Software Engineering and Applications 软件工程与应用, 2020, 9(4), 251-261
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.94028

Hans汉斯

Detection Method of HTTP-DDoS Attack Based on OPRFM

Caixia Guo, Fen Yan

School of Information Engineering, Yangzhou University, Yangzhou Jiangsu
Email: 13040281105@163.com

Received: Jul. 3rd, 2020; accepted: Jul. 17th, 2020; published: Jul. 24th, 2020

Abstract

DDoS attack is a major problem in the field of network security. This paper aims to improve the ability of detection and classification of DDoS attacks, proposes a new improved random forest algorithm, and on this basis, proposes an improved random forest classification model to detect DDoS attacks. The experiments show that compared with the random forest algorithm, decision tree algorithm and support vector machine algorithm, the proposed algorithm shows significant improvement in accuracy, recall rate and F1 value.

Keywords

Machine Learning, Random Forest, OPRFM, HTTP-DDoS

基于OPRFM的HTTP-DDoS攻击检测方法

郭彩霞，严　芬

扬州大学信息工程学院，江苏　扬州
Email: 13040281105@163.com

收稿日期：2020年7月3日；录用日期：2020年7月17日；发布日期：2020年7月24日

摘　要

DDoS攻击是网络安全领域面临的一个重大问题。本文旨在提高DDoS攻击检测分类的能力，提出新的改进的随机森林算法，并在此基础上提出一种检测DDoS攻击的改进随机森林分类模型。利用公开的CICIDS2017数据集进行验证，实验表明，与随机森林、决策树、支持向量机算法相比，提出的算法在精确率、召回率和F1值方面显示出显著的改善。

文章引用: 郭彩霞, 严芬. 基于 OPRFM 的 HTTP-DDoS 攻击检测方法[J]. 软件工程与应用, 2020, 9(4): 251-261.
DOI: 10.12677/sea.2020.94028

关键词

机器学习，随机森林，OPRFM模型，HTTP-DDoS

1. 引言

随着 DDoS 攻击的日益猖獗及其规模的不断增大、攻击方式的日渐复杂，特别是随着大数据及其相关技术的飞速发展，互联网上的网络流量已经出现了高度的复杂性以及访问流量的激增，给传统的 DDoS 攻击检测带来了十分严峻的考验，因此需要一个能应对目前大流量 DDoS 攻击的问题的检测方法。

随机森林(RF)作为一个性能良好的组合分类器算法，已经被广泛用于网络安全领域。文献[1]为了准确预测物联网系统的攻击和异常，比较了几种机器学习模型的性能。使用的机器学习算法有 Logistic 回归(LR)、支持向量机(SVM)、决策树(DT)、随机森林(RF)和人工神经网络(ANN)。分别使用准确性、精确性、召回率、F1 值和 AUC 作为性能比较的评估指标。实验证明，使用决策树、随机森林和人工神经网络的精度达到 99.4%，综合其他指标证明随机森林的性能相对较好。文献[2]中利用网络包的参数 HTTP-GET、POST-Request 和 Delta-Time 进行计算找出可能攻击的准确度，利用朴素贝叶斯、朴素贝叶斯多项式、多层感知、RBF 网络、随机森林等分类器对攻击生成的数据集进行分类，并且通过混淆矩阵，比较了各算法的准确率、真阳性率、假阳性率，实验证明随机森林分类器在各方面的性能存在着优势。文献[2]指出利用机器学习算法可以有效地捕获 DDoS 攻击流量所表现出的流量模式并对一些有监督的机器学习算法进行了评价和排序，性能评估使用多准则决策辅助软件 Visual Promethee，分析证明了基于集成的分类器的有效性。文献[3]综述了利用人工智能技术检测 DDoS 攻击的最新进展，提出了可用于检测 DDoS 攻击的特征，如数据包数量、数据包大小的平均值、时间间隔方差、数据包大小方差、字节数、数据包速率和比特率。在这些人工智能技术中，文献建议使用随机森林树对恶意流量和正常流量进行分类，以获得更好的性能。

