计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600161

基于随机森林的入侵检测分类研究

曹扬晨[1]　朱国胜[1]　祁小云[2]　邹　洁[1]
1 湖北大学计算机与信息工程学院　武汉 430062
2 湖北大学化学化工学院　武汉 430062
(943407866@qq.com)

摘　要　为了有效地检测网络的攻击行为,机器学习被广泛用于对不同类型的入侵检测进行分类,传统的决策树方法通常用单个模型训练数据,容易出现泛化误差大、过拟合的问题。为解决该问题,文中引入并行式集成学习的思想,提出基于随机森林的入侵检测模型,由于随机森林中每棵决策树都有决策权,因此可以很好地提高分类的准确性。利用NSL-KDD数据集对入侵检测模型进行训练和测试,实验结果表明,该模型的准确率可达99.91%,具有非常好的入侵检测分类效果。

关键词:入侵检测;机器学习;随机森林;决策树
中图法分类号　TP181

Research on Intrusion Detection Classification Based on Random Forest

CAO Yang-chen[1],ZHU Guo-sheng[1],QI Xiao-yun[2] and ZOU Jie[1]
1 School of Computer and Information Engineering,Hubei University,Wuhan 430062,China
2 School of Chemistry and Chemical Engineering,Hubei University,Wuhan 430062,China

Abstract　In order to effectively detect the attack behavior of the network,the machine learning method are widely used to classify different types of network intrusion detection. The traditional decision tree methods usually use a single model to training data,which is prone to generalization errors and is prone to over-fitting. To solve this problem,this paper introduces the idea of parallel integrated learning,and proposes an intrusion detection model based on random fo-rest. Since each decision tree in the random forest has decision-making power,it can improve the accuracy of classification very well. By using the NSL-KDD data set to train and test the intrusion detection model,the experimental results show that the accuracy rate can reach 99.91%,which shows that the model has a very good intrusion detection classification effect.

Keywords　Intrusion detection,Machine learning,Random forest,Decision tree

随着互联网的快速发展,新的网络技术数见不鲜,数以千计的软件出现在人们眼前,为传统行业注入了新的力量,给人与人之间的交流、出行、购物和搜索等都带来了极大的便利。在这样的大环境下,网络技术的高速发展一方面推动了社会经济的迅速发展,另一方面也给人类社会带来了前所未有的挑战。网络空间以流量来承载大量有价值的信息,而网络入侵却不断带来网络安全问题,网络入侵可以通过盗取系统管理员密码的方法,来访问之前无权访问的文件或数据,从而控制该主机、修改系统文件、破坏系统的机密性和完整性,不仅如此,攻击者可使用拒绝服务攻击耗尽目标主机的资源、占用网络带宽、破坏系统的可用性,给系统和网络安全构成了严重威胁,对我们的隐私和财产安全也带来了巨大的影响。面对如此复杂的网络环境下的攻击,我们如何及时发现黑客的攻击行为,尽可能抵御并减少网络上的恶意攻击、维护网络安全,使网络流量的监控和入侵检测变得越发重要,建立正确率高的流量检测模型也成为了我们关注的焦点。

入侵检测技术可通过识别网络流量数据中的攻击来保障网络空间的安全。主机型和网络型的入侵检测系统一般比较常见,以日志为数据源的主机型入侵检测系统能很好地识别分析、紧密关注特殊主机事件且成本低廉,可以检测到攻击但不快捷;而网络型入侵检测系统一般通过捕获网络流量的数据包作为数据源,将系统放置在网关或者防火墙之后,对所有的数据包起到监视的作用。

本文提出基于随机森林的入侵检测方法,使用经典的网络入侵检测数据集进行网络流量的攻击识别,选取的特征主要有连接的基本特征、连接的内容特征、在一定时间内的网络流量特征,以100条连接为窗口、同一台主机的连接数量特征建立模型训练数据,通过实验验证了该算法的有效性。

