计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200155

基于 C4.5 决策树的 HTTPS 加密流量分类方法

邹　洁[1]　朱国胜[1]　祁小云[2]　曹扬晨[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
(292370368@qq.com)

摘　要　HTTPS 协议基于原本不具有加密机制的 HTTP 协议。将其与 SSL/TLS 协议组合,在传输数据之前,客户端与服务器端之间进行一次 SSL/TLS 握手,并协商通信过程中使用的加密套件,以安全地交换密钥并且实现双方的身份验证,建立安全通信线路后,对 HTTP 应用协议数据进行加密传输,防止通信内容被窃听和篡改。传统的基于有效载荷的方法已无法处理加密流量,基于流量特征和机器学习的加密流量分类和分析成为目前的主流方法,其通过建立监督学习模型,在保证加密完整性的条件下,基于网络流数据特征工程,应用 C4.5 决策树算法,在局域网环境中对腾讯网中应用 HTTPS 加密数据传输流进行分析,可有效实现对该网站 HTTPS 加密流量进行模块内容的精确分类。

关键词:HTTPS;SSL/TLS;加密流量;决策树;分类
中图法分类号　TP181

HTTPS Encrypted Traffic Classification Method Based on C4.5 Decision Tree

ZOU Jie[1],ZHU Guo-sheng[1],QI Xiao-yun[2] and CAO Yang-chen[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　The HTTPS protocol is based on the HTTP protocol that does not have an encryption mechanism. By combining with the SSL/TLS protocol,an SSL/TLS handshake is performed between the client and the server before the data is transmitted,and the cipher suite used in the communication process is negotiated to securely exchange secret keys and implement mutual authentication. After establishing a secure communication line,the HTTP application protocol data is encrypted and transmitted,preventing the risk of eavesdropping and tampering of the communication content. The traditional payload-based method can't handle encrypted traffic. The classification and analysis of encrypted traffic based on traffic characteristics and machine learning have become the mainstream method. By establishing a supervised learning model,based on network flow data feature engineering,under the condition of ensuring encryption integrity,the C4.5 decision tree algorithm is applied in the LAN environment to analyze the application of HTTPS encrypted data transmission stream in Tencent network,which can effectively realize accurate classification of the website HTTPS encrypted traffic.

Keywords　HTTPS,SSL/TLS,Encrypted traffic,Decision tree,Classification

1　引言

随着互联网被广泛用于网上购物、网上银行、电子交易等商业活动上,数据的价值不言而喻。人们期望网络协议和应用程序通过提供加密、数据完整性来保护关键数据。SSL/TLS 协议套件通常建立在易于理解和彻底分析的加密算法之上,为许多应用程序和协议提供了一定程度的安全性[1]。

作为应用层协议,HTTP 由客户端请求和服务器响应构成,是典型的 C/S 模式,由于 HTTP 协议采用明文传输,当攻击者截获 Web 浏览器与网站服务器之间的传输报文或者冒充 Web 浏览器向服务器发送请求时,传输的信息将不再安全。为了保证数据传输的安全性和完整性,在 HTTP 协议的基础上加入 SSL/TLS 协议构成 HTTPS 协议。经过 TCP 三次握手后,HTTPS 在传输数据之前,客户端与服务器端之间需要进行一次 SSL/TLS 握手,协商通信过程中使用的加密套件,以安全地交换秘钥并且实现双方的身份验证。SSL/TLS 握手成功后,客户端与服务器端进行加密通信,对 HTTP 应用协议数据进行加密传输,HTTPS 加密传输过程如图 1 所示。HTTP 请求处理响应结束后,客户端会发送加密的 alert 消息给服务器端,用来提示服务器端 SSL 传输结束,通过 4 次挥手关闭 TCP 连接。

HTTPS 协议结合对称加密和非对称加密两种加密方式实现数据的安全传输,保护了用户的隐私。本文对于 HTTPS 加密流量的分类研究并不是获取加密的密钥直接对流量进行解密,而是通过监督学习的方法对 Wireshark 采集的数据集进行分析,实现商业网站内容模块的分类,从而获取有用的信息。