随机森林具有自动特征分析的能力，其变量重要性可以作为高维数据处理的特征选择工具，并且基于多维特征的数据分类运行效率更高，实现起来相对简单同时对噪声鲁棒性很强。随机森林由大量的决策树组成，构建过程引入了随机操作过程，同时包括样本子集和特征子集的选择，能提高分类精度并且获得更好的泛化能力。但是传统的随机森林算法只考虑每个特征对于分类的重要性，由于发起 DDoS 攻击的攻击者产生的变量之间的相关性很强，使得随机森林选出的变量之间冗余性很大，这样会严重影响准确性。因此，本文针对随机森林这一不足进行改进，并在此基础上提出一种检测 DDoS 攻击的改进随机森林分类模型(OPRFM)，首先，采用随机森林算法对特征重要性进行排序和降维。其次，将所选特征与改进的随机森林算法相结合，将构建决策树过程中的度量值作为各决策树的权值，建立检测分类预测模型。

本文的主要工作为：

1) 针对 DDoS 攻击者流量特征变量的冗余性较大的问题，对随机森林算法进行改进，并且介绍了详细的改进过程。

2) 在此基础上，提出一种检测 HTTP-DDoS 攻击的改进随机森林分类模型(OPRFM)，并且在公开数据集上进行了验证。

2. 随机森林算法介绍

2.1. 算法分析

OPRFM 的实现是基于传统的随机森林(RF)算法。RF 是 Breiman [4]在 2001 年提出的一种集成算法，是目前在机器学习领域相当热门的集成学习技术之一。

随机森林是一种组合决策树模型，首先是利用 Bagging 方法将原始训练集中进行采样重组，然后再对重组后的数据利用随机特征抽取方法生成决策树，由多棵决策树形成森林。对于随机森林来说，森林中树与树之间的相关性越强，整个分类的性能就会相对较差。但是相比于单个的决策树来说，随机森林可以避免过拟合的现象，分类精度相对较高，稳定性也更好；同时，相比较于其他多分类器集成方法，随机森林在处理数据噪声方面来说更加稳定[5]。

从上述分析可得知，随机森林是通过两个“随机”机制来解决单一分类器容易发生的过拟合的现象。

1) 随机选择训练数据。随机森林从原始训练集数据中有放回地随机抽取样本生成训练数据，重新组合的训练集是原始训练数据集中的子集，因为新的训练子集各不相同，所以，森林中的决策树的生长过程存在随机性。

2) 随机选择训练数据的特征属性变量。随机森林在决策树的构建过程中，不进行任何的剪枝操作，除此之外，也增加了对特征属性变量选择的随机性。由于各决策树有不同的生长过程，因此，增加了随机森林的又一个随机因素。