本文第1节介绍了入侵检测技术的现状,对各种检测和分类模型的准确率进行了比较;第2节给出了本文提出的基于随机森林的入侵检测方法,包括模型与算法步骤;第3节进

基金项目:赛尔网络下一代互联网技术创新项目;基于Cloud VR和IPv6的特殊作业教育培训系统项目(NGII20180507)
This work was supported by the CERNET Innovation Project and Special Operation Education and Training System Based on Cloud VR and IPv6 (NGII20180507).
通信作者:朱国胜(zhuguosheng@hubu.edu.cn)

行了实验，即通过经典的入侵检测数据集对检测方法进行校验与评估。

1 网络入侵检测技术的现状

近年来，机器学习的分类、聚类、降维、回归[1]的方法被广泛用于入侵检测研究，如何建立正确的入侵检测模型成为了研究热点之一，相关的文献也层出不穷。

文献[2]提出基于支持向量机的网络入侵检测模型，此方法的误报率高，对于数据较小的样本，算法性能下降明显，需对其参数进行调试，工作量较大。文献[3]提出一种基于改进BP神经网络的入侵检测模型，虽然该模型的准确率较高，但是训练的收敛速度较慢。文献[4]提出一种基于卷积神经网络的入侵检测模型，但是深度学习的参数较多，调参工作量大，模型较为复杂。文献[5]将随机森林算法应用到入侵检测系统，对特征进行降维处理并优化各个参数，入侵检测系统的最高准确率可达到95.2%。文献[6]以网页数据为样本，采用K近邻算法，提取近4000种特征，包括包长度、顺序、达到时间等，用100个网页样本数据测试可达到85%的真正率和60%的假正率；文献[7]以网页、网站数据为样本，采用支持向量机算法来提取包大小、顺序、方向等特征，平均真正率高于90%；文献[8]以HTTP、SSL协议数据为样本，采用K-Means算法来提取包个数、字节数、间隔时间、包大小等特征，流平均准确率可达95.1%；文献[9]以HTTP、SMTP协议数据为样本，采用DBSCAN聚类算法来提取包大小、间隔时间、流平均包大小等特征，流准确率大于98%；文献[10]以不同APP数据为样本，采用卷积神经网络来提取HTTP请求的特征，http请求包识别准确率高于97%。文献[11]以ET、VoIP数据为样本，采用C4.5决策树算法来提取含25个包的子流包长度、时间间隔等特征，最高精度达到97%。

由以上分析可知，网络流量分类存在准确率低的问题，因此，本文提出基于并行式集成学习的随机森林模型，利用样本的随机选择和特征的随机选择，解决了传统机器学习算法容易过拟合的问题，而且随机森林中的决策树数量多，不但可以解决单个决策树泛化能力弱的问题，对数量大、维度高的数据也能进行较好的分类，提高了入侵检测的准确性。

2 基于随机森林的入侵检测模型

2.1 随机森林算法简介

2.1.1 决策树

决策树(decision tree)是一种非参数的有监督学习方法，它能从一系列有特征和标签的数据中总结出决策规律，以解决分类和回归的问题[1]。

衡量决策树分类效果的指标为不纯度，通常来说不纯度越低，决策树对训练集的拟合效果就越好。决策树中的每个节点都有自己的不纯度值，其中子节点的不纯度比父节点低，由此可知，叶子节点的不纯度是整个决策树中最低的。有两种常用的计算样本集合不纯度的方法，可以通过计算信息熵或是选择基尼系数来选择划分属性。

