计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(6):60−69 [doi: 10.15888/j.cnki.csa.009133]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于平均特征重要性和集成学习的异常检测①

庄　锐[1,2], 张　浩[1,2]

[1](福州大学 计算机与大数据学院, 福州 350116)
[2](福建省网络计算与智能信息处理重点实验室, 福州 350116)
通信作者: 张　浩, E-mail: zhanghao@fzu.edu.cn

摘　要: 异常检测系统在网络空间安全中起着至关重要的作用, 为网络安全提供有效的保障. 对于复杂的网络流量信息, 传统的单一的分类器往往无法同时具备较高检测精确度和较强的泛化能力. 此外, 基于全特征的异常检测模型往往会受到冗余特征的干扰, 影响检测的效率和精度. 针对这些问题, 本文提出了一种基于平均特征重要性的特征选择和集成学习的模型, 选取决策树 (DT)、随机森林 (RF)、额外树 (ET) 作为基分类器, 建立投票集成模型, 并基于基尼系数计算基分类器的平均特征重要性进行特征选择. 在多个数据集上的实验评估结果表明, 本文提出的集成模型优于经典集成学习模型及其他著名异常检测集成模型. 且提出的基于平均特征重要性的特征选择方法可以使集成模型准确率平均进一步提升约 0.13%, 训练时间平均节省约 30%.

关键词: 网络入侵检测; 异常流量; 特征选择; 集成学习; 异常检测

引用格式: 庄锐,张浩.基于平均特征重要性和集成学习的异常检测.计算机系统应用,2023,32(6):60–69. http://www.c-s-a.org.cn/1003-3254/9133.html

Anomaly Detection Based on Average Feature Importance and Ensemble Learning

ZHUANG Rui[1,2], ZHANG Hao[1,2]

[1](College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China)
[2](Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China)

Abstract: Anomaly detection system plays a significant role in cyberspace security and provides effective protection for network security. Regarding complex network traffic information, the traditional single classifier is often unable to ensure high detection accuracy and strong generalization ability at the same time. In addition, the anomaly detection model based on full features is often disturbed by redundancy features, which affects the accuracy and efficiency of detection. To address these problems, this study proposes a feature selection and ensemble learning model based on average feature importance. The decision tree (DT), random forest (RF), and extra tree (ET) are selected as the base classifiers to establish a voting ensemble model, and the average feature importance of the base classifiers is calculated based on the Gini coefficient for feature selection. The experimental evaluation results on several datasets show that the proposed model is superior to the classical ensemble learning models and other well-known anomaly detection ensemble models. The proposed model can improve the accuracy of the ensemble model by about 0.13% and save about 30% of training time on average.

Key words: network intrusion detection; anomaly network traffic; feature selection; ensemble learning; anomaly detection

① 基金项目: 国家自然科学基金重点项目 (U1804263, U21A20472); 国家留学基金青年骨干教师出国研修项目; 福建省自然科学基金 (2021J01616, 2020J01130167, 2021J01625)
收稿时间: 2022-11-10; 修改时间: 2022-12-10, 2023-01-06; 采用时间: 2023-01-19; csa 在线出版时间: 2023-04-14
CNKI 网络首发时间: 2023-04-18

随着物联网、大数据和人工智能的发展, 应用服务和数据规模迅速增加, 由此带来的网络安全问题也日益突出, 网络攻击者通常使用软件的混乱和漏洞来攻击网络计算机系统[1], 给我们的网络带来了许多潜在威胁. 入侵检测系统可以通过收集和分析网络数据传输中的信息来检测异常行为[2], 及时发现网络异常流量并进行处理是降低网络攻击危害的有效手段之一. 然而, 随着网络流量不断更新更迭, 其数据规模和数据维度一直在增大, 网络攻击行为也变得更加隐蔽. 攻击者可以在正常网络流量中隐藏攻击, 或将攻击伪装为正常流量. 因此, 入侵检测算法如何高效并准确地从大规模高维度的网络流量中检测异常数据, 一直以来都是网络安全领域研究的热点与挑战.

在当前, 存在着很多较为完备的网络异常检测手段, 其中, 比较成熟的主流网络异常流量检测方案主要包含基于统计的检测方法和基于机器学习的检测方法等. 其中, 统计方法检测速度快, 具有高效率等众多优点, 能够明确地表示和处理数据的异常, 也不需要构建过多的流量特征. 但是其检测准确率较低, 且并非所有异常行为都可以使用统计模型来表达. 对于某些隐藏的攻击方法, 无法通过统计方法来建立相应的行为模型. 基于机器学习的异常检测方法在一定程度上解决了这个问题, 大部分工作提升了检测的准确率并降低了误报率. 近年来, 基于机器学习的异常检测方法已经成为构建异常流量检测系统的重要方法.

其中, 传统的机器学习改进算法广泛应用于异常检测, 包括聚类算法[3], 支持向量机[4], 随机森林[5], 决策树[6], K 近邻[7] 等, 但是, 基于传统机器学习的异常检测过程中需要构建较多的流量特征和参数等, 在一定程度上影响了模型的精度和泛化能力.

作为机器学习的分支, 深度学习可以通过样本的潜在分布规律和表现层次来构建网络, 可以根据问题自动建立模型, 不需要人工进行特征选择, 在异常流量检测领域得到了广泛的应用并取得了不错的效果, 例如基于神经网络的方法[8,9]. 但是, 深度学习的模型是难以解释的, 也存在着模型结构复杂, 训练模型所需时间长, 易于陷入局部最优等问题, 面对流量数据巨大的网络攻击, 难以做到检测的实时性和高效性.

另一方面, 传统的单一的分类器往往不能同时具有较高检测准确率和较强的泛化能力, 而集成学习则能够将多个分类模型结合起来, 形成一个强学习器, 能够同时降低模型预测的偏差和方差, 提升模型的性能[10], 可以更好地融合到实际中去解决一系列异常检测问题. 与依赖单个分类器的现有方法相比, 集成学习能够在只增加有限计算开销的同时, 表现出优异的性能和鲁棒性, 在异常检测领域得到了广泛的应用[11].

