

基于改进极端随机树的异常网络流量分类

韦海宇^a, 王 勇^{a,b,c,d}, 柯文龙^b, 俸 皓^{c,d}

(桂林电子科技大学 a. 计算机与信息安全学院; b. 信息与通信学院;
c. 广西高校云计算与复杂系统重点实验室; d. 广西可信软件重点实验室, 广西 桂林 541004)

摘 要: 为更有效地识别网络流量中少量的异常流量样本, 提出一种基于改进极端随机树的异常流量分类方法。计算数据中每个特征的信息增益率, 获得较低维度的特征集。在此基础上, 使用随机训练方法训练分类模型, 对一部分基分类器使用全部样本进行训练, 对另一部分则使用经过重采样的数据进行训练, 并使用加权统计的方法修改其最后的投票规则。实验结果表明, 该方法在 NSL-KDD 数据集上可达到 0.995 6 的精确率, 与 ET 和 RF 集成分类算法相比, 其在数据样本较少的类别上分类效果更好。

关键词: 异常网络流量; 流量分类; 特征选择; 随机训练; 极端随机树

中文引用格式: 韦海宇, 王 勇, 柯文龙, 等. 基于改进极端随机树的异常网络流量分类[J]. 计算机工程, 2018, 44(11): 33-39.

英文引用格式: WEI Haiyu, WANG Yong, KE Wenlong, et al. Abnormal network traffic classification based on improved extremely random tree[J]. Computer Engineering, 2018, 44(11): 33-39.

Abnormal Network Traffic Classification Based on Improved Extremely Random Tree

WEI Haiyu^a, WANG Yong^{a,b,c,d}, KE Wenlong^b, FENG Hao^{c,d}

(a. School of Computer Science and Information Security; b. School of Information and Communication;

c. Guangxi Colleges and Universities Key Laboratory of Cloud Computing and Complex Systems;

d. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)

[Abstract] In order to identify a small number of abnormal traffic samples in network traffic more effectively, an abnormal traffic classification method based on improved extremely random trees is proposed in this paper. The information gain rate of each feature in the data is calculated and the feature set of lower dimensions is obtained. On this basis, the classification model is trained with the use of random training method. For parts of the base classifiers, all training samples are used, and for the others, they are trained with the resampling data. At the same time, the weighted statistical method is used to modify the final voting rules for those base classifiers using the resampling data. Experimental results show that the proposed method can achieve the accurate rate of 0.995 6 on the NSL-KDD data sets. Meanwhile, compared with other ensemble classification algorithms such as ET and RF, this method obtains better classification results when dealing with fewer data samples.

[Key words] abnormal network traffic; traffic classification; feature selection; random training; extremely random tree
DOI: 10.19678/j.issn.1000-3428.0048489

0 概述

随着计算机网络服务技术的迅速发展, 计算机网络安全已经成为一个备受关注的问题, 如何快速、准确地发现网络中的异常流量已经成为目前网络安全管理的主要研究内容。异常网络流量通常定义为

偏离正常的网络连接, 主要包括网络扫描、网络攻击、病毒感染等^[1]。检测、监控网络安全状态, 进行网络流量分类十分有必要, 但是在海量的网络流量数据下, 异常网络流量相对来说非常少, 因此, 识别少量异常流量样本是一项具有价值的研究。

近年来, 机器学习算法快速发展, 并被广泛应用

基金项目: 国家自然科学基金(61662018, 61163058); 广西自然科学基金(2016GXNSFAA380153); 广西高校云计算与复杂系统重点实验室研究课题(14103, 15208); 广西云计算与大数据协同创新中心研究课题(YD16303)。

作者简介: 韦海宇(1992—), 男, 硕士研究生, 主研方向为大数据、机器学习; 王 勇, 教授、博士生导师; 柯文龙(通信作者), 博士研究生; 俸 皓, 副教授、博士。

收稿日期: 2017-08-31

修回日期: 2017-12-05

E-mail: 984577327@qq.com

到网络流量分类中。机器学习分类是一种有监督学习,首先需要提取数据样本,根据标签从数据集中学习,通过训练生成分类模型,然后将新的数据导入到分类模型中进行自动分类,分类的效果取决于分类样本的特征属性和分类模型的训练。本文提出一种基于改进极端随机树的异常网络流量分类方法,以期在大量网络流量情况下提高对少量异常网络流量样本的识别率。

