜索算法，解决麻雀搜索算法(Sparrow Search Algorithm, SSA)的不稳定性和容易陷入局部极值的问题，将其应用到 BP 神经网络(Back Propagation Neural Network, BPNN)进行改进，找到最佳适应度个体并取得最优权值和阈值，提高了网络安全态势评估的准确性和模型收敛速度。杨宏宇[4]等提出了一种基于并行特征提取网络(Parallel Feature Extraction Network, PFEN)和注意力机制改进双向门控循环单元(Attention-based Bidirectional Gated Recurrent Unit, ABiGRU)的深度学习网络安全态势评估方法。使用 PFEN 模块对不同网络威胁信息进行差异化提取并与原始信息相融合，利用 ABiGRU 模块的注意力机制对关键特征加权确性，提高了模型的精确率和召回率。吴海涛等[5]提出了一种基于 RBF-SVM 智能配变终端网络安全态势评估方法。将提取到的安全检测指标数据归一化，构建基于径向基函数(Radial Basis Function, RBF)的非线性支持向量机(Support Vector Machine, SVM)分类器，采用 k 折交叉验证和网格搜索法寻找分类器的最优参数，提高了网络安全态势评估的准确率。王金恒等[6]提出了一种基于遗传优化概率神经网络的网络安全态势评估。利用遗传算法优化概率神经网络的修正因子，使其避免了因网络安全态势参数细粒度评估而导致的收敛速度变慢问题，提高了概率神经网络的稳定性和训练速度，提高了网

络安全态势评估的准确度。李欣等[7]提出了一种改进隐马尔可夫模型的网络安全态势评估方法，利用人群搜索算法(Seeker Optimization Algorithm, SOA)随机搜索能力强的特点与 Baum-Welch (BW)参数优化算法相结合，解决传统参数优化算法产生的容易陷入局部极值最优问题，将优化后的参数代入隐马尔可夫模型(Hidden Markov Model, HMM)中，进行网络安全态势评估，提高了模型的准确率。

为了解决以上网络安全态势评估中出现的态势要素样本数据少、评估准确性不足和模型训练耗时长等问题，提出了一种基于深度森林的网络安全态势评估方法。首先，将获取到的多源态势要素数据融合量化为更适合深度森林中级联森林的有效特征。然后，将特征输入多粒度扫描模块，提取出具有强表征性的特征。最后将特征向量输入特征优化后的级联森林模块逐层训练，完成网络安全态势评估。

2. 深度森林

深度森林是一种基于决策树(Decision Tree, DT)的集成学习模型，由决策树森林级联构建的多层有监督学习模型。不同于传统深度神经网络模型，完整的深度森林包含多粒度扫描和级联森林两个相互独立的环节。深度森林基于非微分基学习器集成的深度学习模型，具有完全非神经网络模式的深度结构，能够有效避免传统神经网络出现的大量超参数、训练困难、耗时严重的问题。具体深度森林结构如图 1 所示。