此外, 网络空间数据流包含大量的时间、空间、负载和统计信息[12], 它可能包含一些不完整或冗余的信息. 异常可能只存在于几个相关的特征维度中, 噪声数据可能会严重混淆分类模型. 在异常流量检测领域中, 处理这样大量且多样的特征可能会导致以下问题: 1) 过多的特征需要漫长的训练过程, 而预测精度没有明显提高, 甚至有所下降. 2) 某些特征可能会在分类过程中引入偏差, 特别是那些与要分类的数据流量相关性较小的特征[13]. 因此, 基于全特征的异常检测模型往往无法捕获隐藏在局部数据特征中的异常信息, 如何通过合适的特征选择方法筛选出相关性较高的特征, 也成为当前研究的重点与挑战.

针对这些问题, 本文提出一种基于平均特征重要性的特征选择和集成学习的模型, 以解决异常检测问题. 选取决策树、随机森林、额外树作为基学习器, 建立投票集成模型, 并基于基尼系数计算基学习器的平均特征重要性进行特征选择, 去除相关性较低的特征, 以提高检测的精度, 节省检测时间. 在 CICIDS2017、UNSW-NB15、MIX 数据集上的实验评估结果表明, 本文提出的集成模型优于常见的经典机器学习模型、经典集成学习模型及最新异常检测集成模型. 此外, 采用基于平均特征重要性的特征选择方法后, 本文所提模型的检测精度和检测效率都有所提升.

本文的主要贡献如下.

(1) 提出一种基于平均特征重要性的特征选择方法, 与常规的只基于一种模型的相关性或重要性的特征选择方法不同, 本文提出的特征选择方法基于集成学习的多个基学习器计算特征的平均重要性, 能够更加全面地反映出特征与集成学习中所有基学习器的关系, 从而准确筛选出对每个基学习器都相对重要的特征, 解决了单一模型特征选择在集成学习中的局限性.

(2) 考虑树模型的性能及可并行特点, 选取决策树、随机森林、额外树作为基学习器, 建立加权软投票集成模型, 在多个数据上的评估显示, 该集成模型优于堆栈集成模型、经典集成模型及最新的集成模型.

(3) 将基于平均特征重要性的特征选择方法与基于树的集成学习模型相结合, 应用于异常流量检测, 得到了兼具检测精确度和检测效率的检测模型.

1 研究现状

1.1 特征工程

机器学习领域中, 特征工程是前期工作的重中之重, 为了减少特征空间并仅保留最重要的特征, 特征选择成为异常检测中的关键预处理步骤, 特征选择的效果很大程度上决定着模型检测结果.

在特征选择中, 若特征维度过大, 可能会出现过拟合的现象, 影响模型的精度和训练的效率. 若特征维度太少则不足以很好地区分类别之间的差异性, 容易导致欠拟合的现象, 使得检测过程中漏报现象严重.

为了解决上述问题, 学者们提出了许多具有创新性的特征选择方法. 其中, Al-Yaseen 等人[14] 提出了一种 Wrapper 特征选择方法, 采用差分评估算法来选择有用的特征, 以提高模型的性能并缩短其处理时间; Yang 等人[15] 通过基于信息增益和快速相关滤波的特征选择方法对数据集进行处理, 去除不相关和冗余的特征, 然后传递给核主成分分析模型, 进一步降维和降噪特征; 刘新倩等人[16] 采用基于递归消除特征算法选择特征, 获得理想的特征子集, 提升入侵检测模型的性能; Li 等人[17] 提出了一种使用多卷积神经网络的深度学习方法, 根据相关性将特征数据分为 4 部分, 并对数据集的不同部分使用相同的 CNN 结构. Demir 等人[18] 采用随机森林进行特征选择, 并结合多个特征子集生成 100 个模型. 根据准确性、信息增益和召回率 3 种评估方法, 将模型缩减为 10 个, 并将优化后的特征模型作为第 2 层的输入. 他们的研究可以取得良好的检测效果, 但中间特征选择需要花费大量的训练时间.

综上所述, 特征选择在异常检测领域中的运用广泛, 取得了显著的成果. 因此, 在异常检测的研究中, 通过合适的特征工程提高模型的性能是十分重要的.

1.2 集成学习

与传统的依赖单一分类器的方法相比, 集成学习能够将多个分类模型结合起来, 从而形成强学习器, 进一步提高模型的精度, 具有优异的性能和鲁棒性. 此外, 集成学习的合并策略通常很简单, 因此集成多个基础学习器通常仅花费非常低的计算成本. 集成学习算法会训练多个基学习器并将多个基学习器的分类结果合并起来. 通常一个结合了多个分类器的集成学习模型会比其单一的基分类模型性能更好, 泛化能力更强. 为了使集成模型的集成效果更好, 每个基分类器应该尽可能准确, 同时尽可能最大化其多样性.

近年来, 学者们研究了很多创新性的集成学习方法. Yang 等人[19] 提出了一种新的集成策略, 与许多集成技术使用的预定义或静态权重不同, 该模型根据基础学习器的实时性能为其分配动态权重. Dutta 等人[20] 提出了堆叠神经网络的模型, 以提高传统机器学习方法在网络异常检测中的效率. Olasehinde 等人[21] 提出了一种具有两个元学习器的堆栈集成方法. 第 1 个元学习器基于经典堆叠集成, 第 2 个元学习器基于多特征堆叠集成. 改进的堆栈模型优化了元特征组合的选择.

此外, 在异常检测领域, 也有很多学者提出了特征工程与集成学习结合的相关创新工作, 其中, Yang 等人[22] 提出一种基于特征重要性的堆栈集成方法, 通过计算基学习器的特征重要性来进行特征选择, 以提高堆栈集成模型的效率; Zhang 等人[23] 提出了一种基于多维特征融合与堆栈集成的异常检测方法, 通过排列组合的数据特征融合方法, 突出不同特征之间的相互支持或互补关系, 提升堆栈集成模型的性能.

综上所述, 基于集成学习的异常检测方法具有较高的可行性及有效性, 并且能够通过良好的特征工程, 进一步提升集成学习模型的性能.