1 相关工作

文献[2]提出一种基于 C4.5 决策树和支持向量机(Support Vector Machine, SVM)检测异常流量的组合算法。首先根据 C4.5 算法对异常网络流量进行分类,将其分为正常网络流量和异常网络流量,然后使用 SVM 算法对异常的网络流量进行细分类,将其分为不同攻击类别。该算法在 KDD CUP 99 数据集上能够获得较好的效果,但该数据集存在大量的重复数据,对模型的结果影响较大。

为了减少分类模型的训练时间,文献[3]首先根据主成分分析(Principal Component Analysis, PCA)方法对数据进行降维,然后使用随机森林对异常网络流量分类。虽然 PCA 能够有效地减少网络流量特征,但也会在一定程度上丢失一些重要数据特征。

文献[4-5]通过设计基于 C4.5 的决策树对异常流量数据进行分类。该方法简单、计算速度快、分类结果较好,但对新的网络环境适应性比较差。

文献[6-7]基于层次聚类方法对数据进行特征选择,其中文献[7]基于低维数据使用 SVM 算法对流量数据进行分类,该方法在 NSL-KDD 数据集上可得到 95.2% 的分类准确率,但是在少量样本的 U2L 类别上分类结果相对较差。

文献[8]提出 PSO-Random Forest 的异常流量检测算法,首先根据粒子群优化(Particle Swarm Optimization, PSO)算法获得最优的特征组合,然后使用随机森林算法对这些特征进行分类。该算法在 KDD CUP 99 数据集上能够获得很好的效果。

在异常网络流量分类的问题上,研究者关注分类准确率和时间效率,致力于结合各个算法的优点提高分类的精度,减少分类模型训练时间。

文献[9]首先提出一个基于密度比的数据特征转换算法以获得更好的特征表示,然后基于 SVM 实现网络流量的分类。该方法在 NSL-KDD 数据集上得到了 99.28% 的准确率。

文献[10]提出超图遗传算法和 SVM 算法的组合算法,其中超图遗传算法对模型和数据特征选择进行参数优化。该文在实验中对使用特征选择算法后的数据和原始数据进行分类结果比较,结果表明特征选择可以更好地提高分类检测率。

文献[11]提出基于词袋模型聚类的异常流量识

别方法,首先根据 K-Means 算法对流量做均值聚类,得到网络流量的关键特征,然后利用 SVM、C4.5 等分类算法进行识别,平均识别率达到 90.9%。

文献[12]研究集成学习方法在异常网络流量中的应用,重点介绍 Adaboost 算法、Boost 算法等,该文总结得出集成方法在异常网络流量分类中效果更为明显,能达到 81% 左右的准确率。

文献[13]首先根据分类与回归树(Classification and Regression Tree, CART)算法将网络流量分为异常流量和正常流量 2 种流量类型,然后比较 SVM、Naive Bayes、Neural Network 在攻击类别分类中的效果。利用该方法在 NSL-KDD 数据集上对 DoS、R2L 和 Probe 的分类获得了 96% 的准确率,但该文未研究 U2R 攻击类别的分类。

综上所述,目前的异常网络流量分类方法研究主要集中在决策树、SVM 等机器学习方法,但是仍存在问题。多数研究只考虑在整体上评估分类效果,而对于少量样本的类别并没有过多关注。此外,决策树对少量样本类别的数据容易过拟合,或者是算法模型复杂、训练时间长、使用的实验数据存在样本重复等。针对以上问题,本文提出一种基于极端随机树的改进算法(IET)。本文的创新点在于:1)结合极端随机树(ET)和随机森林算法(RF)的优点,改进分类模型的训练方法和最后的投票决策^[14]过程;2)更全面地训练少量样本的类别,同时保持算法模型基分类器的随机独立性。

2 极端随机树算法

文献[15]提出极端随机树算法,该算法是一种集成算法,用 $\{T(K, X, D)\}$ 表示,其中, T 表示最后的分类器模型, D 为数据样本集, K 为基分类器的数量,每个基分类器根据输入样本 $X = \{x_1, x_2, \dots, x_M\}$ 产生预测结果,最终通过投票确定最后的分类类别。ET 算法的具体步骤如下:

步骤 1 给定原始数据集 D 、样本数量 N 、特征数量 M ,在极端随机树的分类模型中,每个基分类器使用全部的样本进行训练。

步骤 2 根据 CART 决策树算法生成基分类器。为增强随机性,在每个节点分裂时随机从 M 个特征中选择 m 个特征,对每个节点选择最优属性进行节点分裂,分裂过程不剪枝。对分裂产生的数据子集再迭代执行步骤 2,直至生成一棵决策树。

