软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(10):3254−3265 [doi: 10.13328/j.cnki.jos.006062]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合随机森林和梯度提升树的入侵检测研究*

周杰英, 贺鹏飞, 邱荣发, 陈 国, 吴维刚

(中山大学 数据科学与计算机学院,广东 广州 510006)

通讯作者: 吴维刚, Email: wuweig@mail.sysu.edu.cn

摘 要: 网络入侵检测系统作为一种保护网络免受攻击的安全防御技术,在保障计算机系统和网络安全领域起着非常重要的作用.针对网络入侵检测中数据不平衡的多分类问题,机器学习已被广泛用于入侵检测,比传统方法更智能、更准确.对现有的网络入侵检测多分类方法进行了改进研究,提出了一种融合随机森林模型进行特征转换、使用梯度提升决策树模型进行分类的入侵检测模型 RF-GBDT,该模型主要分为特征选择、特征转换和分类器这 3 个部分.采用 UNSW-NB15 数据集对 RF-GBDT 模型进行了实验测试,与其他 3 种同领域的算法相比,RF-GBDT 既缩短了训练时间,又具有较高的检测率和较低的误报率,在测试数据集上受试者工作特征曲线下的面积可达 98.57%. RF-GBDT 对于解决网络入侵检测数据不平衡的多分类问题具有较显著的优势,是一种切实可行的入侵检测方法.

关键词: 网络入侵检测;数据不平衡;随机森林;梯度提升树;UNSW-NB15 数据集

中图法分类号: TP309

中文引用格式: 周杰英,贺鹏飞,邱荣发,陈国,吴维刚.融合随机森林和梯度提升树的入侵检测研究.软件学报,2021,32(10): 3254−3265. http://www.jos.org.cn/1000-9825/6062.htm

英文引用格式: Zhou JY, He PF, Qiu RF, Chen G, Wu WG. Research on intrusion detection based on random forest and gradient boosting tree. Ruan Jian Xue Bao/Journal of Software, 2021,32(10):3254−3265 (in Chinese). http://www.jos.org.cn/1000-9825/6062.htm

Research on Intrusion Detection Based on Random Forest and Gradient Boosting Tree

ZHOU Jie-Ying, HE Peng-Fei, QIU Rong-Fa, CHEN Guo, WU Wei-Gang

(School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

Abstract: As a security defense technique to protect the network from attacks, the system of network intrusion detection system, as a security defense technology to protect the network from attacks, plays a very important crucial role in the field of guaranteeing computer system and network security. However, for the multi-classification problem of unbalanced data in network intrusion detection data, machine learning has been widely used in intrusion detection so as to achieve high intelligence and accuracy. In this paper, the current multi-classification method for network intrusion detection is improved, and an intrusion detection model RF-GBDT is proposed, which applies based on the random forest model for to feature conversion and classification using the model of gradient boosting decision tree to classification model is proposed. The model is mainly includes divided into three parts: Feature selection, feature conversion, and classifier. The UNSW-NB15 dataset was used for the experimental data set to test; experimental tests were carried out on the RF-GBDT model. Compared with the other three algorithms in the same field, RF-GBDT, this model not only reduces training time, but also has a higher detection rate and a lower false alarm rate. The area under the subject's working characteristic curve on the test data set can reach 98.57%. RF-GBDT, the proposed model has significant advantages in solving the multi-class problem of multi-classification of unbalanced data in network intrusion detection data and is a feasible method for network intrusion detection.

* 基金项目: 国家重点研发计划(2018YFB0203803); 国家自然科学基金(U1711263, U1801266); 广东省自然科学基金(2018A030313492, 2018B030312002)

Foundation item: National Key Research and Development Project of China (2018YFB0203803); National Natural Science Foundation of China (U1711263, U1801266); Natural Science Foundation of Guangdong Province of China (2018A030313492, 2018B030312002)

收稿时间: 2019-09-12; 修改时间: 2020-02-01; 采用时间: 2020-04-13

Key words: network intrusion detection; unbalanced data; random forest; gradient boosting tree; UNSW-NB15 dataset

1 引 言

1.1 介 绍

随着计算机应用的快速发展,网络入侵检测已成为保障计算机安全的一道重要屏障.入侵检测早在 30 年前就已成为研究者们关注的领域,现在依然是研究的热点,不断产生新的技术进展.入侵检测是一种安全机制,通过分析主机审计数据、网络流量数据等特征来监测和过滤网络行为,在网络通信中识别出异常访问,并及时通知网络管理员,以此来达到保护网络信息安全的目的.

入侵检测系统通常可以分为 3 个模块:数据采集模块、入侵检测模块和响应处理模块,如图 1 所示.数据采集模块主要负责从网路中采集数据,采集数据的来源有很多,如系统日志、网络数据流量、主机审计数据等,采集到的数据被送入入侵检测模块;入侵检测模块负责对采集到的数据进行数据处理和建模分析,以判断行为是否是攻击行为、属于哪种攻击等,它是入侵检测系统的核心,直接影响到入侵检测系统的好坏;响应处理模块接收入侵检测模块检测到的攻击数据,并根据其攻击类型采取相应的处理措施来进行处理.