2 平均特征重要性和集成学习的检测模型

基于以上分析, 本文提出了一种基于平均特征重要性和集成学习的异常检测模型, 主要分为以下几个阶段. 首先, 根据数据的性质, 从原始数据或基准数据集中提取若干基本特征数据集. 其次, 如果数据集的类别不平衡, 则执行过采样减少其影响. 在下一阶段, 基于平均特征重要性进行特征选择, 去除相关性较低的特征, 以提高检测的精度, 节省检测时间. 随后, 输入到集成模型对数据进行分类评估.

本文设计的基于平均特征重要性和集成学习的算法流程如算法 1 所示.

算法 1. 基于平均特征重要性和集成学习的检测算法

1) 输入: 训练集$T=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)\}$.
2) 如果类别不平衡: 进行 SMOTE 过采样;
3) 训练基分类器 RF、DT、ET;
4) for j=1 to n: 计算基分类器的特征X_j的重要性$VIM_j^{(\mathrm{RF})}$、$VIM_j^{(\mathrm{DT})}$、$VIM_j^{(\mathrm{ET})}$;
5) for j=1 to n: 计算特征X_j的平均重要性:

$$VIM_j=\frac{VIM_j^{(\mathrm{RF})}+VIM_j^{(\mathrm{DT})}+VIM_j^{(\mathrm{ET})}}{3}$$

6) 依据特征重要性分数从高到低对特征进行排序;
7) 依次选中特征直至特征重要性分数达到所设阈值;
8) 以 RF、DT、ET 为基分类器建立投票集成模型;
9) 将特征筛选后的训练集输入集成模型进行训练;
10) 输出: 预测结果$P=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)\}$.

本文提出的异常检测方法流程如图 1 所示.

2.1 数据预处理

由于网络数据越来越呈现出高维性和大规模的特点, 异常检测系统的计算复杂度可能会急剧增加. 因此, 应对外部网络数据进行进一步的特征分析和处理. 收集到的流量数据将通过几个步骤的预处理, 以便更好地进行异常检测系统的开发. 首先, 对数据进行独热编码, 以便更合理地计算特征之间的距离. 另一方面, 机器学习训练通常对归一化的数据更有效, 因此, 将每个具有数值的特征归一化为 0.0 到 1.0 的范围. 归一化后的每个值可以表示为:

$$X_n = \frac{x - min}{max - min} \tag{1}$$

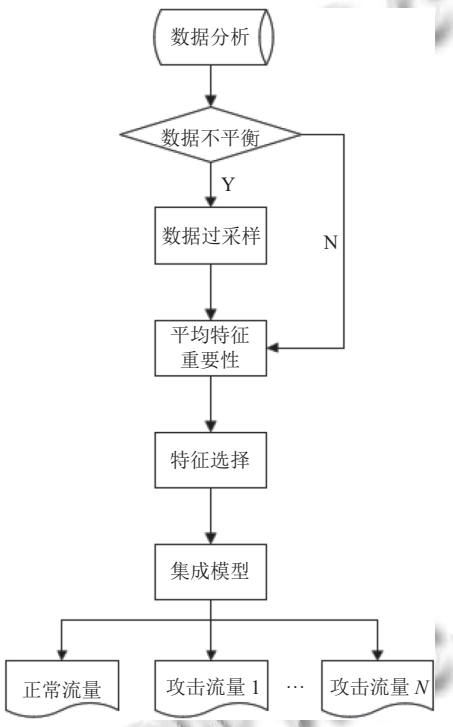

图 1 异常检测流程图

此外, 在现实生活中, 由于网络大部分时间保持正常状态, 且攻击标签实例往往不够多, 因此网络数据往往是类别不平衡的. 这通常会导致低异常检测率, 为了克服类别不平衡数据的问题, 可以使用随机过采样或者少数类合成过采样方法 (SMOTE)[24] 在没有足够数据的少数类别中生成更多数据. 随机过采样的基本策略是制作样本的多个副本以达到样本填充效果, 填充少数类中的数据. 但是, 随机过采样方法存在一定的弊端, 其学到的信息是非常具体的, 而不是一般的, 可能出现过拟合的现象. 而 SMOTE 技术通过 KNN 的思想, 对少数类进行分析, 在少数类样本间插入新样本. 因此, 该方法可以产生高质量的样本, 并用于本文提出的模型中的少数类. 图 2 为利用 SMOTE 算法生成新样本.

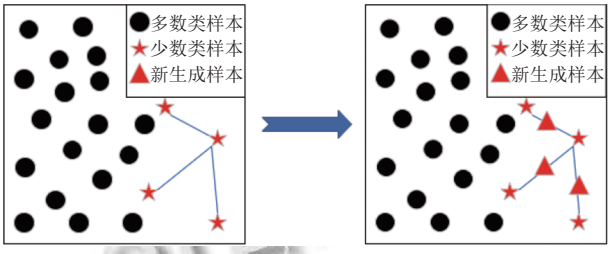

图 2 SMOTE 算法生成新样本

2.2 集成模型

为了检测各种网络攻击, 异常流量检测系统可以被视为一个分类问题. 各种不同的机器学习算法被广泛地应用于解决这些分类问题. 包括决策树 (DT)、额外树 (ET)、随机森林 (RF) 等. 决策树是一种基于分治策略的常用分类方法, 通过信息增益来选择划分属性. 随机森林是一种基于多数投票的集成学习分类器, 选择决策树中得票最高的类作为分类结果. 类似地, 额外树是另一种集成模型, 它处理不同数据集子集生成的随机决策树集合. 除了提出的算法, 其他经典模型例如最近邻 (KNN)、支持向量机 (SVM) 等也常用于异常检测问题.

为了选择合适的基分类器模型, 本文首先对常见的有监督机器学习算法的复杂度进行了近似计算和比较. 假设训练样本数量是 N, 特征数量是 M, 共有 T 颗树, 复杂度的近似计算如表 1 所示.