步骤 3 重复步骤 1 和步骤 2 迭代 K 次,生成 K 棵决策树以及极端随机树。

步骤 4 对生成的极端随机树使用测试样本生成预测结果,对所有基分类器的预测结果进行统计,利用投票决策的方法产生最终的分类结果。

ET 算法具有如下优点:1)是一种集成学习方法,其通过投票决策产生预测结果,具有更好的泛化

能力;2)使用全部数据样本训练基分类器,使得所有样本都被训练;3)在节点分割上随机选择分割属性,增强了基分类器间节点分裂的随机性。但是该算法也有些不足,表现在每棵决策树都使用全部样本进行训练模型,这使得决策树间的随机性有所降低,对结果预测造成影响。为既能保证模型得到充分训练,又能提高基分类器间的独立性,本文提出改进的极端随机树算法。

3 改进的极端随机树算法

本文在训练分类模型的过程中,对一部分基分类器采用全部样本构建决策树,另一部分基分类器采用经过重采样的样本来构建决策树,同时到最后投票结果中,对于使用重采样样本进行训练的基分类器,采用加权投票方法,优化分类器的投票过程,形成改进的极端随机树算法。具体算法框架如图1所示,其中分类过程包括3个部分:特征选择,装配训练模型,投票代理决策。

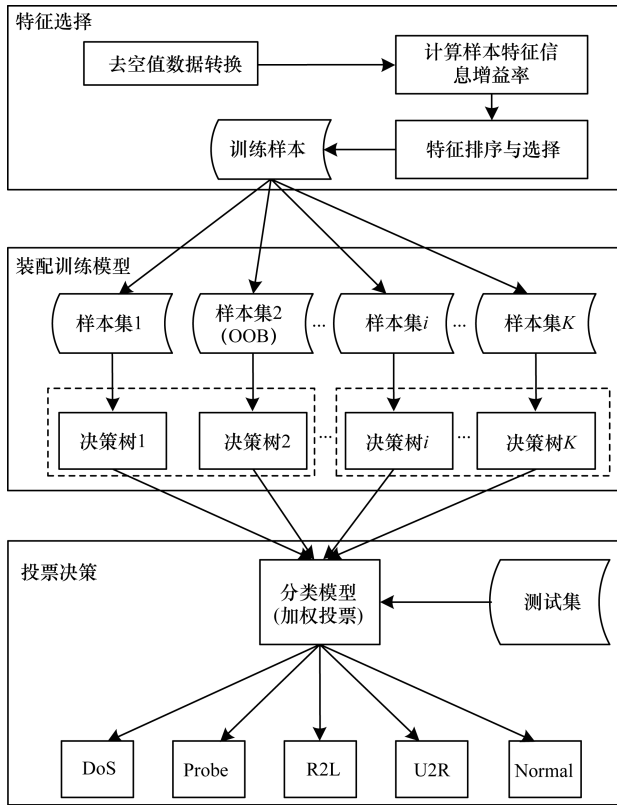


图1 改进的异常网络流量分类算法框架

3.1 特征选择

特征选择的具体步骤如下:

步骤1 特征预处理。首先将数据的空值特征进行替换,在本文中,将空值特征用0替换,消除网络流量类型偏差的影响;其次,为统一数据格式,本文将网络协议、服务类型、网络状态等名义变量映射为数值索引,例如协议内容包括{icmp,tcp,udp},对应的索引地址为{0,1,2}。

步骤2 计算特征值的信息增益率。例如设定A为网络协议的某个特征,数据集D中特征A的信息增益率 $gainRatio(D,A)$ 根据式(1)计算。

$$gainRatio(D,A) = \frac{gain(D,A)}{splitInfo(D,A)} \quad (1)$$

要求得特征A的信息增益率,应已知该属性的信息熵 $gain(D,A)$ 和信息分裂度量 $splitInfo(D,A)$ 。 $gain(D,A)$ 根据式(2)计算。

$$gain(D,A) = entropy(D) - \sum_{a=1}^A p_a entropy(A) \quad (2)$$

其中, p_a 为特征A各类划分的频率, $entropy(D)$ 和 $entropy(A)$ 的计算方法如式(3)所示。

$$entropy(D) = - \sum_{c=1}^C p_c \times \lg(p_c) \quad (3)$$

在式(3)中,C表示在数据集D有C种类别,本文中C取{0,1,2,3,4},分别表示DoS、Probe、R2L、U2R和Normal, p_c 表示 $C=c$ 在总样本中的占比,该表达式表示关于第c个分类的信息度量。相应地,在求 $entropy(A)$ 时,C表示在特征A上可划分的类别, p_c 表示该类别数据子集与总数的占比,最后求得关于特征A的信息熵。

