一种针对木马流量的特征选择方法

张　瑜，刘晓洁，李贝贝
(四川大学网络空间安全学院，成都 610065)

摘　要：针对现有基于会话流异常行为的木马检测方法中，普遍存在所选特征代表性不足、特征间信息冗余导致检测效果差的问题，提出一种特征选择方法. 首先，通过捕捉流量对木马通信行为加以分析，根据各阶段提取相关的属性，并在每一属性上进行派生，得到足够充分的特征集合. 然后，为了衡量特征的重要性和特征间的相关性，提出了改进的特征重要性评价系数和基于关联信息熵的联合相关性评价系数，并设计了基于序列后向选择策略的特征选择算法，以得到自适应规模的特征子集. 算法通过每一轮迭代计算特征的评价系数，通过排序完成选择. 为验证该算法有效性，采用朴素贝叶斯分类和支持向量机分类算法设计了与 FCBF 算法和 IG 算法的对比实验，相较于 FCBF 算法，在两种分类算法上的召回率分别提升 3.76%、1.64%，F_1 值提升分别为 1.04、0.99. 相较于 IG 算法，召回率提升分别为 6.46%、4.96%，F_1 值提升分别为 3.56、3.18. 实验结果表明，提出的特征选择算法能够有效选择木马流量各个属性上的特征，克服特征间关联性带来的影响，在缩减特征维度的同时提升木马通信流量的检测效果.

关键词：木马检测；特征选择；标准化互信息；关联信息熵
中图分类号：TP391　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.012004

A feature selection method for remote access trojan's traffic

ZHANG Yu, *LIU Xiao-Jie*, *LI Bei-Bei*
(College of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Existing Trojan detection methods suffer from poor detection performance because of the information redundancy and lack of representative features. In order to solve the problems, a feature selection method was proposed in this paper. Firstly, in order to get sufficient feature set, features were derived from the relevant attributes by analyzing communication behavior of Trojan. Then, in order to measure the importance of features and the correlation between features, the improved evaluation coefficient of feature importance and the joint evaluation coefficient based on correlation entropy are proposed. Afterwards, a feature selection algorithm based on sequential backward selection was designed to obtain a feature subset of adaptive size. The evaluation coefficient of feature was calculated through each iteration, and the selection was done by sorting those features. In order to verify the validity of the proposed method, we combined the method with naive bayes classification and the SVM classification algorithm, respectively, and conducted experiments to compare the two classification algorithms with FCBF and

收稿日期：2020-05-11
基金项目：国家重点研发计划(2020YFB1805400)；中央高校基本科研业务经费(YJ201933)；国家自然科学基金(U1736212, U19A2068, 62032002, 62002248)；中国博士后科学基金特别资助项目(2019TQ0217)；四川省重点研发计划(20ZDYF3145)
作者简介：张瑜(1996－)，男，硕士研究生，研究方向为网络安全技术及应用. E-mail：onlythen@icloud.com
通讯作者：刘晓洁. E-mail：liuxiaojie@scu.edu.cn

IG. Compared with FCBF, the recall rate of the two classification algorithms increased by 3.76% and 1.64%, and the F1 increased by 1.04 and 0.99. Compared with IG, the recall rate increased by 6.46% and 4.96%, and the F1 increased by 3.56 and 3.18, respectively. The results of the experiments showed that the proposed algorithm can effectively select the characteristics of each attribute from Trojan traffic, and overcome the influence of the correlation between features, reducing feature dimension and improving Trojan traffic detection.

Keywords: Trojan detection; Feature selection; Normalized mutual information; Correlation information entropy

1 引　言

网络攻击中,木马作为一种十分隐蔽的恶意程序,常被攻击者用来窃取信息、远程控制他人主机并借此构建僵尸网络来发动大规模的攻击,其中远程控制型木马危害较大,其大多对通信数据进行加密,在目标机器上通过多种方式隐藏自身,检测难度较高. 国家计算机网络应急技术处理协调中心在《2018 年中国互联网网络安全报告》中指出,2018 年境内共有 659208 个 IP 地址的主机被植入木马或僵尸程序,给网民、企业以至国家造成了巨大损失[1].