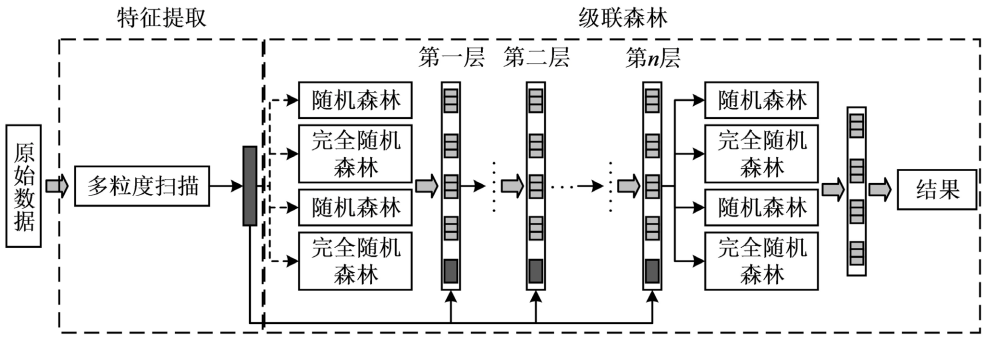

Figure 1. Structure diagram of deep forest
图 1. 深度森林结构图

2.1. 随机森林

决策树是一种树型结构的分类模型，决策树的每个节点表示一个属性，根据属性的划分，会依次进入节点的子节点，直至叶子节点，同时每个叶子节点都表示一定的类别，从而达到完成分类任务。随机森林(Random Forest, RF)是一种重要的集成学习模型，每个随机森林分类器均由多个决策树组成，其中每个决策树中的所有参数均独立同分布。对于每个输入样本，每颗决策树都会对其进行投票，最终将获得最高票数的类别决定为该样本的类别。随机森林使用 bootstrap 方法对训练集进行采样，在取样阶段，随机森林并不每次都选取信息增益率最大或基尼指数最小的特征，而是每次在当前节点 d 个特征的特征集中按比例随机选取一部分数量为 d' 的特征子集，并在其中找到最优解。其中，d' 取值如下：

$$d' = \log_2 d \tag{1}$$

随机森林使得每颗决策树都使用完全不同的训练集，有效避免了重复采样和样本未采样，差异化基分类器提高了模型的泛化能力和分类性能。完全随机森林(Completely Random Forest, CRF)是随机森林的一种极端情况，完全随机森林中的每棵树包含所有的特征，不选取最优特征值作为划分点，而是随机选择其中的一个特征值来划分决策树。

2.2. 多粒度扫描

多粒度扫描(Multi-Grained Scanning)环节为了增强级联森林，在前置阶段对特征做出一定的处理。与卷积神经网络(Convolutional Neural Networks, CNN)的滑动卷积核特征提取类似，多粒度扫描方法使用滑动窗口扫描原始特征，并生成输入特征。假设滑动窗口的尺寸为s，原始特征向量维度为M，经过窗口滑动转换为$(M-s)+1$维的特征向量，从相同大小的窗口中提取的特征向量训练随机森林和完全随机森林，获得类分布向量并将串联作为变换后的特征向量。重复以上步骤，最终获得多个增强特征向量。由于使用了时间窗口，多粒度扫描对时间序列的特征提取更为有效。

2.3. 级联森林

级联森林(Cascade forest)实现了特征信息在森林集合中逐级传递并处理的结构，将随机森林和完全随机森林作为基学习器构成一个级联层，并将变换的特征向量与原始特征相拼接作为下一级联层的训练数据，以减少随级联层堆叠深度加深而产生的特征向量信息退化问题以及在逐层传递特征过程中产生的过拟合风险。在每个级联层生成时都利用 k 折交叉验证检验模型的准确率，若准确率不再提升，则将该层视为模型的最后一级，最终得到输出结果。级联森林的层数可以根据数据集的规模来自适应确定。级联结构在不额外引入参数的同时增加了模型的深度并实现特征的逐层传递和处理，有利于模型的高效训练。

随着级联层堆叠深度的加深，深度森林模型的表征特征向量维度会逐渐变大，对分类结果可能产生影响，进而增加了模型的时间复杂度与时间开销。为了解决上述问题，使得模型更有效地进行训练的同时还能尽可能在特征传递过程中保留有用信息，本文将级联森林每层中上一级输出的变换类分布向量的平均值与原始向量相融合，作为本层的输入向量，进行特征优化。在充分考虑上级变换向量与原始向量影响的同时，有效解决表征特征向量随级联层堆叠深度加深而不断变大的问题。

3. 基于深度森林的网络安全态势评估

整个网络安全态势评估过程分为三个阶段，具体网络安全态势评估模型如图 2 所示。

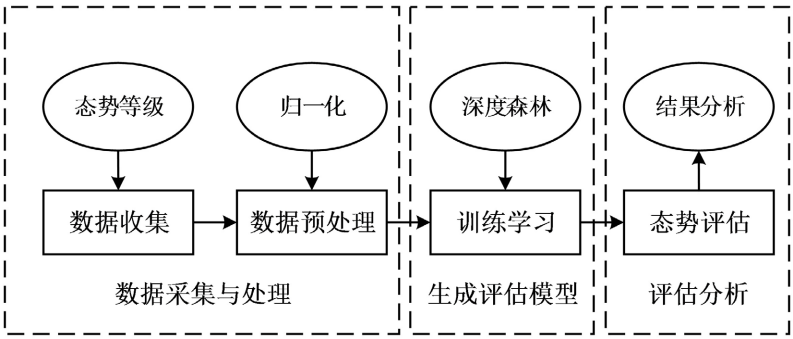

Figure 2. Network security situation assessment model based on deep forest
图 2. 基于深度森林的网络安全态势评模型图