2.2. 算法特点

1) 能够计算特征对于分类目标的重要性，并且由于单个决策树的生长速度快，所以总体的分类速度就快，对于大样本数据能够高效的处理。

2) 对噪声数据有较强的健壮性。

3) 通过在结点处随机选择特征进行分类，能够提高分类精度，同时有能很好的解决过拟合问题。

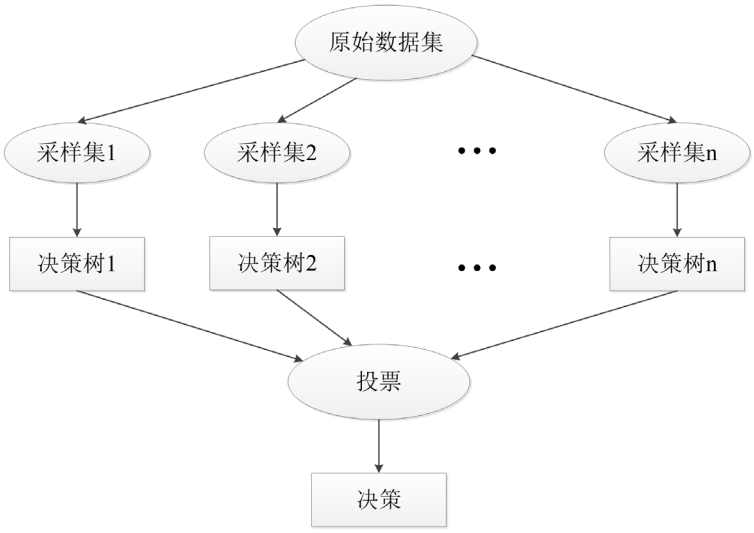


Figure 1. Frame diagram of random forest classifier
图 1. 随机森林分类器的框架

如图 1 所示，以下是建立随机森林分类器的主要步骤：

1) 选取训练集：每棵决策树对应一个训练集，从原始训练数据集中随机抽取 N 个训练子集构建 N 棵决策树，剩余的未被抽取到的数据则作为测试集。

2) 构建每棵决策树：随机森林随机的从所有特征中随机选择一定数量的特征构建子树而不使用所有的特征，说明随机森林可以处理很高维度的数据，然后根据某种策略进行树的分裂。

3) 多数投票：输入到随机森林里的样本进行整个森林的遍历，每个分类器都得出一个结果，最终由这多个结果进行多数投票得到。

3. 随机森林算法的改进

一般的随机森林算法只是简单地将所有的子决策树组合在一起，所有的决策树在进行分类投票时具有相同的权重值，然而发起 DDoS 的攻击者所控制的傀儡机之间产生的流量的相关性很强，造成特征之间的冗余性很大，这会使得真实的数据集上的冗余特征导致随机森林的性能降低。因此，我们对随机森林算法进行改进，通过对不同的决策树赋予不同的权值，同时对每个决策树赋予不同的权重。改进的随机森林算法结构如图 2 所示。

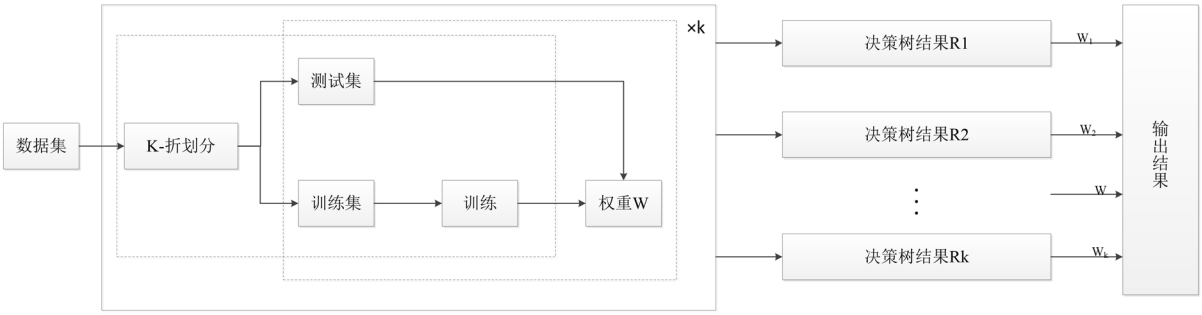


Figure 2. Structure diagram of improved random forest algorithm
图 2. 改进随机森林算法结构图

1) 在 K-折交叉验证的启发下，采用 K-折划分的方法，将原始数据集 D 划分为 K 个子集，如图 3 所示，然后，其中 $K-1$ 个子集作为训练集用来训练 n 个决策树，通常使用 CART 树或者 C4.5 树，最后一组数据则用来作为测试集，从而得到 n 个子决策树的分类准确率 P_n，$P_n=\{p_1,p_2,p_3,\cdots,p_n\}$。不断地重复上述步骤，直到每个子集在一次迭代中都被用作测试集，在每一次迭代中，我们利用公式计算每个子决策树的权重 w_i。