根据式(4)计算特征A中的分裂信息度量:

$$splitInfo(D,A) = - \sum_{a=1}^A p_a \times \lg(p_a) \quad (4)$$

步骤3 特征排序和选择。根据样本特征的信息增益率对其进行排序,选择相关性较大的特征,同时去除相关性较小的特征,在本文中根据阈值选择网络流量特征。

3.2 改进的训练模型

本文算法主要是改进训练方法和模型,具体步骤如下:

步骤1 选择特征向量。计算数据样本特征值的信息增益率,并对其建立特征值的索引信息。

步骤2 生成字典对象,在训练极端随机分类树时使用字典保存基分类树模型。

步骤3 装配训练模型。采用与其他算法不同的组装方法,在构建的训练模型中,一部分基分类器使用全部的训练样本,另一部分基分类器使用重采样的方法产生训练样本。该过程的目的是既能全面训练少量样本,又能增强基分类器间的独立性,以达到更好的预测结果。本文首先生成10以内的随机整数,对2求余,若为0,则采用抽回样本进行训练;若为1,则采用全部样本进行训练。在训练过程中,使用字典 $T \leftarrow \{t_1, t_2, \dots, t_k\}$ 记录每棵决策树的训练,其中 t_i 表示每个基分类器的分类规则,同时记录训练样本的信息。

步骤 4 训练分类器。如果输入样本低于节点最低样本或者经过规则可以确定分类类别,则生成叶子节点,否则随机选择若干个候选属性 $A \leftarrow \{a_1, a_2, \dots, a_m\}$, 产生相应的候选样本子集 $D^m = \{D_1, D_2, \dots, D_m\}$ 。

步骤 5 选择分裂的属性值,该过程与 CART 的分裂规则一致,但是在选择节点分裂的特征值时根据分裂分数最高,即 $\max(score(A, D))$ 的属性进行分裂,产生 2 个数据子集数据。评估分裂分数是特定归一化后的信息增益计算方法,通过式(5)计算。

$$score(A, D) = \frac{2i_A^D(D)}{entropy(D) + entropy(D, A)} \quad (5)$$

其中, $entropy(D)$ 为数据集 D 中的分类信息熵,计算方法见式(3), $entropy(D, A)$ 为属性 A 的分裂信息, $i_A^D(D)$ 为分类信息熵和分裂熵的相乘值。

步骤 6 分别对左子树和右子树迭代执行上述步骤,生成决策规则,直到满足条件为止。

步骤 7 对上述步骤重复 K 次。

3.3 改进的投票方法

在改进的分类器模型中,对于使用重采样数据进行训练的基分类器采用加权投票表决方法。在抽回采样中,使用 bagging 方法从原始训练集中有放回地抽取样本,形成新的样本子集,该样本被称为 Out-of-Bag(OOB)数据。对于原始数据样本,每个未被抽取的概率为:

$$p = \left(1 - \frac{1}{N}\right)^N \quad (6)$$

将当前决策树的 OOB 数据记为 O_t , 由于决策树训练过程是有监督的,通过比较当前分类器的分类结果和真实样本,得到 O_t 中分类正确的样本个数,记为 O_n 。由此,可得到当前决策树 t 的分类准确率:

$$p_t = \frac{O_n}{O_t} \quad (7)$$

将决策树 OOB 数据的分类准确率 p_t 作为对应分类决策树的权重,对样本 x 的分类结果进行加权统计,对于预测的类别投票记为:

$$vote = \max\left(\sum_{i=1}^T T_{i,x} p_i\right) \quad (8)$$

其中, $T_{i,x}$ 表示分类正确或者错误,取值 1 或者 0,最后得票最多的即为最后的分类结果。

4 实验与结果分析

4.1 实验数据

KDD99 数据集于 1999 年产生,由于该数据集重复样本比较多,因此本文使用 NSL-KDD 数据集^[16]。该数据集是在 KDD99 数据集上进一步整理得到的数据集,共有 41 个网络流量特征,包括连接

特征、时间特征、数据包统计等信息。从数据类型上看,该数据集有 3 个名义数据、6 个二分类属性和 32 个数值类型。协议类型有 ICMP、UDP、TCP 3 种;状态类型有 OTH、REJ、RST、OSO 等 16 种;服务类型有 IRC、FTP、whois 等 70 种;攻击类型有 DoS、Probe、R2L 和 U2R 4 种。该数据集的攻击类别信息如表 1 所示。