基金项目:赛尔网络下一代互联网技术创新项目(NGII20180411)
This work was supported by CERNET Innovation Project (NGII20180411).
通信作者:朱国胜(zhuguosheng@hubu.edu.cn)

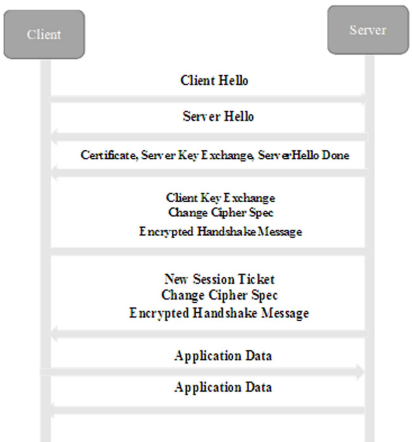

图 1 HTTPS 加密传输过程

Fig. 1 HTTPS encrypted transmission process

2 加密流量分类问题及描述

对网络流量进行加密传输，虽然极大程度地保护了用户隐私，但识别加密流量中的潜在威胁却为网络安全带来了一系列挑战，随着 HTTPS 的使用量超过 HTTP，通过加密网络通道传递的恶意软件将变得越来越多，因此加密流量分类对于有效的网络分析和管理至关重要。传统的基于有效载荷的方法已无法处理加密流量，基于流量特征和机器学习的加密流量分类和分析成为目前的主流方法。

文献[2]结合了基于签名的方法和统计分析方法对加密流量进行分类，通过使用签名匹配方法识别 SSL/TLS 流量，通过真实数据训练贝叶斯分类模型，以获取用于识别不同应用协议的统计信息。实验结果表明，该方法能够识别超过 99%的 SSL/TLS 流量，并且在 F-score 中实现 94.52%的协议识别。文献[3]基于监督决策树方法、半监督 k-means 方法和无监督多目标遗传算法(MOGA)，在 SSH 加密流量中，比较选取了 46 个流量特征，对来自不同网络的一组流量进行跟踪基准测试。实验结果表明，当在与训练网络轨迹相同的网络上捕获的交通轨迹上测试每种方法时，基于多目标遗传算法的训练模型能够在 3 种方法中实现最佳性能；当在与训练轨迹完全不同的网络上捕获的交通轨迹上测试时，C4.5 决策树算法在 3 种方法中实现了最佳性能。文献[4]对超过 6000 个网页进行流量分析攻击，这些网页覆盖了医疗保健、金融、法律服务和流媒体视频等领域中 10 个广泛使用的行业领导网站的 HTTPS 部署，通过构建机器学习模型，对同一网站中的个人页面进行攻击识别，识别出访问的特定页面，其准确率达到 89%，进而能推断出更多的个人信息。文献[5]提出了一种基于统计协议标识(SPID)的 Skype 加密流量分类方法，通过分析流量和应用层数据的统计值，对通过 TLS 协议进行隧道传输的加密 Skype 服务 TCP 流进行了分类。文献[6]基于 K 近邻算法选取了 4000 个流量特征，将 100 个受监控的网站作为目标，来识别用户的网络活动，确定用户正在访问 100 个受监控网页中的具体网页，其真实阳性率达到 85%，而假阳性率仅为 0.6%。文献[7]提出了基于支持向量机的加密流量识别方法 SVM-ID，其选取 50 个加密通信流和 100 个正常数据流组成的训练样本对 SVM 模型进行训练，构建 SVM 分类器后再选取 73 个加密通信流和 276 个正常数据通信流，使用 SVM-ID 方法对这些会话数据流进行检测。其查准率和查全率分别达到了 94.03%和 91.31%。文献[8]提出了一种快速的网络流量识别方法，该方法对加密流量进行载荷特征提取，基于决策树算法对加密流量进行识别，实现了加密网络应用协议和部分恶意流量的准确分类。