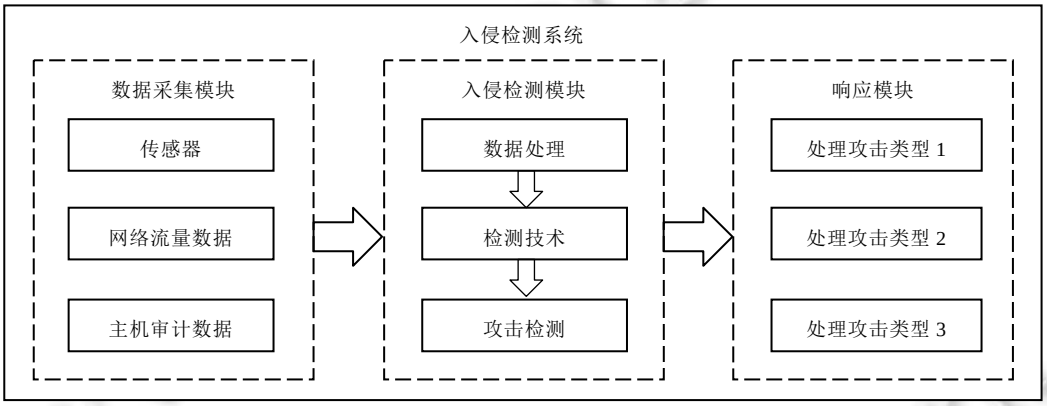

Fig.1 Intrusion detection system

图 1 入侵检测系统

按照入侵检测技术的不同,入侵检测可以分为基于异常的检测和基于滥用的检测.每种检测方式都有各自的局限性,并不适用于识别所有的攻击行为[1].基于滥用的检测是指由专家介入,分析各种常用的攻击手段,建立入侵特征模式库,通过对定义好的入侵模式进行规则匹配,来判断新的数据是否是攻击数据.滥用检测对于已知类型的攻击行为具有很高的准确率,且误报率很低.常用的滥用检测技术包括专家系统、模式匹配等.但是滥用检测需要频繁更新数据库的规则和签名,且很难检测出新型的零日漏洞攻击.基于异常的检测则关注网络和系统的异常情况,对正常状态下的网络和系统进行建模,在定义了正常模式之后,偏离正常模式的数据即被判断为异常.因此,异常检测可以检测出新型的攻击类型.常用的异常检测技术包括基于统计模型的异常检测、基于机器学习的异常检测和基于免疫算法的异常检测等.鉴于每种方法都有局限性,基于滥用检测与异常检测混合的技术[2,3]也被广泛加以研究,并应用于入侵检测系统中.

网络入侵检测主要面临的问题有以下几个.

(1) 多样性:入侵检测数据中数据的攻击类型往往有很多种,检测时需要分辨其具体属于哪种攻击类型.因此,入侵检测问题是一个多分类问题,检测难度要大于单纯的二分类问题;

(2) 数据不平衡:在入侵检测数据中,攻击类型的样本非常少,存在严重的数据不平衡问题.这样会严重影响数据的建模与训练.

针对上述多分类与类别不平衡的问题,本文提出一种基于随机森林(RandomForest,简称 RF)进行特征转换、

使用梯度提升决策树(gradient boosting decison tree,简称 GBDT)模型进行分类的入侵检测模型框架(RF-GBDT).RF-GBDT 框架属于网络入侵检测多分类方法,具有预测精度较高、收敛速度较快以及泛化性能好的特点,可以较好地解决网络异常检测中数据不平衡的多分类问题.

1.2 相关工作

入侵检测系统已逐渐发展成为商业化产品,国内对入侵检测的研究起步较晚,目前,国外的商业化入侵检测产品保持着领导地位.2018 年,致力于信息技术研究和分析的 Gartner 公司公布了入侵检测与防御魔力象限[4],思科(Cisco)、趋势安全(trend micro)和 Intel 安全(McAfee)保持着在入侵检测与防御方面的领导地位.中国绿盟科技公司的入侵检测产品跃升“挑战者”象限,成为亚太地区首个进入该象限的厂商,入围象限的中国厂商还有启明星辰、山石网科.由此可见:在入侵检测产品的实力方面,中国与国外领先的研究机构还有一定的差距.

近年来,机器学习算法被广泛地应用于解决网络入侵检测的问题.谢潇雨[5]结合当下应用广泛的深度神经网络,提出了一种基于批归一化的卷积神经网络模型(BN-CNN),该模型在每一层卷积神经网络中加入对数据的批归一化处理,然后经过网络的全连接层得到最终的分类结果;还提出了一种基于焦点损失函数的入侵检测模型(FL-CNN),该模型使用卷积神经网络的方法对入侵检测数据进行训练,使用焦点函数作为模型的损失函数,通过降低数据的非均衡性对模型检测结果的影响,降低正确分类样本的损失以,提高分类结果的准确性.池亚平等人[6]设计了一种基于增益率算法和卷积神经网络算法的网络入侵检测模型,采用增益率筛选数据集数据特征,在保证入侵检测准确率的同时,缩短了卷积神经网络的训练时间.该模型相比其他基于机器学习的入侵检测模型具有较高的准确率和较强的泛化能力,同时优化了卷积神经网络的训练方式,在保证准确率的同时缩短了神经网络训练的时间.但是机器学习算法会出现过拟合情况,导致入侵检测准确率降低.为解决该问题,夏景明等人[7]提出了一种改进的随机森林分类网络入侵检测方法,通过高斯混合模型聚类算法,将数据分成不同的簇,为每一个簇训练不同的随机森林分类器,通过这些训练好的随机森林分类器进行网络入侵检测.针对网络入侵检测数据不平衡问题,研究人员主要在两个方面加以展开:基于数据层面的方法[8]和基于算法层面的方法[9,10].