表 1 NSL-KDD 数据集攻击类别

攻击类别	具体攻击
DoS	back, land, neptune, pod, smurf, teardrp, mailbomb, processtable
Probe	ipsweep, nmap, portsweep, satan, saint
R2L	ftp_write, imap, guess_passwd, multihop, phf, spy, warezclient, warezmaster
U2R	buffer_overflow, loadmodule, perl, spy, rootkit, ps, xterm, sqlattack, mscan

本文对 NSL-KDD 数据进行混洗后按 6:4 划分进行实验,即训练集占总数据的 60%,测试集占 40%。在本文中,所有对比的实验结果都是基于同组数据的情况下进行,数据样本分布如表 2 所示。

表 2 实验数据样本分布

攻击类别	训练样本数	测试样本数
Normal	46 310	30 880
DoS	32 057	21 328
Probe	8 443	5 634
R2L	2 234	1 511
U2R	65	54

实验的硬件配置为 IntelCore(TM) i5-4210H@2.90 GHz,4 GB RAM,软件配置为 Windows 10 64 位操作系统,Scikit-Learn 0.18。

4.2 分类性能评估方法

在特征选择过程中,主要是计算流量特征的信息增益率,根据信息增益率值从大到小进行排序和筛选。

为有效评估分类效果,实验中对于分类结果的评价指标使用精确率和召回率。分类精确率计算公式如下:

$$p = \frac{TP}{TP + FP} \quad (9)$$

其中, TP 表示在正样本中分类正确的样本, FP 表示将负样本分为正确的样本, FN 表示正样本数分为负样本数。召回率计算公式如下:

$$r = \frac{TP}{TP + FN} \quad (10)$$

4.3 特征选择结果

本文实验计算属性的信息增益率,并对其进行排序,图 2 列出了数据集中各个特征信息增益率的值以及排序的结果。从图 2 中可以看出:将信息增益率阈

值调整到 0.01 时,最后获得 39 个数据特征;将阈值调整到 0.1 时,获得 33 个数据特征,其中 `srv_serror_rate` 的信息增益率最高,达到 0.507,而 `num_outbound_cmds` 和 `is_host_login` 的信息增益率为 0。