通过网络会话流的异常行为识别木马是当前的研究热点,通过采集木马流量并统计特征来构建异常检测模型,部署在网络出口节点上,从而实现对未知木马的检测,此方式避免了对加密流量载荷的分析,同时克服了基于主机行为的特征码检测方式[2-3]的滞后性. 目前,对木马通信流量进行异常检测的研究工作主要集中在以下两方面.

(1) 对于木马会话流特征提取阶段的改进:李巍等[4]将木马通信过程划分成建立连接、命令交互、保持连接三个阶段,分别提取出代表性特征后建模验证特征的有效性. Jiang 等[5]提出一种在木马通信早期阶段进行检测的方法,将会话从 TCP 三次连接开始到数据包间隔大于 1 s 这段时期定义为流的早期阶段,通过提取该阶段的特征进行模型构建及识别. 但 UDP 会话无法划定早期阶段,且该研究选择的特征不够具有代表性. 胥攀等[6]在时间维度上对木马通信流进行聚类生成通信流簇,在簇上提取特征能够更精确地描述木马流量. 该方法需要对提取到的数据聚合多次,增加了计算代价且损失了实时性.

(2) 对于检测阶段的改进:兰景宏等[7]提出一种木马流量检测集成分类模型以增加分类精度和泛化能力,先对旋转随机森林算法中的主成分变换进行均值化改进,接着采用此旋转森林算法对原始数据集进行旋转处理,再选取朴素贝叶斯、C4.5 决策树和支持向量机构建集成分类模型. 张兆林等[8]引入人脸识别领域的 Adaboost 算法模型,选择支持向量机、C4.5 决策树和神经网络建立集成分类模型,提高了单一算法的检测效果. 汪洁等[9]提出多层集成分类器的方法检测恶意流量,首先采用无监督学习框架对数据进行预处理并将其聚成不同的簇,并对每一个簇进行噪音处理,然后使用随机森林、bagging 和 Adaboost 构建三层分类器进行检测,达到了较好的检测效果. 此类方法[10-11]选择的特征较少,代表性不足,且存在特征间信息冗余的缺点.

针对以上问题,本文提出一种子集规模自适应特征选择方法. 在提取并派生出充分的特征后,先对提取的特征计算重要性评价系数,接着在每一轮迭代中更新特征的联合相关性评价系数,同时做出排序,使得筛选后的特征具有足够的代表性,并减小子集中特征的冗余,最后选择另外两种特征选择算法在真实木马流量上采用朴素贝叶斯、支持向量机两种分类算法进行对比实验.

2 木马通信行为分析

木马大多采用 C-S 架构部署,服务端运行在受控主机上,客户端运行在控制主机上,这种木马称为远程控制型木马. 在 Windows 平台上,木马具有以下行为:磁盘文件操作,包括远程运行、删除、修改、上传及下载;注册表读写操作;进程管理操作;屏幕监控和鼠标控制;键盘记录及远程操作;远程执行 CMD 命令;摄像头及声音设备控制. 这些行为从网络流的角度可以划分成四类:下行短数据流(如控制命令)、上行短数据流(如命令执行结果)、下行长数据流(如文件传输)、上行长数据流(如屏幕监控),这里的上行指服务端向客户端发送的方向,下行则是指客户端向服务端发送的方向,长短

表示流的持续时间. 在程序通信中，网络数据流指按照五元组（源 IP、目的 IP、源端口、目的端口、协议）对数据包划分后得到的数据包集合，本文将一条网络数据流定义为一条会话，通过对多种木马运行并分析其会话数据后，划分以下 5 类会话属性，共提取 43 个会话特征作为初选特征集，用以描述木马流量与正常流量的差异.

2.1　上下行流特征

相比于正常应用程序，木马服务端作为受控端，提供窃取信息和执行命令的功能，而正常应用程序的网络行为是获取信息和发送请求，反映在流量统计上则是上行流量远高于下行流量，例如攻击者在下载服务端上的文件或监控服务端主机的屏幕时. 例如采集到的正常通信流和木马流在上下行数据量比上的取值分布统计对比（如图 1 所示），从图 1 可以看出木马流量和正常流量的分布差异. 本文在此属性上派生出的 6 个会话特征见表 1.