基于数据层面的方法主要采用数据抽样的技术使得每个类别所占比例接近来进行模型构建.近年来,基于抽样算法的分类器模型的构建成为一个研究热点[11,12].Chawla[13]提出了基于 Synthetic Minority Over-sampling Technique(SMOTE)方法的提升(boosting)方法,在每一次迭代中,使用 SMOTE 进行上采样.Nekooeimehr[14]提出了一种自适应半监督加权抽样方法,用于解决不平衡数据的分类问题.该算法采用半监督合成聚类方法对少数类样本进行聚类处理,然后根据分类的复杂性或者交叉验证方式来对每个簇进行自适应的过抽样.该算法能够很好地识别少数类样本.然而,基于采样的方法改变了数据的分布,许多机器学习算法是建立在训练集和测试集数据同分布的假设下的,改变数据分布会使得最终在训练集的效果和测试集的效果有偏差.

基于算法层面的方法主要是通过改进算法训练过程以及采用多种集成策略来提升分类性能,使用单一的机器学习算法应用于入侵检测问题往往效果不尽如人意,人们开始研究融合多个机器学习算法应用到入侵检测问题中,比如 K 均值算法和决策树算法的融合[15]、K 均值算法和贝叶斯算法的融合[16,17],采用特征选择技术和分类器集成技术来构建模型.集成分类器是指通过多数投票、提升和装袋的方法结合多个弱分类器来提高多分类性能的模型,集成或混合检测模型的构建,能够有效避免单个分类器出现的资源消耗和分类偏差等问题,能够提高检测模型的分类性能并且能够降低方差,防止过拟合.

本文所提出的 RF-GBDT 模型框架就是从算法层面来提高入侵检测系统中多分类的效果:首先使用 GBDT 模型对特征进行重要性排序,并利用递归消除特征(recursive feature elimination,简称 RFE)[18]方法进行特征选择;然后使用随机森林模型对选出的最优特征子集进行特征变换;最后再使用 GBDT 分别对特征转换后新的训练集和测试集进行训练和预测.

2 梯度提升决策树算法原理

2.1 梯度提升决策树模型

梯度提升决策树模型[19]是一种加法模型:

$$F(x)=\sum_{t=1}^{T}\alpha_t h_t(x) \tag{1}$$

其中,x 为输入样本,$h_t(x)$为分类回归树(classification and regression trees,简称 CART),T 是梯度提升决策树中需要构建的树的数量,α_t 是第 t 棵树的权重.

梯度提升决策树算法采用前向分布算法,首先确定 $F_0(x)$为模型初始值,通常为常数,第 m 步的模型是

$$F_m(x)=F_{m-1}(x)+\alpha_m h_m(x) \tag{2}$$

其中,$F_{m-1}(x)$为当前模型.

新添加的分类回归树 $h_m(x)$通过最小化损失函数求得:

$$h_m=\arg\min_{h}\sum_{i=1}^{N}L(y_i,F_{m-1}(x_i)+h(x_i)) \tag{3}$$

其中,N 是样本个数.梯度提升决策树采用梯度下降法来求解最优模型,将损失函数在当前模型 $F_{m-1}(x)$的负梯度值作为梯度下降的方向:

$$F_m(x)=F_{m-1}-\alpha_m\sum_{i=1}^{N}\nabla_F L(y,F_{m-1}(x_i)) \tag{4}$$

其中,α_m 通过线性搜索(line search)求得:

$$\alpha_m=\arg\min_{\alpha}\sum_{i=1}^{N}L(y_i,F_{m-1}(x_i))-\alpha\frac{\partial L(y_i,F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{5}$$

梯度提升决策树的正则化,可以通过设置学习率(learning rate)来控制:

$$F_m(x)=F_{m-1}(x)+v\alpha_m h_m(x) \tag{6}$$

其中,v 表示学习率.学习率越小,则需要更多的 CART,最终误差会更小,但也会增加训练的时间.所以,需要同时控制学习率和 CART 的个数以确定一个速度快且精度高的模型.

2.2 特征重要性评估

梯度提升决策树模型可以基于特征重要性评估来进行特征选择,以此选择与目标变量更加相关的特征进行训练.它不仅能缩短计算的时间、加快训练的速度,还可以提高模型预测的精度.一个特征在所有决策树中被选择用来划分数据的次数越多,则该特征对于预测结果起的作用越大.

假设有 C 个特征 $X_1,X_2,X_3,\ldots,X_C$,使用基尼指数来评估特征的重要性.在第 i 棵树中,节点 m 的基尼指数是

$$GI_m(p)=\sum_{k=1}^{K}p_{mk}(1-p_{mk})=1-\sum_{k}^{K}p_{mk}^2 \tag{7}$$

其中,K 表示有 K 个类别,p_{mk} 表示节点 m 中类别 k 所占的比例.

节点 m 分枝前后的基尼指数变化量为

$$\Delta G=GI_m-GI_l-GI_r \tag{8}$$

其中,GI_l 表示分枝后左节点的基尼指数,GI_r 表示分枝后右节点的基尼指数.

特征 X_j 在节点 m 的重要性为

$$VIM_{j_m}^{(G_i n_i)}=w_m\times\Delta G \tag{9}$$

其中,$w_m=\dfrac{n}{N}$ 表示节点 m 的样本量 n 占总样本量 N 的比例.