政府和企业对大规模电子监控和数据挖掘的研究推动了 HTTPS 的普及，HTTPS 采用混合加密技术，中间无法直接查看明文内容，通过抓包可以看到数据不是明文传输的[9]。为了方便用户浏览信息，更加迅速地找到感兴趣的模块内容，网站服务器通常会将所属资源进行模块分类以达到简化用户检索的目。在一个局域网环境中，当用户访问某一特定网站时，我们将用户发起的请求进行抓包，筛选出需要的数据包并对其进行一定的处理，通过建立监督学习模型，对这些加密的数据包进行分类，从而知道用户访问已知网站的相应模块内容。

3 加密流量分类方法

在流量分类问题上，朴素贝叶斯算法(NB)、支持向量机算法(SVM)和 C4.5 决策树算法(C4.5)是经常使用的机器学习算法。

朴素贝叶斯方法是一种基于概率的参数估计方法，该方法应用的前提在于：待分类样本的先验概率在样本空间保持稳定。然而，在真实的网络环境中，网络流样本的分布是动态变化的，应用朴素贝叶斯方法的前提条件无法满足[10]。文献[10]从理论上证明了朴素贝叶斯算法在流量分类问题上具有潜在的不稳定性，通过在 Moore_Set 和 CAS_Set 数据集上的对比实验表明，利用 C4.5 决策树来处理流量分类问题在分类稳定性上具有明显的优势。

支持向量机算法是一种以统计学习理论为基础的机器学习方法，处理问题时通过使用非线性变换技术，将样本空间的分类问题转换为高维特征空间的分类问题，并在高维空间中构造线性判别函数取代原空间的非线性判别函数，遵循结构风险最小化原则，将分类问题转化为在约束条件下的最优超平面的二次寻优问题[11]。通过实验发现，支持向量机方法在本实验采集的数据集上对加密流量进行分类相对 C4.5 决策树算法的准确率较低，在某些类型的加密数据流上的查全率和查准率过低，因此，本实验重点讨论基于 C4.5 决策树的 HTTPS 加密流量分类方法。

一般来说，一棵决策树由一个根节点、若干个内部节点和若干个叶节点组成。其中，根节点包含样本全集，叶节点对应于决策的结果，其他节点对应于一个属性测试，每个节点包含的样本集合根据属性测试的结果被划分到子节点中，从根节点到每个叶节点的路径对应一个判定测试序列。

为了产生一棵泛化能力强的决策树，在决策树划分过程中，我们需要选择最优划分属性使决策树分支节点的纯度尽可能高，所包含的样本尽可能属于同一类别。信息熵是度量样本集合纯度的常用指标，假设样本集合 S 中第 k 类样本所占比例为 $p_k(k=1,2,\cdots,|y|)$，则信息熵定义为：

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

信息熵的值越小，样本集合的纯度越高。

C4.5 决策树算法使用增益率来选择最优划分属性，假定离散属性 A 有 V 个可能的取值$\{A_1,A_2,\cdots,A_v\}$，若使用属性 A 对样本集 S 进行划分，则会产生 V 个分支节点，其中第 v

个分支节点包含了 S 中在属性 A 上取值为 A_v 的样本，记为 S_v，将每个分支节点的权重记为 $\frac{|S_v|}{|S|}$，则增益率定义为：

$$GR(S,A)=\frac{G(S,A)}{IV(A)} \tag{2}$$

其中，$G(S,A)$ 表示用属性 A 对样本集合 S 进行划分所获得的信息增益，$IV(A)$ 表示属性 A 的固有值。

$$G(S,A)=ENT(S)-\sum_{v=1}^{V}\frac{|S_v|}{|S|}ENT(S_v) \tag{3}$$

$$IV(A)=-\sum_{v=1}^{V}\frac{|S_v|}{|S|}\log_2\frac{|S_v|}{|S|} \tag{4}$$

C4.5 决策树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，减少了信息增益准则对可取数值数目较多的属性有所偏好而可能带来的不利影响[12]。