假定当前在样本集合D中第k类样本所占的比例为p_k($k=1,2,\cdots,|y|$)，则D的信息熵定义为：

$$Ent(D)=-\sum_{k=1}^{|y|}p_k\log_2 p_k \tag{1}$$

其中，$Ent(D)$的值越小，D的纯度就越高[1]。

数据集D的纯度用基尼系数来度量：

$$Gini(D)=\sum_{k=1}^{|y|}p_k(1-p_k)=1-\sum_{k=1}^{|y|}p_k^2 \tag{2}$$

与信息熵一样，$Gini(D)$越小，数据集D的纯度就越高[1]。

在构造决策树时，随着树深度的增加，节点的熵值不断地下降，下降越快，决策树的深度就越小，这样我们就可以得到一颗高度最矮的树。假定离散属性a有V个可能的取值$\{a^1,a^2,\cdots,a^V\}$，若使用a来对样本集D进行划分，则会产生V个分支节点，其中第v个子结点包含了样本集合D中所有在类别a^v上的样本，表示为D^v，我们可根据式(1)计算出D^v的信息熵，考虑到不同的分支节点所包含的样本数不同，因此给分支节点赋予权重$\frac{|D^v|}{|D|}$，以划分所获得的“信息增益”[1]。

$$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v) \tag{3}$$

一般来说，用于分类的属性a信息增益越大，基于该属性进行划分使得纯度的提升值也越大，但是，在属性下类别数量较多的情况下，会造成信息增益很大但分类效果不好的情况，为了减小这种情况带来的影响，对C4.5决策树算法进行了改进，使用“增益率”来选择最优属性，增益率的定义为：

$$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)} \tag{4}$$

其中，

$$IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|} \tag{5}$$

$IV(a)$称为属性a的固有值，属性a的可能取值数量越多，V越大，$IV(a)$的值就越大[1]，而用信息增益率准则来评估取值数量少的属性的分类效果时，会出现信息增益率很大但分类效果不好的情况，因此C4.5算法并不是直接选择增益率最大的候选划分属性，而是先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的属性[1]。

2.1.2 随机森林

随机森林是通过集成学习的思想将多棵决策树集成的一种算法[1]，集成学习的主要思想为分而治之，主要分为Boosting和Bagging两大流派，Boosting是将弱学习器提升为强学习器的集成方法，用于提高预测精度，Bagging即通过随机采样的方法生成众多并行的分类器，通过少数服从多数的原则来确定最终的分类结果[1]。随机森林利用bagging的思想，先对数据进行随机采样，假如有包含m个样本的数据集D，从数据集D中有放回地抽取m次数据，其中有的数据可能被抽到了几次，有的数据可能没被抽到，样本在m次采样中始终没有被抽到的概率为$(1-\frac{1}{m})^m$，取极限得到$\lim_{m\to\infty}(1-\frac{1}{m})^m=\frac{1}{e}\approx0.368$，由上式可知，初始数据集$D$中约有36.8%的数据未出现在采样集中，这些数据就成为了袋外数据(Out-of-bag，OOB)，可用于测试模型，而约有63.2%的数据会出现在采样集中，其中有一部分数据重复。按照这种方法，我们对数据集D有放回地进行m次采样，重复这个操作T次，则可以得到T个含m个样本的采样集，其中每一个含m个样本的采样集使用一个决策树来构建模型，传统的决策树在属性的选择上通常选用最优的属性，而这种方法存在过拟合的问题，因此随机森林算法在属性的选择上选用随机选择属性的方法，对于

决策树的每个结点的选取，先从该所有属性集合 d 中随机选取包含 k 个属性的子集，然后从子集中选择最优属性用于划分，一般情况下，$k=\log_2 d$。最后，每棵决策树都会基于一个采样集训练出基学习器，再引入集成学习的思想将基学习器进行结合，通过投票选择，按照少数服从多数的原则来确定最终分类的结果[1]，如图 1 所示。