如果特征 X_j 在第 i 棵树中出现的节点在集合 M 中,那么特征 X_j 在第 i 棵树的重要性为

$$VIM_{ij}^{(G_i n_i)}=\sum_{m\in M}VIM_{jm}^{(G_i n_i)} \tag{10}$$

假设梯度提升树一共有 n 棵,那么特征 X_j 的特征重要性为

$$VIM_{j}^{(G_i n_i)}=\sum_{i=1}^{n}VIM_{ij}^{(G_i n_i)} \tag{11}$$

最后归一化处理:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{C} VIM_i} \tag{12}$$

特征重要性是特征在所有树的重要性评分的累加和,评分越高,则该特征对预测结果起的作用越大.

3 RF-GBDT 模型框架

RF-GBDT 由 3 部分构成,分别是特征选择、特征变换和分类器.训练数据首先通过 GBDT 进行训练,得到特征重要性按从大到小排序的特征,利用递归消除特征方法进行特征选择,选择出最优特征子集;选择好最优特征子集之后,使用随机森林模型训练特征子集,将样本落到每棵树的叶子索引作为最终分类器的输入,若有 m 个样本,随机森林模型有 n 棵树,则转换之后的数据大小是 $m \times n$;最后使用 GBDT 模型对转换后的特征进行训练与预测.模型框架的总体结构如图 2 所示.

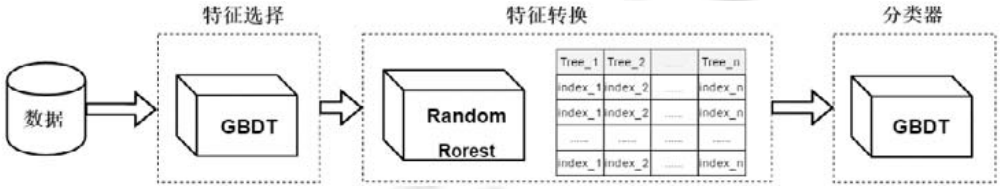

Fig.2 Model and framework of RF-GBDT

图 2 RF-GBDT 模型框架

第 1 部分是特征选择.本文基于递归消除特征[18]方法提出了基于梯度提升决策树的递归消除特征的方法,简称 GBDT-RFE.GBDT-RFE 属于包装法特征选择算法的一种,使用 GBDT 模型进行多轮训练,每轮训练记录损失值,并且消除特征重要性最小的特征,再基于新的特征集进行下一轮训练,直到特征集全部消除完毕为止.如 Algorithm 1 所示:首先使用 GBDT 进行训练,利用 GBDT 的特征重要性评估方法计算出所有特征的特征重要性,从大到小排序,同时记录损失值;然后删除特征重要性最小的特征,即特征重要性排序后的最后一个特征;接着更新特征集,基于新的特征集进行下一轮训练,直到特征集全部消除完毕为止;最后,选取最小损失值所对应的特征集为最优特征集.

Algorithm 1. GBDT-RFE.

1: Input: *X*, *y*;
2: Output: *F*.
3: **while** *features_list*:
4: Train GBDT (*X*,*y*) on *features_list*:
5: *RecordLoss*:
6: Get *ranked_feature_importances*;
7: del *ranked_feature_importances* [−1];
8: *features_list=ranked_feature_importances*;
9: **end while**
10: $F = \arg\min_{F} Loss$

第 2 部分是特征变换.He 在文献[20]中提出了决策树特征转换方法,使用一棵决策树转换的特征数量比使用多棵决策树转换的特征数量要少,包含的信息量更少.本文在此基础上提出了基于随机森林模型进行特征转换的方法.使用随机森林模型对训练数据进行训练,然后把训练数据的每个样本点落到所有树的叶子索引作为分类器的训练集,把测试数据的每个样本点落到所有树的叶子索引作为分类器的测试集.样本 x 遍历所有树,首先经过第 1 棵树,落到索引为 *index*_1 的叶子上;然后样本 x 经过第 2 棵树,落到索引为 *index*_2 的叶子上;直到样

本经过最后一棵树,落到索引为 *index_n* 的叶子上,将所有叶子索引组合起来构成新的训练数据.随机森林模型的树的个数越多,转换后的特征越多,分类效果就越好.但是,随着特征维度的增加,模型训练时间也会增加.文献[20]在特征转换之后还进行了独热编码(one-hot encoding),但本文直接使用叶子索引作为特征,不进行独热编码.若有 m 个样本,随机森林模型有 n 棵树,则转换之后的数据大小是 $m \times n$(如图 3 所示).

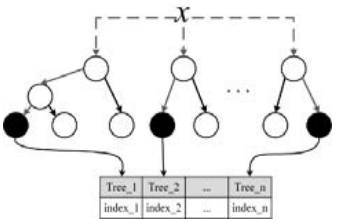

Fig.3 Index example of the decision tree

图 3 决策树索引示例

第 3 部分是分类器.使用 GBDT 对第 2 部分得到的数据大小为 $m \times n$ 的训练集进行训练,对第 2 部分得到的测试集进行预测.在训练集上使用交叉验证技术,调整树的个数和学习率,选择一个最优的模型.需要指出的是:树模型对稀疏数据分类效果没有那么好,所以在使用随机森林模型进行特征转换之后没有进行独热编码,直接使用叶子索引作为数据的输入.

本文所提出的 RF-GBDT 模型框架主要分为 3 部分.