4　加密流量分类模型

4.1　模型构建

根据 HTTPS 加密流量分类的实际问题，对可以获取到的原始数据进行采集，同时利用已经确定的标签数据，提取出决策树样本全集，通过特征工程对原始数据进行预处理，筛选出较为显著的特征。将样本全集根据合适比例划分为训练集和测试集后对模型进行训练，通过测试集验证模型的有效性，得到决策树分类模型。决策树分类模型如图 2 所示。

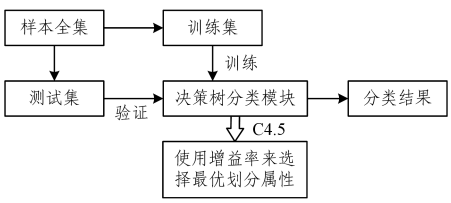

图 2　决策树分类模型

Fig. 2　Decision tree classification model

4.2　参数设置

利用 python 的第三方机器学习模块 scikit-learn 包提供 DecisionTreeClassifier()函数提供的算法构建决策树。将特征选择标准 criterion 设为信息熵 entropy，则得到对 ID3 算法优化过的 C4.5 决策树算法。C4.5 将训练的树(即 ID3 算法的输出)转换成 if-then 规则的集合，然后评估每个规则的准确性，以确定应用它们的顺序。DecisionTreeClassifier()函数并没有直接提供相关剪枝算法，而是通过一系列参数的设置达到剪枝的效果。其主要参数设置如表 1 所列。

表 1　参数设置

Table 1　Parameter settings

Parameter	Description	Settings
criterion	特征选择标准	entropy
splitter	特征划分点选择标准	best
max_features	划分时考虑的最大特征数	None
max_depth	决策树最大深度	7
min_samples_split	内部节点再划分所需的最小样本数	2
max_leaf_nodes	最大叶子节点数	None
class_weight	类别权重	None
presort	数据是否预排序	False

由于本实验模型样本量和样本特征不多，因此在特征的所有划分点中选择最优划分点，将划分时考虑的最大特征数、最大叶子节点数设置为默认值 None，内部节点再划分所需最小样本数设置为默认值 2，决策树最大深度设置为 7；因为样本类别分布没有明显的偏倚，将类别权重设置为默认的 None；数据是否预排序设置为默认值 False 不排序。

5　实验环境

在局域网环境中，选择旁路监控模式配置本地端口镜像对通过交换机的网络流量进行监听和分析。旁路镜像端口的设置如图 3 所示，通过配置交换机将一个或多个端口的数据转发到镜像端口上，镜像端口将收发的报文转发到观察端口，将监控设备直接连接在观察端口上，观察端口将复制过来的报文发送给监控设备。此时，在不改变现有网络拓扑结构、不影响网络通讯速度的情况下，用一台设备对整个局域网的网络流量进行抓取，即使旁路监控设备出现故障或者停止运行，现有网络也不会受到影响。

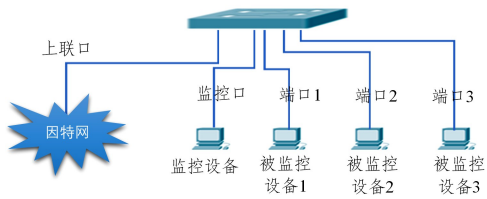

图 3　旁路镜像端口设置

Fig. 3　Bypass mirror port settings

5.1　实验数据说明与处理

我们在监控设备上通过网络封包分析软件对 HTTPS 加密流量进行抓取，获取 SSL/TLS 握手成功后加密传输的 HTTP 应用数据。

本实验以腾讯网[1)]为数据采集网站，对该网站的娱乐、时尚、财经、军事、体育及科技模块内容进行访问，采集了 3 835 条网络流样本作为实验的数据集，该数据集被分为 6 个类型，每类网络流的数目和所占比例如表 2 所列。