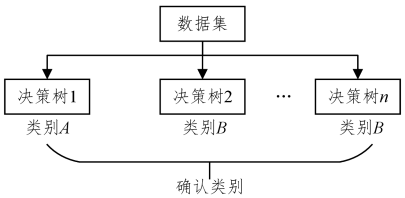

图 1 随机森林简图

Fig. 1 Random forest simplified

在样本和特征选择随机的情况下，随机森林对噪声和异常值有较好的容忍度，提高了模型的泛化性能。

2.2 数据特征的选取

下面详细介绍 NSL-KDD 数据集的特征。

TCP 连接的基本特征共包含 9 项，是判断入侵检测的基础依据，如表 1 所列。

表 1 TCP 连接的基本特征

Table 1 Basic characteristics of TCP connection

No.	Features	Description
1	duration	持续连接时间，单位为 s
2	protocol_type	协议类型(tcp，udp，icmp)
3	service	目标主机的服务类型
4	flag	连接状态
5	src_bytes	源主机到目标主机的字节数
6	dst_bytes	目标主机到源主机的字节数
7	land	连接的源地址和目的地址是否相同
8	wrong_fragment	错误分段的数量
9	urgent	紧急数据包个数

TCP 连接的内容特征有 13 项。对于嵌入在数据包中的数据负载中的攻击，单从数据包的连接特征来看，与普通的数据包并无差异，如 U2R 和 R2L 之类的攻击，需要从 TCP 连接的内容特征中筛选相应内容特征来检测是否存在攻击行为，如表 2 所列。

表 2 TCP 连接的内容特征

Table 2 Content characteristics of TCP connections

No.	Features	Description
1	hot	访问系统敏感文件和目录的次数
2	num_failed_logins	登录失败的次数
3	logged_in	是否登录成功
4	num_compromised	compromised 条件出现的数量
5	root_shell	是否已 root 权限执行 shell
6	su_attempted	是否执行 su root 指令
7	num_root	root 用户访问次数
8	num_file_creations	创建文件的数量
9	num_shells	shell 命令使用次数
10	num_access_files	访问控制文件的次数
11	num_outbound_cmds	一个 ftp 会话出站连接次数
12	is_hot_login	登录用户是否存在于 hot 列表
13	is_guest_login	是否是 guest 用户登录

基于时间的网络流量统计特征有 9 项。网络攻击行为通常在时间上有一定的联系，统计当前连接与之前一段时间的连接之间的关系可以对攻击行为进行较好的识别，以 2 s 为时间节点，统计过去 2 s 内，在与当前连接具有相同目标主机或是有相同服务的连接数量和百分比，如表 3 所列。

表 3 基于时间的网络流量统计特征

Table 3 Statistical characteristics of network traffic based on time

No.	Features	Description
1	count	相同的目标主机的连接数
2	srv_count	相同服务的连接数
3	serror_rate	相同目标主机的连接中，出现“SYN”错误的连接的百分比
4	srv_serror_rate	相同服务的连接中，出现“SYN”错误的连接的百分比
5	rerror_rate	相同目标主机的连接中，出现“REJ”错误的连接的百分比
6	srv_rerror_rate	相同服务的连接中，出现“REJ”错误的连接的百分比
7	same_srv_rate	相同目标主机的连接中，与当前连接具有相同服务的连接的百分比
8	diff_srv_rate	相同目标主机的连接中，与当前连接具有不同服务的连接的百分比
9	srv_diff_host_rate	相同服务的连接中，与当前连接具有不同目标主机的连接的百分比

基于主机的网络流量统计特征有 10 项。对于慢速攻击如 Probe，基于时间的网络流量统计特征已无法很好地识别出相应的联系，而需要以 100 个连接为窗口，统计前 100 个连接中，与当前连接具有相同目标主机或是相同服务的连接数量和百分比，如表 4 所列。

表 4 基于主机的网络流量统计特征

Table 4 Statistical characteristics of network traffic based on host

No.	Features	Description
1	dst_host_count	相同目标主机的连接数
2	dst_host_srv_count	相同目标主机相同服务的连接数
3	dst_host_same_srv_rate	相同目标主机相同服务的连接所占的百分比
4	dst_host_diff_srv_rate	相同目标主机不同服务的连接所占的百分比
5	dst_host_same_src_port_rate	相同目标主机相同源端口的连接所占的百分比
6	dst_host_srv_diff_host_rate	相同目标主机相同服务的连接中，与当前连接具有不同源主机的连接所占的百分比
7	dst_host_serror_rate	相同目标主机的连接中，出现 SYN 错误的连接所占的百分比
8	dst_host_srv_serror_rate	相同目标主机相同服务的连接中，出现 SYN 错误的连接所占的百分比
9	dst_host_rerror_rate	相同目标主机的连接中，出现 REJ 错误的连接所占的百分比
10	dst_host_srv_rerror_rate	相同目标主机相同服务的连接中，出现 REJ 错误的连接所占的百分比