- 第 1 部分,首先使用 GBDT 获得按特征重要性排序后的特征;然后使用递归消除特征方法进行特征选择,删除无关的特征,加快训练速度,提升模型效果;
- 第 2 部分,使用随机森林模型进行特征变换,将样本落在叶子的索引作为新的数据输入.若随机森林模型中树的个数太少,树的深度太浅,则转换的特征包含的信息较少,最终分类的效果就没有那么好;若树的个数太多,树的深度很深,转换的特征虽然包含的信息很多,但是数据量变大了,那么最终的训练时间也会变长;
- 第 3 部分,使用 GBDT 进行分类,需要调节的参数有树的个数和学习率,学习时需要的树越多,虽然检测率越高,但是训练时间越长,学习率越低.需要同时调节树的个数和学习率,两者相互平衡找到一个效果较好的模型.

4 有效性评估指标

网络安全领域的网络流量数据具有样本分布不平衡的特点,为了正确反映模型的真实效果,应该选取合适的评估指标.混淆矩阵是用于计算评估模型有效性的指标,根据表 1 所示的混淆矩阵,介绍 4 种可能发生的情况.

Table 1 Confusion matrix

表 1 混淆矩阵

真实情况	预测结果	
	正例(positive)	反例(negative)
正例(positive)	真正例(true positive)	假反例(false negative)
反例(negative)	假正例(false positive)	真反例(true negative)

表 1 中:真正例(true positive,简称 TP)是将异常样本预测为异常的数量;真反例(true negative,简称 TN)是将正常样本预测为正常的数量;假反例(false negative,简称 FN)是将异常样本预测为正常的数量,属于漏检的样本数量;假正例(false positive,简称 FP)是将正常样本预测为异常的数量,属于误检的样本数量.

网络安全领域的网络流量数据的样本分布不平衡,正常样本多,异常样本很少,如果采用准确率作为评估指标,则无法正确地表现出模型在真实环境的效果.比如:正常样本有 99 个,异常样本有 1 个,将所有样本预测为正常,那么准确率是 99%,检测率却只有 0.入侵检测系统的目标是一方面检测出所有异常样本,另一方面尽量减少将正常样本错误地判定为异常的数量.入侵检测系统采用以下的评估指标来评估异常检测系统的有效性.

- 检测率(detection rate,简称 DR):正确检测出的异常样本占所有异常样本的比例,也称为召回率(recall);
- 误报率(false alarm rate,简称 FAR):正常样本被误判为异常样本的比例;
- 精确率(precision):正确检测的异常样本占检测为异常样本的比例;
- $F1$ 分数值:精确率和召回率的调和平均数.

模型评估指标的具体计算公式如下(其中,C 是类别的数量):

$$DR=\frac{\sum_{i=1}^{c}TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)} \tag{13}$$

$$FAR=\frac{\sum_{i=1}^{c}FP_i}{\sum_{i=1}^{C}(FP_i+TN_i)} \tag{14}$$

$$Precision=\frac{\sum_{i=1}^{c}TP_i}{\sum_{i=1}^{C}(TP_i+FP_i)} \tag{15}$$

$$F1\text{-}score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{16}$$

5 UNSW-NB15 数据集

2015 年,Nour 和 Slay 提出的 UNSW-NB15 数据集[21]作为入侵检测领域新的基准数据集,能够正确反映当今多样的攻击类型和复杂的网络情况.UNSW-NB15 数据集一共有 2 540 044 条数据,包含 49 个特征,一共有 10 个类别,分别是正常样本“Normal”和 9 种攻击类型:“Fuzzers”“Analysis”“Backdoors”“DoS”“Exploits”“Generic”“Reconnaissance”“Shellcode”和“Worms”.UNSW-NB15 数据集[21]有一个子集版本,训练集有 175 341 条数据,测试集有 82 332 条数据,包含 41 个特征.子集数据特征见表 2.

Table 2 Subset data feature of UNSW-NB15

表 2 UNSW-NB15 子集数据特征

序号	特征名称	序号	特征名称	序号	特征名称	序号	特征名称	序号	特征名称
1	dur	10	dttl	19	swin	28	trans_depth	37	ct_ftp_cmd
2	proto	11	sload	20	stcpb	29	res_bdy_len	38	ct_flw_http_mthd
3	service	12	dload	21	dtcpb	30	ct_srv_src	39	ct_src_ltm
4	state	13	sloss	22	dwin	31	ct_state_ttl	40	ct_srv_dst
5	spkts	14	dloss	23	tcprtt	32	ct_dst_ltm	41	is_sm_ips_ports
6	dpkts	15	sintpkt	24	synack	33	ct_src_dport_ltm	–	–
7	sbytes	16	dintpkt	25	ackdat	34	ct_dst_sport_ltm	–	–
8	dbytes	17	sjit	26	smeansz	35	ct_dst_src_ltm	–	–
9	sttl	18	djit	27	dmeansz	36	is_ftp_login	–	–

经过实验发现,UNSW-NB15 数据子集具有冗余样本,冗余的表现是:特征值相同,攻击类型却不同.也就是数据相同,类别却不同.其中的原因可能是 Moustafa 等人[21]对原始数据进行标注的时候产生了一些误差.这类冗余数据对于模型来说属于噪声数据,会影响模型的效果.为此,在数据预处理时,本实验进行了数据清洗,将这些噪声数据全部删除.为了实验效率,将数据子集进行采样,划分为训练集和测试集.训练集和测试集的分布见表 3.