表 2　不同网站模块类型的网络流数目及比例

Table 2　Network flow number of each type of module and its proportion

Type	Num	Percent/%
Ent	578	15.072
Fashion	717	18.696
Finance	834	21.730
Military	470	12.256
Sport	500	13.038
Tech	736	19.192
Total	3835	100.000

将 Wireshark 捕获的 pcap 数据包解析成 json 数组并存入表格文件中，通过数据预处理对属性值进行规范化，去掉网络流数据中属性值相同的特征，对缺失值进行处理，用十六进制表示 tcp 有效荷载，并将应用数据转换成相应的字符串长度。

对网络数据流中的每个属性进行具体含义的分析，选取网络数据流中的时间相关特征、网络数据包中与字节长度相关的分组长度特征及端口特征得到本实验数据集包含的 15

1) https://www.qq.com

项网络流属性，具体的属性说明如表 3 所列。

表 3 网络流属性及描述

Table 3 Network flow attributes and descriptions

No.	Attribution	Description
1	frame_time_delta	Time delta from previous captured frame
2	frame_time_delta_displayed	Time delta from previous displayed frame
3	frame_len	Frame length
4	frame_cap_len	Frame capture length
5	tcp_srcport	Source port
6	tcp_stream	Stream index
7	tcp_seq	Sequence number
8	tcp_nxtseq	Next sequence number
9	tcp_ack	Acknowledgment number
10	window_size_value	Window size value
11	window_size	Calculated window size
12	initial_rtt	Initial Round-Trip Time
13	tcp_payload	The length of TCP payload
14	record_length	The length of record
15	app_data	The length of application data

5.2 模型性能评估指标

对决策树泛化能力进行评估，需要有一定的评价标准，假定样本全集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中，$y_i$ 表示 x_i 的实际标记。假设决策树预测的结果是 p_i，则分类正确的样本数占样本总数的比例精度定义为：

$$acc=\frac{1}{n}\sum_{i=1}^{n}II(p_i=y_i) \tag{5}$$

对于多分类问题，令 TP_i 表示实际类型为 i 的样本中被分类模型预测正确的样本数，FP_i 表示实际类型为非 i 的样本中被分类模型误判为类型 i 的样本数量，FN_i 表示实际类型为 i 的样本中被分类模型误判为其他类型的样本数，则类型 i 的查准率和查全率分别定义为：

$$P=\frac{TP_i}{TP_i+FP_i} \tag{6}$$

$$R=\frac{TP_i}{TP_i+FN_i} \tag{7}$$

类型 i 的查准率表示类型 i 中被预测正确的样本数在样本全集中被预测为类型 i 的样本数中所占的比例，查全率表示类型 i 中被预测正确的样本数在样本全集中实际类型为 i 的样本数中所占的比例。虽然精度是分类任务中常用的性能度量，但是并不能满足所有的任务需求，当精度不足时，就需要查准率、查全率或其他的性能度量。

5.3 实验结果

我们采用留出法将样本全集划分为两个互斥的集合，一个作为训练集，一个作为测试集。为了保持数据分布的一致性，需要保证样本中各类别比例相似，我们采取分层采样的方式，将大约 1/5～1/3 的样本作为测试集，由于测试集样本数目不同，分层采样下的精度也不同，如图 4 所示。

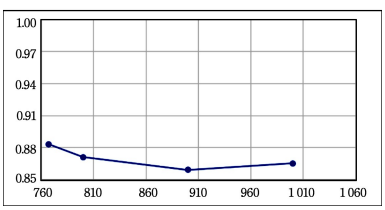

图 4 分层采样下精度变化折线图

Fig. 4 Line chart of accuracy changes under stratified sampling

可以看到，分层采样下的精度并不会随着测试集数目的增加而持续减小，在实验过程中选择合适比例的训练集和测试集对实验结果有一定的影响。当训练集过大时，可能出现过拟合的情况，精度反而会降低。