2.3 基于随机森林的入侵检测模型

本文提出的入侵检测算法的训练步骤如下：

(1)首先导入 NSL-KDD 网络入侵数据集的训练集部分，并将其划分为 70%的训练集和 30%测试集。

(2)将非数值型特征进行数值转换。

(3)对于没有关联的特征分类数值进行 one-hot 编码。

(4)调整随机森林中基评估器的数量 *n_estimators*，这个参数对随机森林的精确度的影响是单调的，*n_estimators* 越大，模型的效果往往越好，但是需要的计算量和内存也越大，因此需要调整参数，使得训练难度和模型效果之间取得平衡。

(5)将数据导入随机森林算法中，让模型对网络流量的数据进行训练，这样可得到一个随机森林分类器。

(6)比较不同的 *n_estimators* 对精确度的影响，选择分类

结果最好的一个作为最终的分类器。

本文提出的入侵检测算法的流程如图 2 所示。

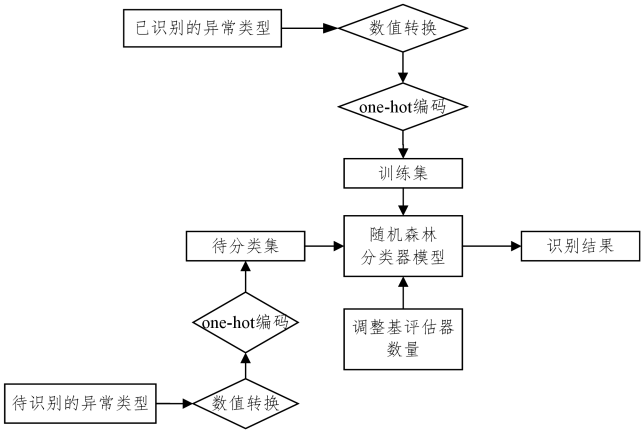

图 2 随机森林识别模型

Fig. 2 Random forest identification model

3 实验与分析

3.1 数据描述

NSL-KDD 数据集是目前最常用的用于入侵检测研究的数据集，数据集中的每条记录均被标注为正常或异常，包括 Normal，Probe，DoS，R2L 和 U2R 5 种类型的数据，训练集中有 22 种攻击类型，总数据规模达到 125 973 个样本。与 KDD99 数据集相比，NSL-KDD 数据集一方面删除了数据集中重复记录的数据，解决了记录冗余的问题，另一方面优化配置，使数据合理分布，解决了训练出的模型对攻击记录数量较少的类别学习能力降低，而对大量相似的数据的分类效果较好的问题。NSL-KDD 数据集已经逐步替代 KDD99 数据集，成为了评估分类模型的性能上应用得最广泛的数据集。

表 5 NSL-KDD 数据集入侵数据分布

Table 5 Intrusion data distribution of nsl-kdd dataset

Intrusion type	Data set
Normal	67 343
Probe	11 656
DoS	45 927
R2L	995
U2R	52

NSL-KDD 数据集有 43 项特征，其中前 42 项与 KDD99 特征相同，第 43 项代表分类的难易程度，数值越大就越容易分类。

3.2 数据预处理

因为 sklearn 规定导入模型的数据只能为数值型，所以利用 Preprocessing. LabelEncoder 模块将分类标签转成分类数值，这里 23 个标签数据编码为 0－22；同理，用 Preprocessing. OrdinalEncoder 将分类特征转换成分类数值，将协议类型、主机的服务类型、连接状态 3 个特征值转换为数值类型，而将分类转换成数字时，如果忽略数字中自带的数学性质，则会给算法传达一些不准确的信息，影响建模过程。因此，我们使用独热编码，将这 3 个特征转换为哑变量覆盖原来的数值，将处理好的数据作为随机森林算法的输入。