$$w_i = P_i - \min(P_n) \tag{1}$$

因此利用式(1)可以得到每个字决策树的累计权重为：

$$Sumw_i = \sum_{i=1}^{K} w_i \tag{2}$$

2) 然后将样本 x 通过随机森林分类器进行检测分类，然后进行加权统计，属于 C 类别的总得票数记为 L_c，则

$$L_c = \sum_{h=1}^{H} H_{c,x}(x) * Sumw_i \tag{3}$$

式(3)中，$H_{c,x}(x)$ 的取值根据样本 x 经过决策树的分类结果而定，若结果为 C 类，则取值为 1，否则取值为 0。

3) 经由以上步骤，选出得票数最多类别 C_x 作为样本 x 的最终类别

$$C_x = \arg\max\left(L_c\right) \tag{4}$$

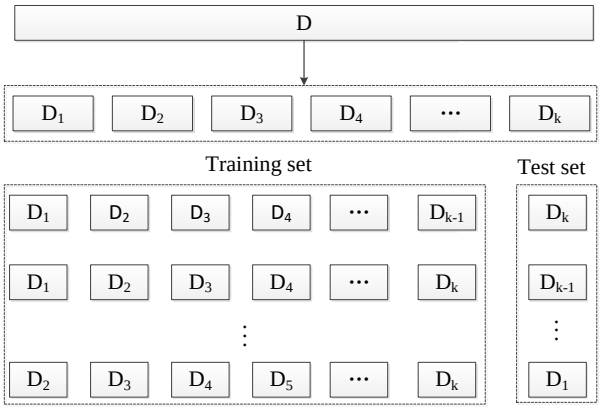


Figure 3. Data set segmentation graph [7]
图 3. 数据集分割图[7]

4. 基于 OPRFM 的 DDoS 攻击检测分类模型

4.1. 模型设计

针对 HTTP-DDoS 攻击，本文提出的改进随机森林分类模型(Optimized Random Forest Model, OPRFM)主要包括特征向量提取模块和检测分类模块两部分。其检测模型框架如图 4 所示。

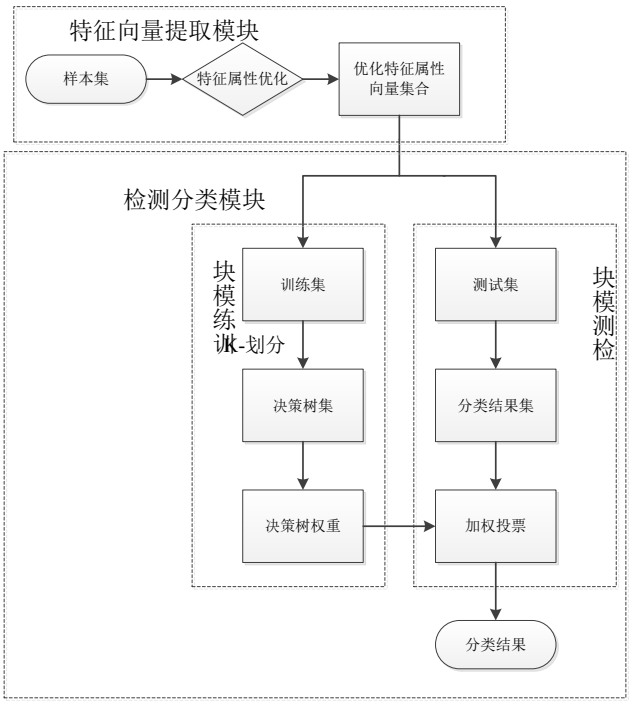


Figure 4. Framework diagram of improved random forest classification model
图 4. 改进随机森林分类模型框架图

1) 特征向量提取模块。将特征属性进行优化，生成优化特征属性向量集合，通过特征排序提取具有代表性的特征属性并生成特征向量集合，具体步骤如下：

① 使用当前特征集通过应用 RF 算法建立模型。

② 计算当前特征集中每个特征的变量的重要性得分 I，并根据特征变量的重要性得分 I 按降序排列每个变量。

③ 从当前特征变量中删除无意义的和不重要的特征以获取新的特征集。

采用随机森林排序法来衡量攻击者的特征，并减少攻击者预测的维度，有效的提高预测效率。

2) 检测分类模块。该模块使用改进的随机森林算法分类器，该分类器由训练子模块、检测分类子模块组成，旨在利用改进的方法生成改进随机森林分类器并对测试集进行检测分类。步骤如下：

① 将特征向量提取阶段生成的特征向量集合作为训练模块的输入。按比例抽取一部分数据作为待训练子集训练分类器并生成训练后的决策树集，将剩余数据作为测试数据。

② 决策树构造选择模块将训练数据集 D 划分为 K 个子集，然后，$K-1$ 个子集作为训练集用来训练 n 个决策树，最后一组数据作为测试数据，得出每棵决策树的分类正确率，每次迭代过程中得出每棵决策树的权重值，最后得出决策树的累计权重。