Table 3 Data distribution of training set and test set

表 3 训练集和测试集的数据分布

序号	类别	训练集		测试集	
		数量	比例(%)	数量	比例(%)
1	Normal	6 522	43.79	1 673	44.92
2	Exploits	1 868	12.53	432	11.60
3	Fuzzers	1 389	9.33	353	9.48
4	Reconn	1 294	8.69	301	8.08
5	DoS	578	3.88	146	3.92
6	Generic	1 449	9.73	355	9.53
7	Shellcode	1 118	7.51	287	7.71
8	Analysis	275	1.85	74	1.99
9	Backdoor	274	1.84	74	1.99
10	Worms	130	0.87	29	0.78
合计		训练集数量	14 895	测试集数量	3 724

6 实验结果分析

实验硬件配置为 Intel 四核处理器,主频 2.5GB,8GB 内存,64 位 Windows 10 操作系统.实验在 Anaconda 平台上使用 Python 语言,通过调用 Scikit-learn[22]工具包实现.将 RF-GBDT 框架模型与以下 4 种算法进行了对比:Adaboost、随机森林算法、*K* 最近邻(*K*-nearest neighbor,简称 *K*-NN)算法、逻辑回归(logistic regression,简称 LR)算法.

6.1 数据预处理

在预处理阶段,对数据集进行去除冗余操作,然后对类别特征进行独热编码.为了缩小特征值之间的大小差异,避免较大数量级数据对较小数量级数据造成干扰,保证结果的有效性,对数值型特征进行归一化处理,将所有特征值映射到[0,1]之间.归一化的计算如公式(17)所示:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{17}$$

其中,X 为某一列特征的值,$X_{\max}$ 为特征 X 的最大值,$X_{\min}$ 为特征 X 的最小值.最后,将经过独热编码后的类别特征与归一化后的数值特征拼接起来.

6.2 特征选择

使用 GBDT-RFE 算法进行特征选择,一共有 41 个特征,因此一共运行了 41 轮训练,得到 41 组特征子集.每一轮训练都记录一下训练损失,损失函数采用对数损失(log loss)函数,计算如公式(18)所示:

$$L_{\log}(Y,P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k}\log p_{i,k} \tag{18}$$

其中,

- N 表示样本的数量,K 表示类别的数量;
- $y_{i,k}$ 表示第 i 个样本属于 k 类别;
- $p_{i,k}$ 表示第 i 个样本属于 k 类别的概率;
- Y 表示经过独热编码的矩阵,大小为 $N\times K$,即:如果第 i 个样本属于 k 类别,则 $y_{i,k}=1$;如果第 i 个样本不属于 k 类别,则 $y_{i,k}=0$;
- P 表示概率矩阵,大小为 $N\times K$.

使用 GBDT-RFE 算法进行特征选择,训练损失的变化情况如图 4 所示.当特征数量大于 12 个特征时,训练损失变化不大,图中的虚线表示最小训练损失所对应的特征子集数量为 20 个.本实验选取 20 个特征作为最优特征进行特征转换,按特征重要性排序的特征子集是:“*dbytes*”“*dmeansz*”“*synack*”“*dintpkt*”“*ct_state_ttl*”“*sload*”“*ct_srv_src*”“*service*”“*dur*”“*smeansz*”“*dloss*”“*dttl*”“*sttl*'”“*sbytes*”“*dload*”“*ct_dst_sport_ltm*”“*res_bdy_len*”“*ct_srv_dst*”“*ct_dst_src_ltm*”和“*tcprtt*”.

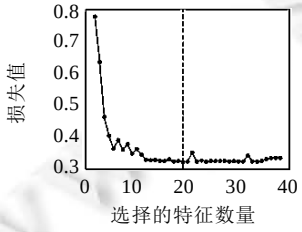

Fig.4 Training loss change

图 4 训练损失变化图

6.3 特征转换

使用随机森林算法进行特征转换.随着树的个数的增加,转换之后的数据也随之变多,GBDT 模型的测试集

损失逐渐收敛,但是模型的训练时间呈线性增长,如图 5 所示.

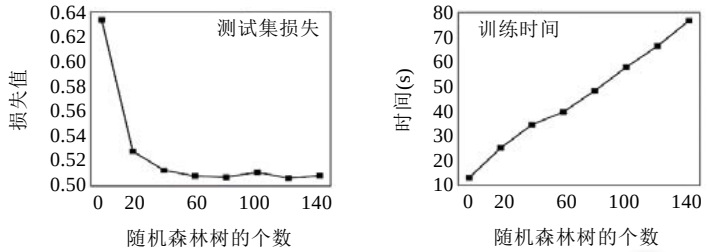

Fig.5 Loss of test set and the training time

图 5 测试集损失和训练时间

6.4 结果对比

表 4 显示了 RF-GBDT 与 *K*-NN、AdaBoost、LR 在训练集十折交叉验证的性能表现,以及与其他文献的对比.其中,N/A 表示结果不存在或者评价方法不同,无法进行对比.RF-GBDT 的检测率是 83.78%,误报率是 1.8%,*F*1 分数值是 83.78%,三者都是最高的.RF-GBDT 作为多分类模型,在每个类别上的检测率也有不错的表现,表 5 显示了 RF-GBDT 与 *K*-NN、AdaBoost、LR 这 4 种算法在每种类别上的检测表现,其中,“Worns”“Reconnaissance”“Shellcode”和“Generic”样本数量都很少,但是检测率都在 84%以上.