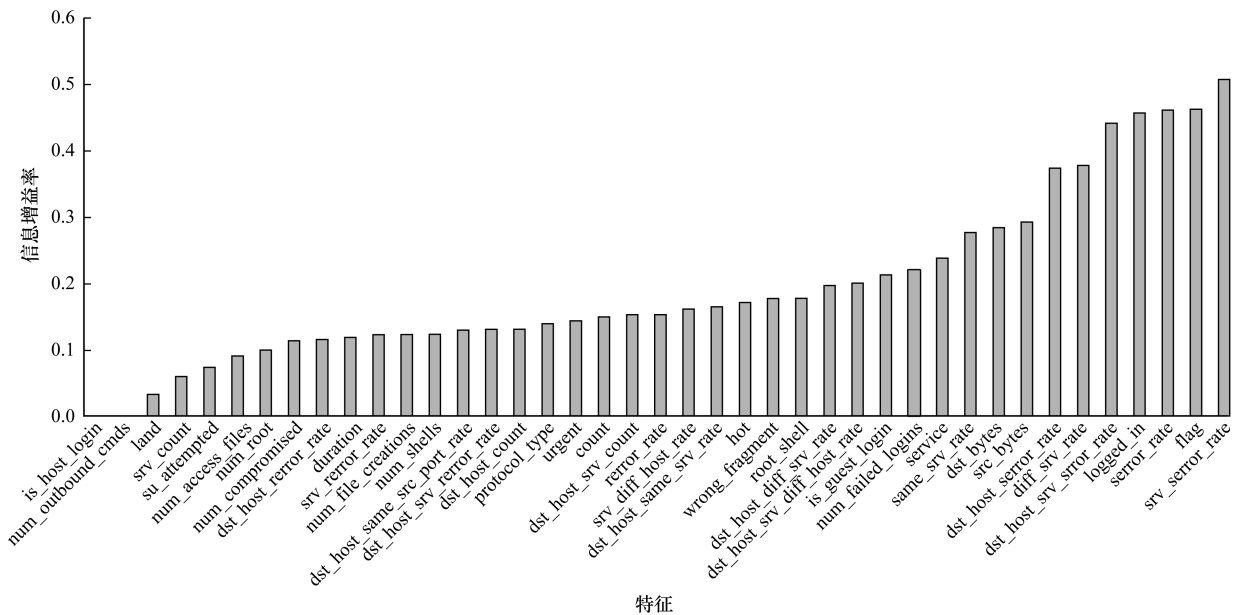


图 2 特征选择排序结果

4.4 分类结果对比

根据不同的数据集验证极端随机树 (ET)、随机森林 (RF) 和本文提出的改进极端随机树方法 (IET),比较这三种集成算法的分类效果,从各个网络流量类别分类的精确率、召回率和时间上进行比较,最后验证基分类树个数与精确率的关系以及基分类器与时间的关系。本文提出的算法基于开源机器学习库 Scikit-Learn^[17],对其中的极端随机树算法进行改进并编码实现。

将本文算法同其他集成算法的分类结果相比较,取 Scikit-Learn 中的极端随机树和随机森林算法作为对比,分别计算未经特征选择的数据、特征的信息增益率大于 0.01 的数据,以及数据特征的信息增益率大于 0.1 的数据,比较同类数据集下 3 种算法的精确率 p 和召回率 r 。

从表 3 ~ 表 5 中可以看出,对于 5 种流量类别的分类结果,3 种算法对 DoS、Probe 这两种攻击类型的数据集上分类结果总体相差不大,精确率都能达到 0.99,主要是该类数据的训练样本比较多,训练过程比较充分,而 R2L 和 R2R 这两类数据样本则有所不同,具体数据比较如表中加粗部分所示。

表 3 3 种算法在未经特征选择的数据集上的结果

攻击类型	IET 算法		ET 算法		RF 算法	
	r	p	r	p	r	p
Normal	0.99	0.99	0.99	0.99	0.99	0.98
DoS	0.99	0.99	0.99	0.99	0.99	0.99
Probe	0.99	0.99	0.99	0.99	0.98	0.99
R2L	0.91	0.96	0.90	0.94	0.91	0.96
U2R	0.71	0.77	0.69	0.74	0.64	0.75

表 4 3 种算法在信息增益率大于 0.01 数据集上的结果

攻击类型	IET 算法		ET 算法		RF 算法	
	r	p	r	p	r	p
Normal	0.99	0.99	0.99	0.99	0.99	0.98
DoS	0.99	0.99	0.99	0.99	0.99	0.99
Probe	0.99	0.99	0.99	0.99	0.98	0.99
R2L	0.92	0.96	0.91	0.95	0.92	0.96
U2R	0.52	0.86	0.56	0.79	0.41	0.90

表 5 3 种算法在信息增益率大于 0.1 数据集上的结果

攻击类型	IET 算法		ET 算法		RF 算法	
	r	p	r	p	r	p
Normal	0.99	0.99	0.99	0.99	0.99	0.98
DoS	0.99	0.99	0.99	0.99	0.99	0.99
Probe	0.99	0.99	0.99	0.99	0.98	0.99
R2L	0.92	0.97	0.91	0.95	0.92	0.96
U2R	0.46	0.86	0.46	0.77	0.40	0.84

由表 3 可见:对于攻击类别为 R2L 的网络流量类型,本文算法召回率达到 0.91,精确率达到 0.96;对于 U2R 类型的攻击类别,本文算法召回率达到 0.71,精确率达到 0.77,分别为该网络类别最好的划分结果。

由表 4 可见:对于攻击类别为 R2L 的网络流量类型,本文算法召回率达到 0.92,精确率达到 0.96;对于 U2R 类型的攻击类别,召回率达到 0.52,精确率达到 0.86;在 U2R 上,IET 算法的表现介于 ET 算法和 RF 算法之间,RF 算法的精确率要高于 IET 算法,但是 IET 算法的召回率要高于 RF 算法,而 IET 算法与 ET 的算法相比,召回率稍低,但是精确率较高。

由表 5 可见:在 R2L 类别上,本文算法召回率达到 0.92,精确率达到 0.97;在 U2R 类别上召回率为 0.46,精确率为 0.86。

在建模时间上,本文算法和原始算法在使用实现 4 个线程同时工作时,设定基分类器的个数为 180 个,选不同的数据集进行实验,训练时间和预测时间的时间控制在 1 min 以内,结果如图 3 所示。可以看出,本文算法建模时间始终保持在极端随机树和随机森林之间,能够保证算法在时间上的可行性。