③ 检测分类模块将步骤①里的剩余数据作为训练后的决策树集的输入，得出分类结果集，同时使用训练模块得到的权重值进行加权投票，得出最终分类结果。

4.2. 特征向量提取

攻击者有许多特征，可分为两大类：流量特征、行为特征。不同的攻击规模会影响这些特征的集合。如果将所有的特征(维度)都考虑在内，这将导致算法在应用时在时间和空间上的巨大成本，进而对算法的性能产生严重的影响。因此，首先使用 RF 来减小数据的尺寸，提高模型检测的准确率，构造一个快速低消耗的模型。

RF 方法可以根据特征(变量)的重要性对特征(变量)进行排序，从而减少维数，删除不重要的特征。其核心思想是通过在每个特征中加入噪声来计算预测的精度下降程度。RF 算法中每个特征 x 的重要性计算如下：

1) 对于 RF 中的每个决策树，使用相应的袋外(OOB)数据来计算 OOB 误差记为 $errOOB1$，袋外(OOB)数据指的是在构建决策树的过程中未使用的其余数据。OOB 数据约占所有数据的三分之一，OOB 数据可以用来评价决策树的性能。

2) 对所有 OOB 数据在特征 x 中随机加入噪声干扰，然后再次计算 OOB 误差记为 $errOOB2$。

假设 RF 中有 N 棵树，特征 x 的重要得分可以通过以下公式计算：

$$I_x = \frac{\sum_{j=1}^{N}\left(errOOB2_j - errOOB1_j\right)}{N} \tag{5}$$

通过将噪声随机添加到特征中，袋外数据的精度显著降低，这说明该特征对样本的分类结果影响很大，即具有较高的重要性

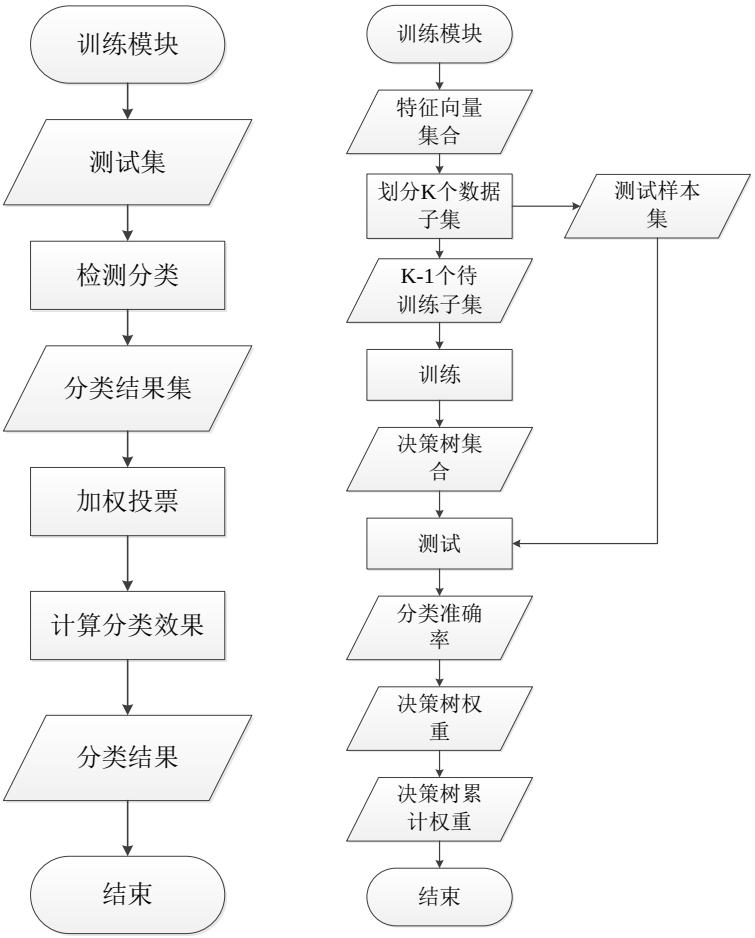
4.3. 检测分类

检测分类模块由两个子模块组成，分别是训练子模块和检测子模块。

1) 训练模块的作用是构造决策树集并计算相对应的权重值，具体过程为：将特征提取阶段生成的特征向量集合划分为 K 个子集，然后，$K-1$ 个子集作为训练集用来训练 n 个决策树，最后一组数据作为测试数据，得出每棵决策树的分类正确率，然后每次迭代得出每棵决策树的权重值并最终得出累计权重值。