当选择测试集样本数目为 767 时，网站各类型模块的查准率和查全率如表 4 所列。

表 4 网站各类型模块的查准率和查全率

Table 4 Precision and recall rate of each type of module on website

Type	Support	*Precision*	*Recall*
Ent	102	0.85	0.70
Fashion	173	0.90	1.00
Finance	162	0.82	0.88
Military	93	0.96	0.80
Sport	110	0.98	0.85
Tech	127	0.86	0.98

当已知测试集中类型 i 的支持数时，根据类型 i 的查准率和查全率，可以计算出类型 i 中预测正确的结果数目以及其他类型中被错误预测为类型 i 的样本数目。从表 4 可以看到，类型为 Fashion 的样本集召回率达到了 1，说明该类型样本集全部被正确预测，而其他类型被错误预测为该类型的样本数目仅为 19。从表中数据可以看出，网站各类型模块的准确率和召回率很难得到完美兼顾，因此，在实际应用中，应该根据对查准率和查全率的重视程度选择合适的性能度量。

结束语 随着机器学习的兴起和广泛应用，且传统破解并解密网络流量的方法需要部署额外设备导致成本和部署难度较高，基于流量特征和机器学习的加密流量分类和分析成为目前的主流方法。本实验通过 C4.5 决策树监督学习算法，基于网络流量的特征，对加密流量进行所属网站模块分类，准确率达到了 88.3%。

相对于一些经典网络流量数据集来说，本实验的样本数目远远不够，下一步会继续采集样本数据，不断完善本实验，尝试将用于显示数据包在物理层上传输的十六进制数据转换成合乎要求的十六进制位图数据，进而将其压缩成图片进行分析。

参 考 文 献

[1] HOLZ R, BRAUN L, KAMMENHUBER N, et al. The SSL Landscape: A Thorough Analysis of the X.509 PKI Using Active and Passive Measurements[C]// Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC '11). New York, NY, USA, ACM, 2011: 427-444.

[2] SUN G, XUE Y, DONG Y, et al. An Novel Hybrid Method for Effectively Classifying Encrypted Traffic[C]// Global Telecommunications Conference (GLOBECOM 2010). IEEE, 2010: 1-5.

[3] ARNDT D J, ZINCIR-HEYWOOD A N. A Comparison of Three Machine Learning Techniques for Encrypted Network Traffic Analysis[C]// IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA). 2011: 107-114.

[4] MILLER B, HUANG L, JOSEPH A D, et al. Tygar. I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis[C]// Privacy Enhancing Technologies, volume 8555 of Lecture Notes in Computer Science. Springer Interna-

tional Publishing，2014：143-163.
[5] KORCZYNSKI M A. Classifying Service Flows in the Encrypted Skype Traffic[C]// 2012 IEEE International Conference on Communications (ICC). 2012：1064-1068.
[6] WANG T，CAI X，NITHYANAND R，et al. Effective attacks and provable defenses for website fingerprinting[C]// 23rd {USENIX} Security Symposium ({USENIX}. 2014：143-157.
[7] CHENG G，CHEN Y X. Encrypted Traffic Identification Method Based on Support Vector Machine[J]. Journal of Southeast University(Natural Science Edition)，2017(4)：655-659.
[8] CHEN W，HU L，YANG L. Fast Identification Method of Encrypted Traffic Based on Load Characteristics[J]. Computer Engineering. 2012(12)：22-25.
[9] ZHANG B Y. Analysis of the Principle and Application of HTTPS Protocol[J]. Network Security Technology and Application，2016(7)：36-37.
[10] XU P，LIN S. Traffic Classification Method Based on C4.5 Decision Tree [J]. Journal of Software，2009(10)：2692-2704.
[11] LIU K. Research on feature selection in network flow classification [D]. Yangzhou：Yangzhou University，2013：18-19.
[12] ZHOU Z H. Machine Learning [M]. Beijing：Tsinghua University Press，2016：73-79.