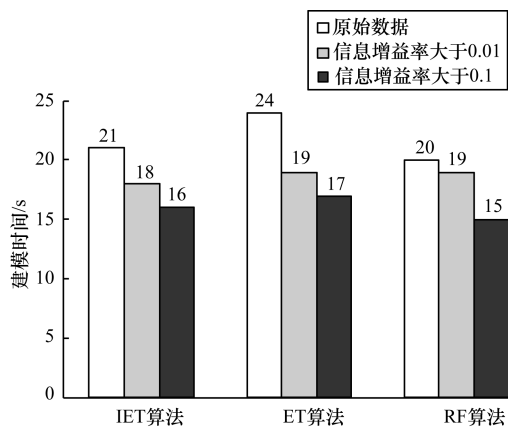


图 3 3 种算法在不同数据集上的建模时间

图 4 显示了特征维度与总体分类精确率的关系。可以看出,本文算法在阈值为 0.01 的数据上,整体分类准确率的表现比原始数据维度的结果更优,达到 0.995 6,而当筛选的阈值为 0.1 时,分类结果相比于原始数据与阈值为 0.01 时略差一些。

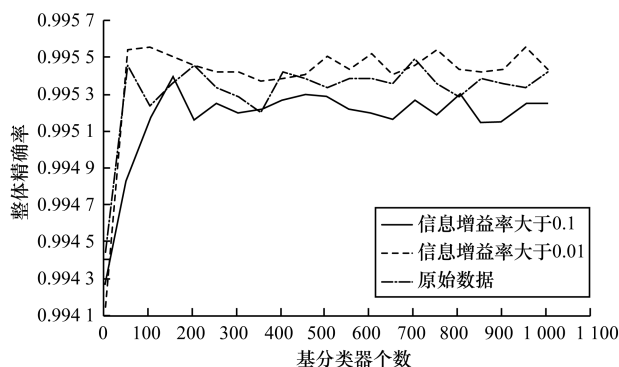


图 4 基分类器个数与精确率的关系

图 5 显示了基分类器个数与时间的关系。可以看出,本文算法在设定阈值为 0.01 时,建模时间相对更短,比选择筛选阈值为 0.1 的数据建模时间更长,表明在异常网络流量的特征选择过程中,使用信息增益率进行特征选择是有效且可行的。

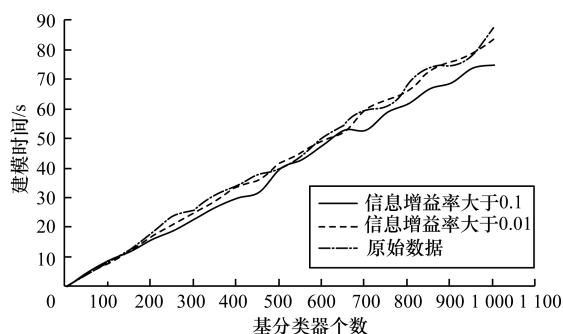


图 5 基分类器个数与建模时间的关系

5 结束语

为提高对少量异常网络流量样本的分类准确率,本文提出一种基于改进极端随机树的分类方法。首先在特征选择部分约减数据维度,除去不相关的数据特征,然后基于改进的极端随机树构建异常流量分类模型。实验结果表明,本文方法在重组的 NSL-KDD 数据上可达到 0.995 6 的分类精确率,在异常流量分类中相对于其他集成分类方法准确度更高,同时在少量样本条件下的分类结果表现更优。

本文算法尚有一些改进的地方,例如在原始数据样本上与其他算法相比,算法的鲁棒性相对较好,但不显著,原因在于数据分布很不均衡。未来将在实际的网络流量上验证本文算法,进一步提升模型的性能,同时对其进行并行化训练和预测,减少训练和预测时间。