训练模块流程如图 5(a)所示。

2) 检测模块的作用是对测试样本集进行检测分类并对分类效果进行评估。具体流程为：将测试集输入到决策树集合中得到分类结果，并且应用子决策树的累计权重对分类结果进行加权投票，得出最终分类结果，计算各值来评价分类器的分类效果。检测分类模块流程图如图 5(b)所示。



(b) 测试模块流程图

Figure 5. Flow chart of detection process
图 5. 检测过程流程图

5. 实验结果与分析

本小节，我们通过实验来验证所提出的用于 HTTP-DDoS 攻击检测的 OPRFM 模型在精确率、召回率以及 F1 值方面的优越性。我们实验的软件和硬件配置如表 1 所示。

Table 1. Experimental environment configuration
表 1. 实验环境配置

硬件配置	软件环境
Intel core i7-8750 CPU @2.20GHz	Windows10 x64 操作系统
8.0GB RAM	Anaconda3

5.1. 实验数据

实验数据集是 CICIDS2017，由加拿大网络安全研究所开发[6]，用于检测网络异常。该数据集包含良性和最新的常见攻击。该数据集基于各种应用层协议(如 HTTP，FTP，HTTPS 和 SSH)使用不同的可用工具生成 25 个用户的抽象行为，为期 5 天。最初，他们在周一捕获了正常的网络流量并标记为 Benign。从周二到周五，他们捕获了针对各种类型攻击的入侵流量，例如 DoS/DDoS，暴力攻击和 Web 攻击等。本章中，我们使用了包含 DDoS 攻击数据和良性数据的周五的数据集。数据被标记为两类，分别为 DDoS 攻击类和良性类，每条数据有 79 个特征项。

5.2. 特征向量提取

为了选取具尽可能多的无关特征属性，本文采取随机森林算法对特征属性进行优化排序。随机森林算法是通过在每个特征中加入噪声来计算预测的精度下降程度，从而计算特征(变量)的重要性值 I。采用上述算法的排序结果如表 2 所示。表 2 列出了 79 个特征属性经过排序后排名在前 22 的特征属性。

Table 2. Value of feature I
表 2. 特征 I 值

No.	Feature	I
1	Avg Fwd Segment size	0.067
2	Packet Length Variance	0.065
3	Packet Length Mean	0.056
4	Fwd Packet Length Mean	0.051
5	Min_seg_size_forward	0.047
6	Max Packets Length	0.03
7	Bwd Packet Length Std	0.027
8	Fwd Packet Length Std	0.027
9	Flow Bytes/s	0.026
10	Init_Win_bytes_backward	0.021
11	Fwd Packet Length Min	0.021
12	Fwd Packet Length Max	0.021
13	Avg Bwd Segment Size	0.020
14	Init_Win_bytes_forward	0.019
15	Bwd Packet Length Mean	0.019
16	Subflow Fwd Bytes	0.018
17	Subflow Bwd Packets	0.017
18	Min Packets Length	0.017
19	Total Length of Fwd Packets	0.017
20	Total Backward Packets	0.016
21	Bwd Header Length	0.014
22	Fwd Header Length	0.013