Table 4 Comparison of 10-fold cross-validation results of training set and compared with other algorithms

表 4 训练集十折交叉验证结果比较以及与其他算法比较

算法	检测率(%)	误报率(%)	*F*1 分数(%)
K-NN	64.35±1.05	3.96±0.12	64.35±1.05
AdaBoost	73.85±0.82	2.91±0.09	73.85±0.82
LR	63.16±0.84	4.09±0.84	63.16±0.84
深度神经网络算法[23]	80	N/A	76
RICSA-KELM[24]	N/A	2.12	N/A
RF-GBDT	**83.78±0.91**	**1.80±0.91**	**83.78±0.91**

Table 5 Comparison of detection rates of four algorithms in each category

表 5 4 种算法在每个类别的检测率对比

类别	比例(%)	检测率(%)			
		K-NN	AdaBoost	LR	RF-GBDT
Analysis	1.85	52.71	52.43	25.74	**70.52**
Backdoor	1.84	14.23	53.65	3.66	**72.65**
DoS	3.88	7.78	21.98	1.39	**35.47**
Exploits	12.53	60.50	62.71	60.29	**76.31**
Fuzzers	9.33	39.02	54.00	33.12	**63.50**
Generic	9.73	56.31	67.90	40.51	**84.27**
Normal	43.79	85.57	86.34	87.58	**92.38**
Reconn.	8.69	56.19	81.68	71.71	**92.89**
Shellcode	7.51	48.49	83.72	45.36	**91.23**
Worms	0.87	14.62	41.54	0.00	**82.31**

另外,在测试集上对比了受试者工作特征曲线(receiver operating characteristic curve,简称 ROC 曲线)和精确率-召回率曲线(precison-recall curve,简称 PR 曲线),如图 6 所示.

图 6 显示了 ROC 曲线下的面积(area under curve,简称 ROC AUC)和 PR 曲线下的面积(area under curve,简称 PR AUC).

- RF-GBDT 的 ROC AUC 达到 98.57%,位列第一;AdaBoost 的 ROC AUC 是 97.31%,排在第 2 位;
- PR AUC 也是 RF-GBDT 的最高,可达 91.48%;第 2 名是 AdaBoost,是 83.48%.

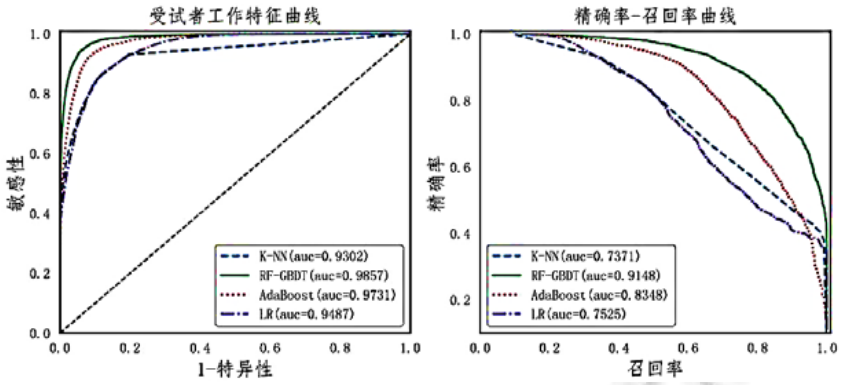

Fig.6 Working characteristic curve and precision-recall curve of subjects

图 6 受试者工作特征曲线和精确率-召回率曲线

实验采用了 UNSW-NB15 数据集训练模型,并且对比了 RF-GBDT、Logistic Regression、AdaBoost 和 *K*-NN 这 4 种算法在训练集上十折交叉验证的表现.

最终实验结果表明,RF-GBDT 模型具有更高的检测率、更低的误报率.

实验结果显示:RF-GBDT 的检测率是 83.78%、误报率是 1.8%、*F*1 分数值是 83.78%、ROC AUC 是 98.57%、PR AUC 是 91.48%.

另外,对于样本量很少的类别,RF-GBDT 的检测率也很高,比如类别“Worns”“Reconnaissance”“Shellcode”和“Generic”样本数量都很少,但是检测率都在 84%以上.

7 结束语

本文针对网络入侵检测数据不平衡的多分类问题,提出了融合随机森林特征变换和梯度提升树的 RF-GBDT 入侵检测分类模型框架,该模型框架主要有 3 个部分:特征选择、特征转换和分类器.

使用 GBDT 的特征重要性参数进行特征选择,丢弃无关特征,不仅能够减少计算量、加快训练的速度,还能提高模型的检测率;使用 RandomForest 训练数据,将样本落到每一棵树的叶子索引作为新的特征;使用 GBDT 进行分类,调整合适的树的个数和学习率,选择最优的模型参数.

实验结果表明:在 UNSW-NB15 数据集上,本文提出的模型 RF-GBDT 具有检测率较高、误报率较低的特点.RF-GBDT 能够较准确地检测出网络流量中的攻击类型,尤其能够更好地检测出样本量少的攻击类型.RF-GBDT 对于解决网络入侵检测数据不平衡的多分类问题,具有较显著的优势.

References:

[1] Lin WC, Ke SW, Tsai CF. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-based Systems, 2015,78:13−21. [doi: 10.1016/j.knosys.2015.01.009]

[2] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 2014,41(4):1690−1700. [doi: 10.1016/j.eswa.2013.08.066]

[3] Karami A, Guerrero-Zapata M. A fuzzy anomaly detection system based on hybrid PSO-*K*-means algorithm in content-centric networks. Neurocomputing, 2015,149:1253−1269. [doi: 10.1016/j.neucom.2014.08.070]

[4] Lawson C, Neiva C. Magic quadrant for intrusion detection and prevention systems. 2018. https://www.gartner.com/doc/3844163?ref=mrktg-srch

[5] Xie XY. Research on intrusion detection model based on convolutional neural network [MS. Thesis]. Najing: Nanjing University of Posts and Telecommunications, 2019 (in Chinese with English abstract). [doi: 10.27251/d.cnki.gnjdc.2019. 000590]

[6] Chi YP, Yang YT, Li GF, *et al*. Design and implementation of network intrusion detection model based on GR-CNN algorithm. Computer Applications and Software, 2019(12):297−302 (in Chinese with English abstract).