ZOU Jie，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 368 页）

测效果的比较，随机森林算法的分类效果要好于 C4.5 决策树，但这并不否定决策树的分类作用，因为不同的数据挖掘算法都有其优势和劣势，我们应当从微博用户的各类特征出发，在提取有效特征的基础上与传统数据挖掘算法结合使用，以起到仅使用其中一种方法所不能达到的良好效果。

结束语　随着对微博用户行为研究的不断深入，对微博上的大量异常用户的检测与识别也引起了学术界和产业界的重视。本文基于用户特征对微博异常用户分类方法进行分析，在获取到的微博用户数据的基础上，通过对用户基本属性、行为模式和文本内容的分析，提取相关特征并进行聚合，完成对数据的预处理。随后选取 Weka 作为分析工具，以微博用户数据集为样本，对微博的传播特点以及用户属性、文本内容等特征进行综合考量，将数据挖掘算法融入对新浪微博异常用户的检测方法当中，建立起微博异常用户检测模型，先后选取 C4.5 决策树和随机森林分类算法完成了模型训练和实验预测，并对其准确率进行了评估。本文结果可以为公安机关开展舆情管控工作提供参考。

参 考 文 献

[1] 中国互联网信息中心. 第 43 次中国互联网络发展状况统计报告[R]. 北京. CNNIC，2019.
[2] FABRICO B，MAGNO G，RODRIGUES T，et al. Detecting spammers and content promoters in online video social networks [C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM，2009：620-627.
[3] BENEVENUTO F，MAGNO G，RODRIGUES T，et al. Detecting spammers on twitter[C]// Collaboration，Electronicmessaging，Anti-Abuse and Spam Conference. Washington，2010：6-12.
[4] STRINGHINI G，KRUEGEL C，VIGNA G. Detecting spammers on socia networks[C]// Proceedings of the 26th Annual Computer Security Applications Conference. ACM，2010：1-9.
[5] 彭希羡，朱庆华，刘璇. 微博客用户特征分析及分类研究——以“新浪微博”为例[J]. 情报科学，2015，33(1)：69-75.
[6] 刘勘，袁蕴英，刘萍. 基于随机森林分类的微博机器用户识别研究[J]. 北京大学学报，2015，51(2)：289-300.
[7] APHINYANAPHONGS Y，RAY B，STATNIKOV A，et al. Text classification for automatic detection of alcohol use-related tweets：A feasibility study[C]// 2014 IEEE 15th International Conference on Information Reuse and Integration(IRI). IEEE，2014：93-97.
[8] 蒋鑫. 基于属性约简的社交网络异常用户识别系统的设计与实现[D]. 北京：北京邮电大学，2016：2-3.
[9] 夏崇欢. 基于行为特征分析的微博恶意用户检测方法[D]. 南京：南京邮电大学，2018：5-6.
[10] 郝亚洲，郑庆华，陈艳，等. 面向网络舆情数据的异常行为识别[J]. 计算机研究与发展，2016，53(3)：611-620.
[11] 张玉清，吕少卿，范丹. 在线社交网络中异常帐号检测方法研究[J]. 计算机学报，2015，38(10)：2011-2027.
[12] 吴大鹏，司书山，闫俊杰，等. 基于行为特征分析的社交网络女巫节点检测机制[J]. 电子与信息学报，2017，39(9)：2089-2096.
[13] 刘琛. 基于行为分析的社交网络异常账号的检测[D]. 北京：北京交通大学，2017.
[14] 孙洋. LBSN 中基于好友聚类的社交搜索系统设计与实现[D]. 南京：东南大学，2017.

YUAN De-yu，born in 1986，Ph.D，lecturer. His main research interests include cyber security，and complex networks.

GAO Jian，born in 1982，Ph.D，lecturer. His main research interests include botnet，malware analysis and cyber crime.