在特征向量提取阶段，将原始数据集的特征向量集合作为输入，选取不同数量的特征进行实验，实验结果图 6 所示。

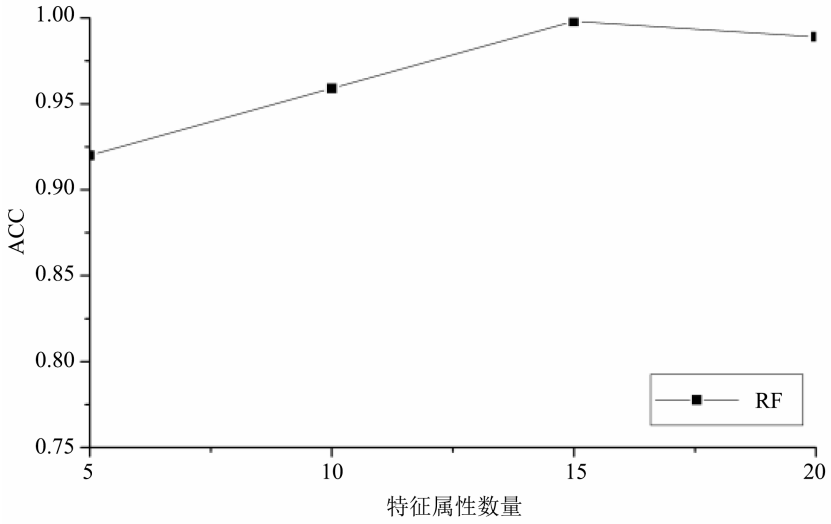


Figure 6. Relationship between the number of features and ACC
图 6. 特征数量与 ACC 关系图

由图 6 可知，在考虑如何选取特征属性的数量时，特征属性数量大于 15 后，ACC 值较高并趋于平稳。

5.3. 评价标准

为了评价所构造的分类器的分类效果，本文采用准确率(Precision，简记 P_r)、召回率(Recall，简记 R_c)及 F1 值作为实验的评价标准。公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$P_r = \frac{TP}{TP + FP} \tag{7}$$

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

$$F1 = \frac{2 * P_r * R_c}{P_r + R_c} \tag{9}$$

TP (true positive)，表示正样本被识别为正样本的比例。

FP (false positive)，表示正样本被识别为负样本的比例。

FN (false negative)，表示负样本被识别为正样本。

TN (true negative)，表示负样本被识别为负样本。

我们使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、*F*1 (F-Measure)值、训练时间和预测时间来对提出的改进随机森林算法分类模型进行评估。

5.4. 实验结果与分析

1) 实验结果

在改进的随机森林算法模型分类实验中，将 5.2 节中的实验结果生成的特征向量数量为 15 个的特征向量集合作为输入。实验结果如表 3 所示。

Table 3. Improved random forest algorithm experiment result table
表 3. 改进随机森林算法实验结果表

	Precision	Recall	F1-score
BENIGN	1	0.99	0.99
DDoS	0.99	0.98	0.97
Avg/total	0.99	0.985	0.98

实验结果由表 3 可知，改进后的随机森林分类器具有较好的检测分类效果，总精度达到 99%，对于正常用户，检测率为 100%，对于 DDoS 攻击，检测率达到 99%。

2) 对比分析

根据 5.2 节中的特征提取阶段生成的优化特征集合作为对比实验的输入，将改进的随机森林算法与其他几种分类算法进行了比较(传统的随机森林算法、决策树、支持向量机)，实验结果如表 4 所示。

Table 4. Performance comparison of each classifier
表 4. 各分类器性能比较

算法	Precision	Recall	F1-Score
随机森林	0.9469	0.9571	0.9483
支持向量机	0.8346	0.84	0.8377
决策树	0.8821	0.9040	0.89
改进随机森林	0.9983	0.9905	0.9890

由表可知，改进的随机森林分类精度高达 99%，优于传统的随机森林算法，并且明显优于其他几类分类算法；是因为随机森林是一种集成机器学习算法，在性能方面，优于单个的分类器。在此过程中，还计算了训练时间，如图 7 所示。改进的随机森林算法相比于传统的随机森林算法建模时间较长，因为是对 RF 算法根据每次的迭代计算权重，导致分类效果提高，但建模时间相对增加。