参考文献

- [1] HMEDM A, MAHMOOD A N, HU J. A survey of network anomaly detection techniques [J]. Journal of Network and Computer Applications, 2016, 60: 19-31.
- [2] KIM G, LEE S, KIM S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection [J]. Expert Systems with Applications, 2014, 41(4): 1690-1700.
- [3] VASAN K K, SURENDIRAN B. Dimensionality reduction using principal component analysis for network intrusion detection [J]. Perspectives in Science, 2016, 8 (C): 510-512.
- [4] 李 强, 严承华, 朱 瑶. 基于决策树的网络流量异常分析与检测 [J]. 计算机工程, 2012, 38(5): 92-95.
- [5] AMMAR A. A decision tree classifier for intrusion detection priority tagging [J]. Journal of Computer and Communications, 2015, 3(4): 52-58.
- [6] 许 倩, 程东年. 基于层次聚类的网络流量异常分类算法 [J]. 计算机工程, 2012, 38(23): 131-136.
- [7] HORNG S J, SU M Y, CHEN Y H, et al. A novel intrusion detection system based on hierarchical clustering and support vector machines [J]. Expert Systems with Applications, 2011, 38(1): 306-313.
- [8] MALIK A J, SHAHZAD W, KHAN F A. Network intrusion detection using hybrid binary PSO and random forests algorithm [J]. Security and Communication Networks, 2015, 8(16): 2646-2660.
- [9] WANG H, GU J, WANG S. An effective intrusion detection framework based on SVM with feature augmentation [J]. Knowledge-Based Systems, 2017, 136: 130-139.
- [10] RAMAN M R G, SOMU N, KIRTHIVASAN K, et al. An efficient intrusion detection system based on hypergraph-genetic algorithm for parameter optimization and feature selection in support vector machine [J]. Knowledge-Based Systems, 2017, 134: 1-12.
- [11] 马林进, 万 良, 马绍菊, 等. 基于词袋模型聚类的异常流量识别方法 [J]. 计算机工程, 2017, 43(5): 204-209.
- [12] ABUROMMAN A A, REAZ M B I. A survey of intrusion detection systems based on ensemble and hybrid classifiers [J]. Computers and Security, 2017, 65: 135-152.
- [13] JI S Y, JEONG B K, CHOI S, et al. A multi-level intrusion detection method for anomaly network behaviors [J]. Journal of Network and Computer Applications, 2016, 62: 9-17.
- [14] 杨宏宇, 徐 晋. 基于改进随机森林算法的 Android 恶意软件检测 [J]. 通信学报, 2017, 38(4): 8-16.
- [15] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees [J]. Machine Learning, 2006, 63(1): 3-42.
- [16] TAVALLAEI M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set [C]//Proceedings of IEEE International Conference on Computational Intelligence for Security and Defense Applications. Washington D. C., USA: IEEE Press, 2009: 53-58.
- [17] PEDREGOSA F, GRAMFORT A, MICHEL V, et al. Scikit-learn: machine learning in Python [J]. Journal of Machine Learning Research, 2013, 12(10): 2825-2830.
- 编辑 金胡考
- ~~~~~
- (上接第32页)
- [3] JIANG Liangxiao, CAI Zhihua, ZHANG H, et al. Naive Bayes text classifiers: a locally weighted learning approach [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25(2): 273-286.
- [4] HE Y L, WANG R, KWONG S, et al. Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis [J]. Information Sciences, 2014, 259: 252-268.
- [5] TAHERI S, YEARWOOD J, MAMMABOV M, et al. Attribute weighted naive Bayes classifier using a local optimization [J]. Neural Computing and Applications, 2014, 24(5): 995-1002.
- [6] 王 辉, 黄自威, 刘淑芬. 新型加权粗糙朴素贝叶斯算法及其应用研究 [J]. 计算机应用研究, 2015, 32(12): 3668-3672.
- [7] 董立岩, 隋 鹏, 孙 鹏, 等. 基于半监督学习的朴素贝叶斯分类新算法 [J]. 吉林大学学报(工学版), 2016, 46(3): 884-889.
- [8] 李楚进, 付泽正. 对朴素贝叶斯分类器的改进 [J]. 统计与决策, 2016(21): 9-11.
- [9] 刘月峰, 苑江浩, 张晓琳. 改进 NB 算法在垃圾邮件过滤技术中的研究 [J]. 微电子学与计算机, 2017, 34(4): 115-120.
- [10] 杨 雷, 曹翠玲, 孙建国, 等. 改进的朴素贝叶斯算法在垃圾邮件过滤中的研究 [J]. 通信学报, 2017, 38(4): 140-148.
- [11] 段宏湘, 张秋余, 张墨逸. 基于归一化互信息的 FCBF 特征选择算法 [J]. 华中科技大学学报(自然科学版), 2017, 45(1): 52-56.
- [12] 徐峻岭, 周毓明, 陈 林, 等. 基于互信息的无监督特征选择 [J]. 计算机研究与发展, 2012, 49(2): 372-382.
- [13] 刘海峰, 姚泽清, 苏 展. 基于词频的优化互信息文本特征选择方法 [J]. 计算机工程, 2014, 40(7): 179-182.
- [14] HODUE N, BHATTACHARYYA D K, KALITA J K. MIFS-ND: a mutual information-based feature selection method [J]. Expert Systems with Applications, 2014, 41(14): 6371-6385.
- [15] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods [J]. Computers and Electrical Engineering, 2014, 40(1): 16-28.
- 编辑 索书志