3.1. 网络安全态势评估等级

建立合理的网络安全态势评估等级至关重要，其目的是将获取到的多源异构态势要素进行分类量化，映射成态势指数，最终分类到不同态势评估等级之中。根据具体网络安全态势要素数据的关联度与重复度，结合《国家网络安全事件应急预案》的总体安全需求，将网络安全态势要素数据分类量化为 5 个等级，具体网络安全态势评估等级如表 1 所示。

Table 1. Network security situation assessment level
表 1. 网络安全态势评估等级

态势等级	态势评价	态势指数	描述
1	安全	$[0.00,0.20]$	网络状况优秀，威胁几乎可忽略
2	低度危险	$(0.20,0.40]$	网络状况良好，威胁可能造成较小损失
3	比较危险	$(0.40,0.60]$	网络状况一般，威胁可能造成一定损失
4	中度危险	$(0.60,0.80]$	网络状况较差，威胁可能造成明显损失
5	高度危险	$(0.80,1.00]$	网络状况差，威胁可能造成严重损失

3.2. 数据预处理

在获取到网络安全态势要素数据后，通过归一化的方式对数据进行预处理。本文使用最小－最大规范化(Min-Max Normalization)对原始数据进行线性变换，将数据映射到同一区间内，可以表示为

$$X_i^* = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

其中：X_i 表示原始特征值，$X_{\min}$ 表示特征最小值，$X_{\max}$ 表示特征最大值，X_i^* 表示归一化后的值。

3.3. 网络安全态势评估

首先将输入的原始数据信号放入多粒度扫描环节中，利用滑动窗口提取数据特征，然后将提取到的数据特征放入随机森林和完全随机森林处理，将随机森林处理后的特征概率类分布向量进行拼接，得到高维增强特征向量。然后将通过多粒度扫描得到的特征向量输入级联森林逐级进行决策，将每级进行特征优化后的表征特征向量传递到下一层中，以此类推。接着将级联森林得到的增强特征向量先进行平均处理，并选取其中概率最大的类别，可获得每层对于测试集的准确率，并与上一层级联森林的结果进行对比。若结果对比相差较大，则说明级联森林未收敛，重复输入级联森林的步骤；若结果对比相差不大说明级联森林收敛，结束训练过程，并输出结果。

4. 仿真与分析

4.1. 仿真环境

仿真试验在某大学网络中心划分的局域网内进行。使用 Netflow，Snort 和 Nessus 等网络监测软件接入主服务器，获取当前服务器各项实时运行的检测、扫描和日志等信息。使用 Sigar 软件实时获取当前网络的平稳性指标数据。试验采用 Python 3.9.12、TensorFlow 2.8.0 和 Matlab R2019a 进行仿真，硬件环境为 Intel(R) Core(TM) i5-10400 CPU，16GB 内存，NVIDIA GeForce RTX 2070 8GB 显卡，操作系统为 Windows10 专业版 20H2 (19042.1466)。

4.2. 态势要素数据

本文使用 Netflow、Snort、Sigar 和 Nessus 等软件获取监控的流量数据、入侵检测日志和扫描分析后的日志等数据作为态势要素的数据来源。数据来源基本覆盖了数据流量、攻击威胁、网络稳定性和潜在系统漏洞等方面的网络安全信息，能较全面地反映当前网络的实时安全状态。数据样本时间跨度为 2021 年 5 月到 2021 年 6 月，样本数量为 12,000 个，训练集、验证集和测试集的比例分别为 70%、10%和 20%，总共包含 372 项安全事件。