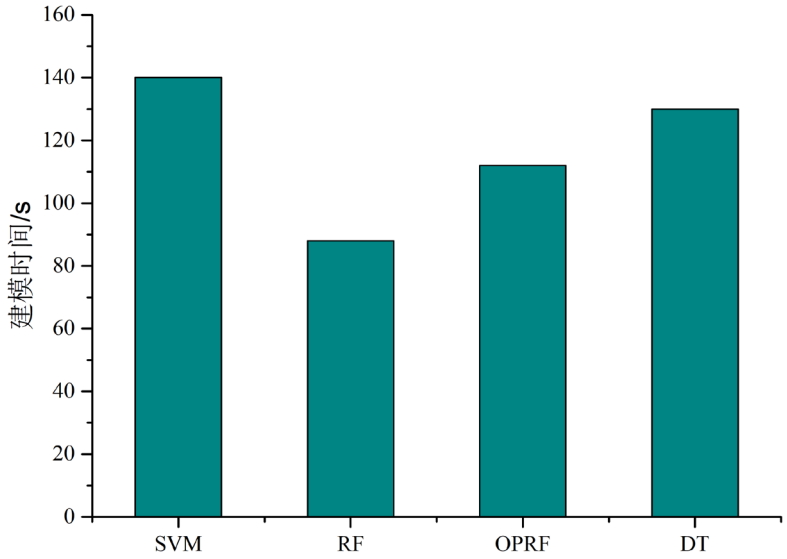


Figure 7. Comparison of modeling time of different classifiers
图 7. 不同分类器建模时间对比

6. 结论

本文首先介绍了随机森林的组合多分类器分类方法，分析了该算法相比于其他分类算法的优越性。根据应用层 HTTP-DDoS 攻击特征原理，改进了传统的随机森林算法，在此基础上设计了基于该算法的检测分类模型(OPRFM)，并对该模型的各个阶段以及各个模块和检测方法检测步骤做了详细的分析，该检测方法对特征属性进行降噪声和消除相关性，从而达到准确检测的目的；最后，通过实验分析，在 HTTP-DDoS 检测方面，与传统随机森林、决策树、支持向量机等算法相比，该算法在各项指标，特别是精确率、召回率和 F1 值方面都有显著的改进。

从研究的局限性来看，加入特征排序和权重计算过程确实意味着建模在时间上比其他算法有更高的成本。在未来的研究中，可以考虑在时间效率和整体预测精度方面的提高，可以采用并行化的方式来提高处理效率。

基金项目

“江苏省博士后科研资助计划”项目(No.1501106C)。

参考文献

[1] Hasan, M., Islam, M.M., Zarif, M.I.I., *et al*. (2019) Attack and Anomaly Detection in IoT Sensors in IoT Sites Using Machine Learning Approaches. *Internet of Things*, **7**, 100059. https://doi.org/10.1016/j.iot.2019.100059

[2] Robinson, R.R.R. and Thomas, C. (2015) Ranking of Machine Learning Algorithms Based on the Performance in Classifying DDoS Attacks. 2015 *IEEE Recent Advances in Intelligent Computational Systems* (*RAICS*), Trivandrum, 10-12 December 2015, 185-190. https://doi.org/10.1109/RAICS.2015.7488411

[3] Singh, K.J. and De, T. (2015) An Approach of DDOS Attack Detection Using Classifiers. In: *Emerging Research in Computing, Information, Communication and Applications*, Springer, New Delhi, 429-437. https://doi.org/10.1007/978-81-322-2550-8_41

[4] Singh, K., Singh, P. and Kumar, K. (2017) Application Layer HTTP-GET Flood DDOS Attacks: Research Landscape and Challenges. *Computers & Security*, **65**, 344-372. https://doi.org/10.1016/j.cose.2016.10.005

[5] 王爱平, 万国伟, 程志全, 等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报, 2011, 22(9): 2059-2074..

[6] Sharafaldin, I., Lashkari, A.H. and Ghorbani, A.A. (2018) A Detailed Analysis of the CICIDS2017 Data Set. *International Conference on Information Systems Security and Privacy*, Springer, Cham, 172-188. https://doi.org/10.1007/978-3-030-25109-3_9

[7] Zhang, Y., Song, B., Zhang, Y., *et al*. (2017) An Advanced Random Forest Algorithm Targeting the Big Data with Redundant Features. *International Conference on Algorithms and Architectures for Parallel Processing*, 642-651.