3.3 实验过程

实验中将 python 作为编程语言，Jupyter lab 作为训练神经网络的工具，用到的 python 模块包括 Scikit-learn，Pandas，Matplotlib。采用 NSL-KDD 数据集作为训练数据和测试数据。

实验将 NSL-KDD 数据集中的 125 973 条数据分为 70% 的训练集和 30% 的测试集，在进行数据的预处理之后，提取数据集中除标签数据外的 42 项特征，构建随机森林的模型，在随机森林调参上，sklearn 中的 *n_estimators* 为基评估器的数量，*n_estimators* 越大，模型的效果越好，鲁棒性就越强，但是 *n_estimators* 达到一定程度后，随机森林的准确率趋于平稳，不再上升，且 *n_estimators* 越大计算量和内存的消耗也越大。

3.4 结果分析

随机森林的学习曲线如图 3 所示。

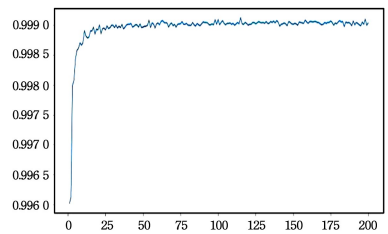

图 3 随机森林学习曲线

Fig. 3 Random forest learning curve

由图 3 可知，当树的数量为 0～15 时，随着随机森林中树的数量的增加，模型分类的准确率呈突发式增长，之后处于一个比较平稳的状态，经过实验可知，准确率最高时，随机森林的决策树个数为 114，达到 99.91%。

随机森林与决策树进行 10 次 10 组交叉验证的效果比较结果如图 4 所示。

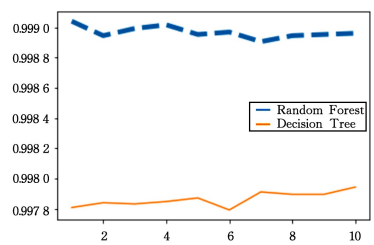

图 4 随机森林与决策树的比较

Fig. 4 Comparison between random forest and decision tree

由图 4 可知，随机森林的预测准确性始终在决策树之上，表现出了非常好的分类性能。

随机森林不仅降低了异常值对模型带来的影响，而且降低了过拟合的可能性。异常值对于单个决策树来说极易导致预测结果不准确，而随机森林采用多棵决策树分类结果投票的方式，使得准确率得到大大提高；随机森林在样本和特征值上的两个随机选择特性，使得决策树容易过拟合的问题也得以解决，极大地增加了模型的泛化性能。

结束语 随机森林在很多方面都有应用，如银行利用随机森林来寻找不同忠诚度的客户，医药行业用随机森林来寻找正确的成分组合以获得新药，随机森林可以对病人的记录进行分析从而确诊病情，随机森林可应用在电子商务的推荐引擎中为客户推荐感兴趣的商品，随机森林在计算机视觉中负责图像的分类，但随机森林在网络流量检测和预测方面的应用还相对较少。本文将随机森林算法应用到网络入侵检测

的分类,理论分析和实验结果表明该方法具有更好的泛化性能,解决了传统机器学习容易过拟合的问题,拥有很高的准确率,可以很好地识别攻击。下一步的工作将基于并行式集成学习随机森林模型建立一个实时、高效的网络流量入侵检测系统。

参 考 文 献

[1] ZHOU Z H. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 27, 75-84, 178-181.

[2] GRIFFITHS W, HAJARGASHT G. On GMM estimation of distributions from grouped data [J]. Economics Letters, 2015, 126: 122-126.