4.3. 评价指标

精确率(Precision)描述了模型正确评估为相应态势等级的样本数量占模型全部评估为相应态势等级的样本数量的比例，可表示为

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

其中：TP 表示正确评估为相应态势等级的样本数量，FP 表示错误评估为相应态势等级的样本数量。

召回率(Recall)描述了模型正确评估为相应态势等级的样本数量占实际相应态势等级的样本数量的比例，可表示为

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

其中：FN 表示错误评估为不对应态势等级的样本数量。

F-score 综合考虑了精确率和召回率，可表示为

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{5}$$

精确率、召回率和 F-score 越高则说明网络安全态势评估越准确。

4.4. 结果与分析

分别将本文深度森林模型 DF 与卷积神经网络模型 CNN 和随机森林模型 RF 进行对比，各模型的精确率和召回率结果对比如表 2 所示。

Table 2. System resulting data of standard experiment
表 2. 标准试验系统结果数据

模型	精确率/%	召回率/%	F-score
RF	82.19	82.35	82.27
CNN	84.38	84.41	84.39
DF	85.77	86.03	85.90

由表 2 仿真结果显示，基于 DF 的网络安全态势评估模型的精确率、召回率和 F-score 均高于其他传统模型。对比 RF 模型和 CNN 模型，DF 模型的精确率分别提高了 4.4%和 1.6%，召回率分别提高了 4.5%和 1.9%，F-score 分别提高了 4.4%和 1.8%。说明深度森林模型具有更高的网络安全态势评估准确性。

5. 总结

本文提出一种基于深度森林的网络安全态势评估方法，将获取到的多源网络态势要素数据进行处理，输入多粒度扫描模块进行强表征特征提取，然后将特征向量输入特征优化后的级联森林模块中进行逐层训练，最后完成网络安全态势评估。仿真结果表明，与其他传统网络安全态势评估模型相比，DF 模型在多个指标上均优于其他对比模型，具有更优的网络安全态势评估性能。

基金项目

哈尔滨师范大学研究生教育教学改革项目“新工科背景下创新创业型研究生多维培养模式的研究——以网络安全方向研究生培养为例”和哈尔滨师范大学本科专业人才培养方案研究改革专项(XJGRYK2022012)。

参考文献

[1] 龚俭, 臧小东, 苏琪, 胡晓艳, 徐杰. 网络安全态势感知综述[J]. 软件学报, 2017, 28(4): 1010-1026. https://doi.org/10.13328/j.cnki.jos.005142

[2] Jin, Y., Shen, Y., Zhang, G., *et al*. (2016) The Model of Network Security Situation Assessment Based on Random Forest. 2016 7*th IEEE International Conference on Software Engineering and Service Science* (*ICSESS*), Beijing, 26-28 August 2016, 977-980. https://ieeexplore.ieee.org/document/7883229/

[3] 张然, 潘芷涵, 尹毅峰, 蔡增玉. 基于 SAA-SSA-BPNN 的网络安全态势评估模型[J]. 计算机工程与应用, 2022, 58(11): 117-124.

[4] 杨宏宇, 张梓锌, 张良. 基于并行特征提取和改进 BiGRU 的网络安全态势评估[J]. 清华大学学报(自然科学版), 2022, 62(5): 842-848. https://doi.org/10.16511/j.cnki.qhdxxb.2022.22.006

[5] 吴海涛, 代尚林, 乔中伟, 梁皓澜, 曾祥君, 刘东奇. 基于 RBF-SVM 智能配变终端的网络安全态势评估[J]. 电力科学与技术学报, 2021, 36(5): 35-40. https://doi.org/10.19781/j.issn.1673-9140.2021.05.005

[6] 王金恒, 单志龙, 谭汉松, 王煜林. 基于遗传优化 PNN 神经网络的网络安全态势评估[J]. 计算机科学, 2021, 48(6): 338-342.

[7] 李欣, 段詠程. 基于改进隐马尔可夫模型的网络安全态势评估方法[J]. 计算机科学, 2020, 47(7): 287-291.