[7] Xia JM, Li C, Tan L, *et al*. Improved random forest classifier network intrusion detection method. Computer Engineering and Design, 2019(8):2146−2150 (in Chinese with English abstract). [doi: 10.16208/j.issn1000-7024.2019.08.009]

[8] Garcia S, Derrac J, Triguero I, *et al*. Evolutionary-based selection of generalized instances for imbalanced classification. Knowledge-based Systems, 2012,25(1):3−12. [doi: 10.1016/j.knosys.2011.01.012]

[9] Sun Z, Song Q, Zhu X, *et al*. A novel ensemble method for classifying imbalanced data. Pattern Recognition, 2015,48(5): 1623−1637. [doi: 10.1016/j.patcog.2014.11.014]

[10] Zhang Z, Krawczyk B, Garcìa S, *et al*. Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. Elsevier Science Publishers B. V. 2016. [doi: 10.1016/j.knosys.2016.05.048]

[11] Sain H, Purnami SW. Combine sampling support vector machine for imbalanced data classification. Procedia Computer Science, 2015,72(Complete):59−66. [doi: 10.1016/j.procs.2015.12.105]

[12] Jian C, Gao J, Ao Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Elsevier Science Publishers B. V., 2016. [doi: 10.1016/j.neucom.2016.02.006]

[13] Chawla NV, Lazarevic A, Hall LO, *et al*. SMOTEBoost: Improving prediction of the minority class in boosting. Lecture Notes in Computer Science, 2003,2838:107−119. [doi: 10.1007/978-3-540-39804-2_12]

[14] Gaddam SR, Phoha VV, Balagani KS. *K*-Means+ID3: A novel method for supervised anomaly detection by cascading *K*-means clustering and ID3 decision tree learning methods. IEEE Trans. on Knowledge and Data Engineering, 2007,19(3):345−354.

[15] Muda Z, Yassin W, Sulaiman MN, *et al*. A *K*-means and Naive Bayes learning approach for better intrusion detection. Information Technology Journal, 2011,10(3):648−655.

[16] Khammassi C, Krichen S. A GA-LR wrapper approach for feature selection in network intrusion detection. Computers & Security, 2017,70:255−277. [doi: 10.1016/j.cose.2017.06.005]

[17] Aburomman AA, Reaz MBI. A novel SVM-KNN-PSO ensemble method for intrusion detection system. Applied Soft Computing, 2016,38:360−372.

[18] Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. Machine Learning, 2002,46(1-3):389−422. [doi: 10.1023/a:1012487302797]

[19] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 2001,29(5):1189−1232. [doi: 10.2307/2699986]

[20] He X, Bowers S, Candela JQ, *et al*. Practical lessons from predicting clicks on ADS at Facebook. In: Proc. of the 20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (ADKDD 2014); the 8th Int'l Workshop on Data Mining for Online Advertising. New York: ACM, 2014. 1−9. [doi: 10.1145/2648584.2648589]

[21] Moustafa N, Slay J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: Proc. of the Military Communications and Information Systems Conf. (MilCIS 2015). IEEE, 2015. 1−6. [doi: 10.1109/MilCIS.2015.7348942]

[22] Swami A, Jain R. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 2012,12(10):2825−2830.

[23] Cai HM, Wang QX. Research on intrusion detection technology based on deep learning. Network Security Technology & Application, 2017(11):62−64 (in Chinese with English abstract).

[24] Ma C. Parallel network intrusion detection method based on ReliefF and improved crow search optimization. Application Research of Computer, 2019(10):3063−3068 (in Chinese with English abstract). http://kns.cnki.net/kcms/detail/51.1196.TP.20180811.1341.098.html

附中文参考文献:

[5] 谢潇雨.基于卷积神经网络的入侵检测模型研究[硕士学位论文].南京:南京邮电大学.2019. [doi: 10.27251/d.cnki.gnjdc.2019.000590]

[6] 池亚平,杨垠坦,李格菲,等.基于 GR-CNN 算法的网络入侵检测模型设计与实现.计算机应用与软件,2019(12):297−302.

[7] 夏景明,李冲,谈玲,等.改进的随机森林分类器网络入侵检测方法.计算机工程与设计,2019(8):2146−2150. [doi: 10.16208/j.issn1000-7024.2019.08.009]

[23] 蔡洪民,王庆香.基于深度学习的入侵检测技术研究.网络安全技术与应用,2017(11):62−64.

[24] 马超.基于 ReliefF 和改进乌鸦搜索优化的并行入侵检测方法.计算机应用研究,2019(10):3063−3068. http://kns.cnki.net/kcms/detail/51.1196.TP.20180811.1341.098.html

周杰英(1966－),女,博士,副教授,CCF 专业会员,主要研究领域为网络空间安全,计算机网络,车联网路由协议,边缘计算,区块链.

贺鹏飞(1996－),男,硕士,主要研究领域为网络空间安全,计算机网络,物联网,车联网.

邱荣发(1993－),男,硕士,主要研究领域为网络安全态势感知,机器学习,深度学习.

陈国(1997－),男,硕士,CCF 学生会员,主要研究领域为网络空间安全,区块链.

吴维刚(1976－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为网络与分布式计算,云计算,分布式机器学习,区块链.