[3] HE W H, LI T S, HUANG R W. Intrusion detection model based on Improved BP algorithm in cloud environment [J]. Computer Technology and Development, 2016, 26(2): 87-90.

[4] WANG M. Network intrusion detection system based on convolutional neural network [D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[5] HOU C, WANG Y, SHAN H, et al. Application and optimization of stochastic forest algorithm in intrusion detection system [J]. Industrial Control Computer, 2019, 32(6): 118-120, 122.

[6] WANG T, CAI X, NITHYANAND R, et al. Effective attacks and provable defense for website fingerprinting [C] // Proc of the 23rd USENIX Security Symposium. 2014: 143-157.

[7] PANCHENKO A, LANZE F, ZINNEN A, et al. Website fingerprinting at Internet scale [C] // Proc of Network and Distributed Sytem Security Symposium. 2016: 1-15.

[8] GLENNAN T, LECKIEC C, ERFANI M S. Improved classification of known and unknown network traffic flows using semi-supervised machine learning [C] // Proc of Australasian Conference on Information Security and Privacy. 2016: 493-501.

[9] XIE G W, ILIOFOTOUS M, FALOUTSOS M, et al. SubFlow: Towards practical flow-level traffic classification [C] // Proc of International Conference on Communications. 2012: 2541-2545.

[10] CHEN Z Y, YU B W, ZHANG Y, et al. Automatic mobile application traffic identification by convolutional neural networks [C] // Proc of IEEE TrustCom/BigDataSE/ISPA. 2016: 301-307.

[11] NGUYEN T T T, ARMITAGE G, BRANCHP, et al. Timely and continuous machine-learning-based classification for interactive IP traffic [J]. IEEE/ACM Transaction on Networking, 2012, 20(6): 1880-1894.

[12] WANG Y S, XIA S T. Overview of stochastic forest algorithm in integrated learning [J]. Information and Communication Technology, 2018, 12(1): 49-55.

[13] FANG K N, WU J B, ZHU J P, et al. Summary of random forest method research [J]. Forum of Statistics and Information, 2011, 26(3): 32-38.

[14] WEI J T, GAO D M. Research on Intrusion Detection System Based on information gain and random forest classifier [J]. Journal of Zhongbei University (Natural Science EditionITION), 2018, 39(1): 74-79, 88.

[15] ZHU K, ZHANG Q. Application of machine learning in network intrusion detection [J]. Data Collection and Processing, 2017, 32(3): 479-488.

[16] ZHAO S, CHEN S H. Overview and Prospect of flow recognition technology based on machine learning [J]. Computer Engineering and Science, 2018, 40(10): 1746-1756.

CAO Yang-chen, born in 1996, postgraduate. Her main research interests include machine learning and network traffic analysis.

ZHU Guo-sheng, born in 1972, Ph. D, professor. His main research interests include next-generation internet and software-defined networks.

(上接第 458 页)

[33] YUAN Y, WANG C R, WANG C, et al. Cloud resource allocation model based on incomplete information game [J]. Computer Research and Development, 2016, 53(6): 1342-1351.

[34] NING P, CUI Y, REEVES D S. Constructing Attack Scenarios Through Correlation of Intrusion Alerts [C] // Proc. of the 9th ACM Conference on Computer and Communications Security. Washington, D. C., 2002: 245-254.

[35] NING P, REEVES D, CUI Y. Correlating Alerts Using Prerequisites of Intrusions: Towards Reducing False Alerts and Uncovering High Level Attack Strategies [R]. North Carolina State University, Department of Computer Science, 2001.

[36] ZHANG H B. Research on IDS alarm correlation model based on description logic [D]. Shanghai: Shanghai Jiaotong University.

YANG Ping, born in 1995, master candidate. Her main research interests include cyber security and reverse engineering.

SHU Hui, born in 1974, postgraduate, Ph. D, professor, Ph. D supervisor. His main research interests include cyber security and reverse engineering.