表 1 常见机器学习算法的近似复杂度

算法	复杂度
DT	$O(N^2M)$
RF	$O(N^2M)$
ET	$O(NMT)$
KNN	$O(N^2M)$
SVM	$O(N^2M)$

然而, 与 KNN 和 SVM 不同的是, 基于树的模型 DT、RF、ET 能够通过多线程来节省训练时间, 实际的计算复杂度将进一步降低.

因此, 本文选取决策树、随机森林、额外树 3 种基于树的算法作为基学习器, 建立投票集成模型, 选择的其他具体原因如下.

(1) 基于树的算法普遍检测能力优于基于概率的方法, 能够更好地处理所提出的网络数据中的非线性和高维数据[22].

(2) 基于树的模型是天生可并行的模型, 可以采用多线程进行计算, 因此计算复杂度较低.

(3) 大多数树结构的模型都使用了集成学习, 因此它们往往比其他单一模型表现出更好的性能.

(4) 基于树的模型在构建模型的过程中就进行了特征重要性的计算, 这有利于我们进行特征选择.

集成学习模型中, 投票集成学习是比较高效的一种集成方法, 只需要获得基学习器在测试集上的预测结果, 而不需要进行重复训练和预测.

投票集成学习算法分为硬投票集成学习算法和软投票集成学习算法. 其中, 硬投票集成学习算法直接以基学习器预测的结果进行投票, 往往会受到预测概率较低的结果的影响, 难以得到最好的预测结果. 而软投票集成学习算法采用的是加权融合方式, 其公式如下:

$$M = n_1 M_1 + n_2 M_2 + \cdots + n_k M_k \tag{2}$$

其中, M 为每个基学习器的预测分类概率, n 为每个基学习器的权重, $n_1 + n_2 + \cdots + n_k = 1$.

具体权重可以按照一定的规则分配到每个基学习器. 最后从具有最高平均概率的类别标签输出最终结果. 由于软投票会给使那些预测概率高的模型更高的权重, 表现往往优于硬投票方法集成方法.

因此, 本文选取决策树、随机森林、额外树 3 种基于树的算法作为基学习器, 建立软投票集成模型.

2.3 特征选择

在网络流量采集中, 收集到的数据中可能有大量的特征, 其中包含大量冗余的, 不相关的特征. 如何通过合适的特征选择方法筛选出对检测结果影响较大的那部分特征, 去除相关性较低的特征, 以提高检测的精度, 节省检测时间, 一直是机器学习领域特征工程中重点关注的问题. 因此, 本文以基尼系数 (Gini index) 作为评价指标, 分别计算选出的随机森林, 决策树和额外树 3 个基学习器的特征重要性分数, 并取其平均值, 从而选出对每个基学习器都相对重要程度高的那部分特征.

基于基尼系数计算特征重要性的思想在于计算各个特征在节点分裂前后基尼系数的改变量, 本文将特征重要性分数用VIM来表示, 将基尼系数用GI来表示, 其计算公式为:

$$GI_m = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^{K} p_{mk}^2 \tag{3}$$

其中, p_{mk}表示节点 m 中类别 k 所占的比例.

特征X_j在节点 m 的重要性, 以其分支前后GI的下降程度来表示, 计算公式为:

$$VIM_{jm}^{(\text{Gini})} = GI_m - GI_l - GI_r \tag{4}$$

其中, GI_l和GI_r分别表示分枝以后左子树和右子树的根节点的基尼系数.

如果特征X_j在树 i 中出现的节点为集合 M, 那么特征X_j在第 i 颗树的重要性为:

$$VIM_{im}^{(\text{Gini})} = \sum_{m \in M} VIM_{jm}^{(\text{Gini})} \tag{5}$$

假设有 n 颗树, 则特征X_j的重要性可以表示为:

$$VIM_{j}^{(\text{Gini})} = \sum_{i=1}^{n} VIM_{ij}^{(\text{Gini})} \tag{6}$$

对重要性评分做归一化处理, 使总的特征重要性为 1:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \tag{7}$$

为了提高所选特征的置信度, 本文采用集成特征选择技术, 分别计算对于随机森林, 决策树, 额外树 3 个基学习器的特征重要性, 以$VIM_j^{(\text{RF})}$、$VIM_j^{(\text{DT})}$、$VIM_j^{(\text{ET})}$表示.

为了筛选出对 3 种模型的重要性都较高的特征, 生成具有说服力的特征重要性, 对 3 种模型计算出来的特征重要性取平均值, 则特征X_j的平均重要性可以表示为:

$$VIM_j = \frac{VIM_j^{(\text{RF})} + VIM_j^{(\text{DT})} + VIM_j^{(\text{ET})}}{3} \tag{8}$$

最后, 在选择特征时, 根据重要性对特征进行排序, 每个特征从重要程度高到低依次选入到特征列表中, 直到选出的特征的重要性之和达到平均特征重要性阈值为止. 剩余的特征则认为是冗余特征, 将其舍去以减少计算成本.

3 实验分析

该系统使用 Python 3.5 实现, 并在一台 i7 处理器

和 8 GB 内存机器上进行实验, 在 Windows 10 上运行. 由于许多经典的流量异常检测数据集已经过时, 不能很好地适应当前的网络流量. 其中一些数据集没有包含较新型的异常数据类型, 不能满足数据的多样性, 以这些数据集来进行异常流量检测在一定程度上缺乏可靠性. 因此, 为了合理准确地评估本文模型的综合性能, 本文选择了 3 个较新的异常检测数据集进行实验, 其中包括两个经典的公共入侵检测数据集 CICIDS2017 和 UNSW-NB15, 以及一个混合了 CICDDoS2019 数据集部分数据和部分真实采集流量混合数据集 MIX. 其中, 实验所涉算法无需特定的超参数设定, 使用默认参数.

3.1 实验数据集

UNSW-NB15 数据集[25] 是一个使用广泛的异常流量检测数据集, 其通过搭建平台, 生成网络正常流量数据和各类网络攻击流量的混合数据集. 数据集包含九种攻击类型以及 49 个特征.

如表 2 所示, 训练集中约有 17 万条记录, 测试集中约有 8 万条记录, 包括正常数据和不同攻击类型的恶意攻击数据.

表 2 UNSW-NB15 数据集

类别	Normal	Attack	Total
Training set	56 000	119 341	175 341
Test set	37 000	45 332	82 332

CICIDS2017 数据集[26] 是一个使用广泛的入侵检测数据集, 包含正常网络流量数据和最新的恶意攻击流量, 接近真实网络流量数据. 如表 3 所示, 该数据集包含 2 830 743 条记录, 每条记录包含 78 个不同的特征及其标签.

表 3 CICIDS2017 数据集

攻击类型	Number	Percentage (%)
BENIGH	2 273 097	80.31
Dos	380 699	13.45
Port-Scan	158 930	5.61
Brute-Force	13 835	0.49
Web-Attack	2 180	0.08
Botnet	1 966	0.07
Infiltration	36	0.001

CICDDoS2019 数据集[27] 是最新的入侵检测数据集之一, 包含正常数据和最新的常见 DDoS 攻击, 类似于真实世界数据. 为了进一步评估本文提出的方法在真实流量上的表现, 本文抽样了一部分 CICDDoS2019 数据集的流量, 并加入了一部分实验室真实采集的流量数据, 将其整合成为的混合数据集 MIX, 每条数据包含 11 个特征, 该数据集的详情如表 4 所示.

表 4 MIX 数据集

类别	Normal	Attack	Total
Training set	1 453	5 915	7 368
Test set	3 582	47 069	50 651

3.2 评估指标

实验通过 4 个性能指标: 准确率 (*ACC*)、召回率 (*Recall*)、精确度 (*Precision*) 和 *F-measure* 对本文提出的方法进行了比较和评估.

准确率 (*ACC*) 是正确预测数据占总数据集比率:

$$ACC = \frac{TP+TN}{IP+TN+FP+FN} \tag{9}$$

精确度 (*Precision*) 是正确识别为正常的数据占所有正常数据的比例, 计算公式为:

$$Precision = \frac{TP}{TP+RP} \tag{10}$$

召回率 (*Recall*) 是正确识别为攻击类型的数据占所有攻击数据的比率, 计算公式为:

$$Recall = \frac{TP}{TP+RN} \tag{11}$$

F-measure 则是精确度和召回率的加权平均, 其计算公式为:

$$F\text{-}measure = \frac{2\times Recall\times Precision}{Recall+Precision} \tag{12}$$

其中, 真正率 (*TP*)、真负率 (*FP*)、假正率 (*TN*)、假负率 (*FN*) 的详细介绍如表 5 所示.

表 5 评价指标

指标	描述
TP	被正确识别的正常数据
TN	被正确识别的攻击数据
FP	被错误识别的攻击数据
FN	被错误识别的正常数据

3.3 实验结果

CICIDS2017、UNSW-NB15 和 MIX 数据集上评估不同算法的结果分别展示在表 6–表 8 中. 其中, 本文对 UNSW-NB15 和 MIX 数据集进行二分类评估, 对 CICIDS2017 进行多分类评估, 多分类的结果采用加权平均进行统计.

表 6 CICIDS2017 数据集全特征实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
SVM	89.17	88.90	89.17	88.72	68.616
KNN	98.78	98.81	98.78	98.79	2.844
RF	99.23	99.24	99.23	99.23	0.815
DT	99.59	99.59	99.59	99.59	1.062
ET	99.21	99.21	99.21	99.20	0.609
XGBoost	94.71	94.83	94.71	94.67	7.944
LightGBM	99.23	99.23	99.23	99.23	1.818
Proposed	**99.66**	**99.66**	**99.66**	**99.66**	2.469

表 7 UNSW-NB15 全特征实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
KNN	80.97	93.15	77.75	84.76	1.389
RF	90.03	**99.06**	86.17	92.17	1.239
DT	89.67	98.11	86.50	91.94	1.132
ET	89.54	98.92	85.56	91.76	0.499
XGBoost	89.87	96.83	**88.00**	92.20	1.441
LightGBM	89.89	98.87	86.12	92.06	0.717
Proposed	**90.06**	98.62	86.61	**92.23**	2.855

表 8 MIX 数据集全特征实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
KNN	99.17	90.17	99.13	94.44	2.923
RF	99.13	89.18	99.83	94.20	0.125
DT	99.92	99.25	99.64	99.44	0.078
ET	99.89	98.46	**99.97**	99.21	0.113
XGBoost	99.84	98.32	99.41	98.86	0.129
LightGBM	99.90	**99.37**	99.16	99.26	0.141
Proposed	**99.93**	99.09	99.89	**99.49**	0.202

如表 6–表 8 所示, 在分别对 3 种数据集进行测试时, 系统中使用的基于树的算法 DT、RF、ET 在上述 4 个精度指标中都显著优于 KNN、SVM 等经典机器学习模型. 此外, 由于基于树的模型 DT、RF、ET 是天生可并行的模型, 可以采用多线程进行计算, 因此与经典机器模型 KNN 和 SVM 相比, 执行时间也明显更短. 以上实验结果也证明了本文选择上述 3 种基于树的模型 DT、RF、ET 作为集成模型的基分类器具有合理性和正确性.

将 DT、RF、ET 三种基于树的模型作为基分类器, 建立投票集成模型进行实验. 实验结果表明, 该集成模型在大部分指标中达到最佳, 整体上优于常见的经典机器学习模型、经典集成学习模型以及基分类器 RF、DT、ET. 本文提出的集成模型具有良好的性能.

随后, 本文分别对 3 个数据集依据本文第 2.3 节的特征选择方法进行平均特征重要性计算及特征选择. 经过多次实验分析, 确定 UNSW-NB15 数据集的最佳特征重要性阈值为 0.8, CICIDS2017 和 MIX 数据集的最佳特征重要性阈值为 0.9. 对于 CICIDS2017 数据集, 本文从所有特征中选择了 38 个特征, 对于 UNSW-NB15 数据集, 从所有特征中选择了 17 个特征, 对于 MIX 数据集, 从所有特征中选择了 7 个特征. 具体筛选特征及其每条特征的平均特征重要性评分如图 3–图 5 所示.

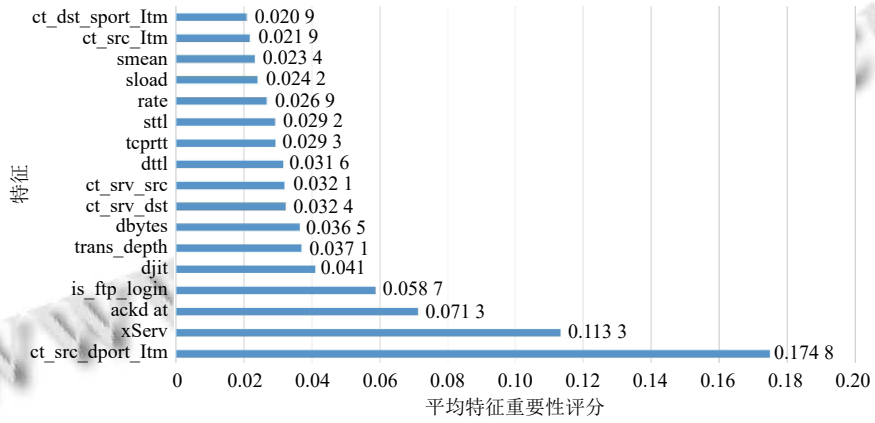

图 3 UNSW-NB15 数据集平均特征重要性

对于 CICIDS2017 数据集, 实验结果如表 9 所示, 基分类器 RF、DT、ET 以及集成模型的训练时间相比特征选择前分别缩短了 16.2%、56.2%、38.1% 和 39.65%. 同时, RF、ET 以及集成模型的精度都有了进一步的提升. 集成模型在各个指标上都达到了最高精度.

对于 UNSW-NB15 数据集, 实验结果如表 10 所, 基分类器 RF、DT、ET 以及集成模型的训练时间相比特征选择前分别缩短了 30.2%、40.7%、24% 和 33%. 同时, DT、RF 以及集成模型在多个指标上的表现有所提升. 集成模型在 *ACC* 和 *F-measure* 指标上达到最佳, 在 *Precision* 与 *Recall* 指标上, 也与最高精度相差无几.

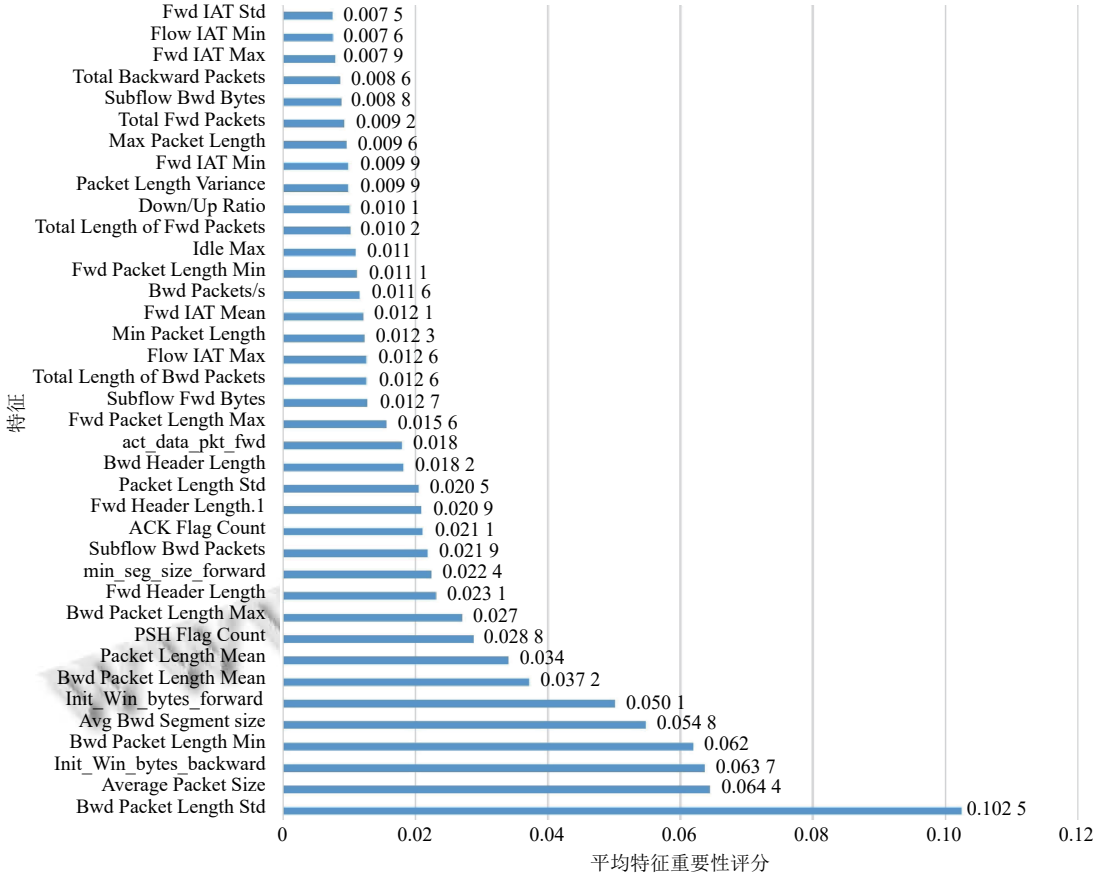

图 4 CICIDS2017 数据集平均特征重要性

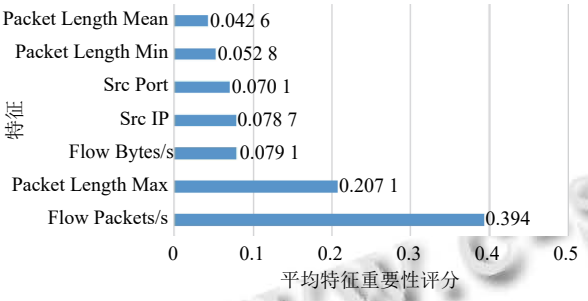

图 5 MIX 数据集平均特征重要性

表 9 CICIDS2017 特征选择后实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
RF	99.58	99.57	99.58	99.57	0.651
DT	99.59	99.59	99.59	99.59	0.500
ET	99.56	99.56	99.56	99.56	0.377
Proposed	**99.74**	**99.74**	**99.74**	**99.73**	1.490

对于混合的 MIX 数据集, 实验结果如表 11 所示, 基分类器 RF、DT、ET 以及集成模型的训练时间相比特征选择前分别缩短了 14.4%、17.9%、7% 和 18.8%, 其中, DT 和集成模型在多个指标上的表现相比特征选择前有了进一步的提升, 集成模型在 *ACC* 和 *Recall* 指标上达到了最佳. 在 *Precision* 与 *F-measure* 指标上也接近最高精度.

表 10 UNSW-NB15 特征选择后实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
RF	90.08	**97.83**	87.36	92.30	0.865
DT	90.26	96.76	**88.65**	92.53	0.671
ET	88.94	97.63	85.83	91.35	0.398
Proposed	**90.36**	97.23	88.36	**92.58**	1.914

表 11 MIX 数据集特征选择后实验结果

算法	*ACC* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)	Time (s)
RF	99.11	88.94	**99.89**	94.10	0.107
DT	99.93	**99.47**	99.66	**99.57**	0.064
ET	99.89	98.57	**99.89**	99.22	0.105
Proposed	**99.94**	99.22	**99.89**	99.55	0.164

应用平均特征重要性进行特征筛选后, 基分类器 RF、DT、ET 和本文提出的集成模型的精确度变化如图 6–图 8 所示. 对于 3 个不同的数据集, 在绝大多数模型下, 基于平均特征重要性的特征选择方法都能够明显提高模型的精确度.

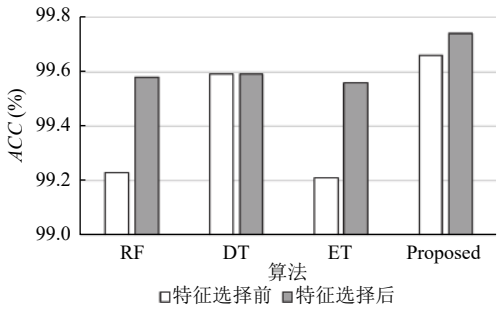

图 6 CICIDS2017 数据集特征选择前后精确度

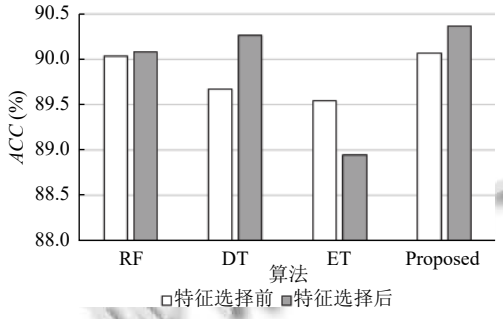

图 7 UNSW-NB15 数据集特征选择前后精确度

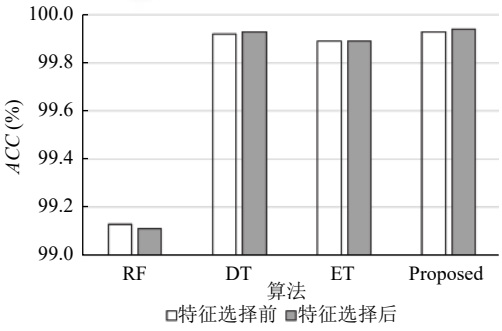

图 8 MIX 数据集特征选择前后精确度

为进一步探索所提出模型的新颖性, 表 12 展示了该模型与当前先进集成模型的性能比较. 包括基于 PCA 特征降维和随机森林的模型 (PCA-RF)[28]、基于树的特征重要性的堆栈集成模型 (Tree)[22]、基于多维特征融合和堆栈集成的模型 (MFFSEM)[23]. CICIDS2017 和 UNSW-NB15 数据集上评估不同算法的结果展示在表 12 中, 评估结果显示, 与目前较先进的集成学习方法相比, 本文提出的模型在各项指标上均有明显优势.

表 12 与当前先进集成学习方案的比较 (%)

数据集	方案	*ACC*	*Precision*	*Recall*	*F-measure*
CICIDS2017	PCA-RF[28]	99.60	99.60	99.60	99.60
	Tree[22]	99.62	99.62	99.62	99.62
	Proposed	**99.74**	**99.74**	**99.74**	**99.73**
UNSW-NB15	MFFSEM[23]	88.85	93.88	80.44	86.64
	Tree[22]	89.67	91.37	**89.67**	89.92
	Proposed	**90.36**	**97.23**	88.36	**92.58**

综合上述实验结果分析, 本文提出的特征选择方法能够在保持高精度的同时, 节省执行时间, 证明该特征选择方法具有有效性, 该投票集成模型具有良好的性能, 优于常见的经典机器学习模型 KNN、SVM、经典集成学习模型 XGBoost、LightGBM、基分类器 RF、DT、ET 以及当前先进的集成学习模型.

4 总结与展望

本文提出了一种基于平均特征重要性的特征选择和集成学习的模型, 用于检测网络入侵威胁. 为了评估提议的入侵检测系统, 在 3 个不同数据集中对其进行评估. CICIDS2017、UNSW-NB15 和 MIX 数据集上评估不同算法的结果表明, 该模型的精度优于常见的经典机器学习模型、经典集成学习模型及先进的异常检测集成模型. 此外, 基于平均特征重要性的特征选择方法能够明显节省模型训练时间, 并进一步提高模型的准确性. 特征选择后, 集成模型准确率分别进一步提高了 0.08%、0.30%、0.01%, 且训练时间分别减少了 39.65%、33% 和 18.8%. 因此, 本文所提基于平均特征重要性的特征选择和集成学习的模型能够有效识别异常流量, 为网络安全提供有效的保障.

在未来的工作中, 可以进一步挖掘不同维度的特征之间的相互支持或互补关系. 目前, 基本特征数据集都是在现有特征数据集的基础上构建的, 而网络的异常数据流可能来自多个不同的网络层. 可以尝试分别在不同的网络层上构建多维子空间数据集, 随后在各个不同的服务层中进行异常检测. 此外, 可以尝试通过粒子群优化或贝叶斯优化等优化方法来调整分类算法的超参数, 对模型进行一些优化调整, 进一步改进所提出的异常检测模型的检测效果.

参考文献

1 Moustafa N, Hu JK, Slay J. A holistic review of network anomaly detection systems: A comprehensive survey. Journal of Network and Computer Applications, 2019, 128: 33–55. [doi: 10.1016/j.jnca.2018.12.006]
2 Al S, Dener M. STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment. Computers & Security, 2021, 110: 102435.
3 Harush S, Meidan Y, Shabtai A. DeepStream: Autoencoder-based stream temporal clustering and anomaly detection. Computers & Security, 2021, 106: 102276.
4 Hiranai K, Kuramoto A, Seo A. Detection of anomalies in working posture during obstacle avoidance tasks using one-class support vector machine. Journal of Japan Industrial

Management Association, 2021, 72(2E): 125–133.
5 Subbiah S, Anbananthen KSM, Thangaraj S, *et al*. Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm. Journal of Communications and Networks, 2022, 24(2): 264–273. [doi: 10.23919/JCN.2022.000002]
6 Nancy P, Muthurajkumar S, Ganapathy S, *et al*. Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks. IET Communications, 2020, 14(5): 888–895. [doi: 10.1049/iet-com.2019.0172]
7 Liu GY, Zhao HQ, Fan F, *et al*. An enhanced intrusion detection model based on improved KNN in WSNs. Sensors, 2022, 22(4): 1407. [doi: 10.3390/s22041407]
8 Kan X, Fan YX, Fang ZJ, *et al*. A novel IoT network intrusion detection approach based on adaptive particle swarm optimization convolutional neural network. Information Sciences, 2021, 568: 147–162. [doi: 10.1016/j.ins.2021.03.060]
9 Lo WW, Layeghy S, Sarhan M, *et al*. E-GraphSAGE: A graph neural network based intrusion detection system for IoT. Proceedings of the NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium. Budapest: IEEE, 2022. 1–9.
10 徐晓芳, 管瑞. 基于神经网络集成学习算法的金融时间序列预测. 计算机系统应用, 2022, 31(6): 29–37. [doi: 10.15888/j.cnki.csa.008551]
11 Kumar G, Thakur K, Ayyagari MR. MLEsIDSs: Machine learning-based ensembles for intrusion detection systems—A review. The Journal of Supercomputing, 2020, 76(11): 8938–8971. [doi: 10.1007/s11227-020-03196-z]
12 Zimba A, Chen HS, Wang ZS, *et al*. Modeling and detection of the multi-stages of advanced persistent threats attacks based on semi-supervised learning and complex networks characteristics. Future Generation Computer Systems, 2020, 106: 501–517. [doi: 10.1016/j.future.2020.01.032]
13 Di Mauro M, Galatro G, Fortino G, *et al*. Supervised feature selection techniques in network intrusion detection: A critical review. Engineering Applications of Artificial Intelligence, 2021, 101: 104216. [doi: 10.1016/j.engappai.2021.104216]
14 Al-Yaseen WL, Idrees AK, Almasoudy FH. Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system. Pattern Recognition, 2022, 132: 108912. [doi: 10.1016/j.patcog.2022.108912]
15 Yang L, Moubayed A, Shami A. MTH-IDS: A multitiered hybrid intrusion detection system for Internet of vehicles. IEEE Internet of Things Journal, 2022, 9(1): 616–632. [doi: 10.1109/JIOT.2021.3084796]
16 刘新倩, 单纯, 任家东, 等. 基于流量异常分析多维优化的入侵检测方法. 信息安全学报, 2019, 4(1): 14–26.
17 Li YM, Xu YY, Liu Z, *et al*. Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. Measurement, 2020, 154: 107450. [doi: 10.1016/j.measurement.2019.107450]
18 Demir N, Dalkiliç G. Modified stacking ensemble approach to detect network intrusion. Turkish Journal of Electrical Engineering and Computer Sciences, 2018, 26(1): 418–433.
19 Yang L, Manias DM, Shami A. PWPAE: An ensemble framework for concept drift adaptation in IoT data streams. Proceedings of 2021 IEEE Global Communications Conference (GLOBECOM). Madrid: IEEE, 2021. 1–6.
20 Dutta V, Choraś M, Pawlicki M, *et al*. A deep learning ensemble for network anomaly and cyber-attack detection. Sensors, 2020, 20(16): 4583. [doi: 10.3390/s20164583]
21 Olasehinde OO, Johnson OV, Olayemi OC. Evaluation of selected meta learning algorithms for the prediction improvement of network intrusion detection system. Proceedings of the 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS). Ayobo: IEEE, 2020. 1–7.
22 Yang L, Moubayed A, Hamieh I, *et al*. Tree-based intelligent intrusion detection system in Internet of vehicles. Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa: IEEE, 2019. 1–6.
23 Zhang H, Li JL, Liu XM, *et al*. Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection. Future Generation Computer Systems, 2021, 122: 130–143. [doi: 10.1016/j.future.2021.03.024]
24 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]
25 Moustafa N, Slay J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). Proceedings of the 2015 Military Communications and Information Systems Conference. Canberra: IEEE, 2015. 1–6.
26 Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy. 2018, 1, 108–116.
27 Sharafaldin I, Lashkari AH, Hakak S, *et al*. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. Proceedings of the 2019 International Carnahan Conference on Security Technology. Chennai: IEEE, 2019. 1–8.
28 Abdulhammed R, Faezipour M, Musafer H, *et al*. Efficient network intrusion detection using PCA-based dimensionality reduction of features. Proceedings of the 2019 International Symposium on Networks, Computers and Communications (ISNCC). Istanbul: IEEE, 2019. 1–6.

(校对责编: 孙君艳)