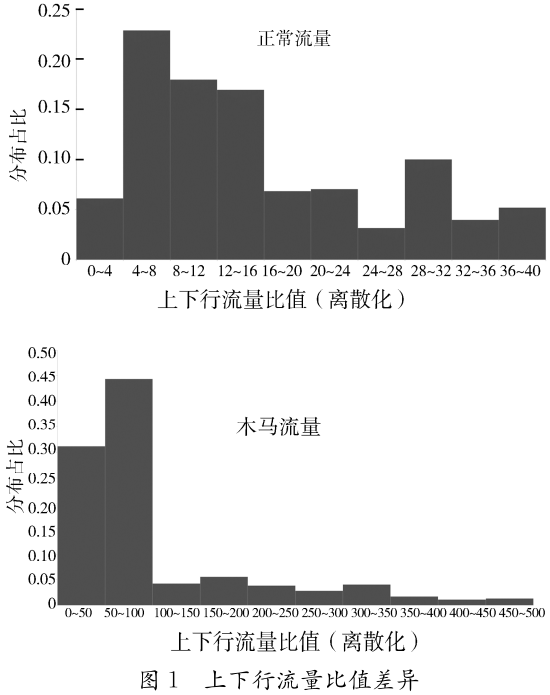

图 1　上下行流量比值差异

Fig. 1　The difference of the ratio of up-stream and down-stream flow size

2.2　上下行数据包特征

在木马连接和通信过程中，控制端会发送大量的命令到服务端执行，服务端会返回执行结果，而命令数据大多是较短指令构成的小数据包（100 字节内），返回的内容大多是大数据包（文件、CMD 返回内容、音视频数据），例如正常流量与木马流量在上行大包数量上的差异如图 2 所示，本文在此属性上派生出的 12 个会话特征见表 2.

表 1　上下行流特征

Tab. 1　The features of up/down stream's traffic

序号	特征名称
1	总上行包数量
2	总上行包长度
3	总下行包数量
4	总下行包长度
5	上下行包数量比
6	上下行数据量比

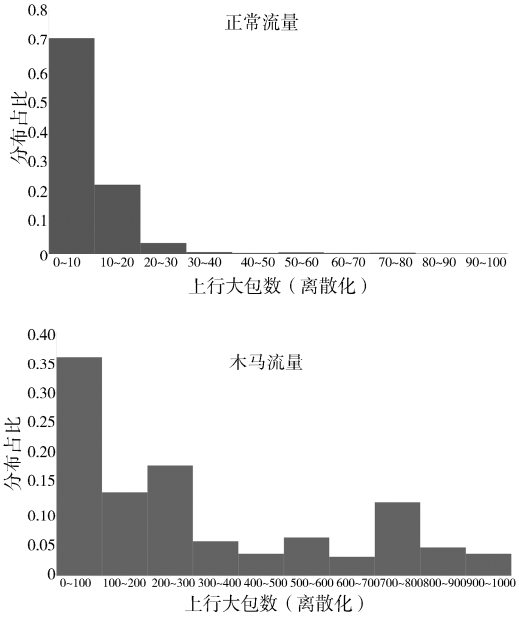

图 2　上行大包数量差异

Fig. 2　The difference of up-stream's big packet count

表 2　上下行数据包特征

Tab. 2　The features of up/down stream's packet

序号	特征名称
7	上行包最小长度
8	上行包最大长度
9	上行包平均长度
10	上行包长度标准差
11	下行包最小长度
12	下行包最大长度
13	下行包平均长度
14	下行包长度标准差
15	下行小包数量
16	上行大包数量
17	下行小包数占小包总数比例
18	上行大包数占大包总数比例

2.3　流标志位特征

木马服务端在通过 DNS 解析到客户端 IP 后，会向该地址不断发送连接请求，直到成功连接到客户端，在这一过程中，服务端会发起大量的 TCP 连接请求，产生了大量的带有 SYN 标志位的数据包. 同时为了使两端的通信延迟更小，发送方会在发送控制数据时将该次连接的 PSH 标志位置 1，这样接收方便会在执行完成后立即返回结果数据，而不必等待其他数据，这也使得会话中带有 PSH 标志位的数据包占比较正常会话高，本文在此属性上提取的两个会话特征见表 3.

表 3　流标志位特征

Tab. 3　The features of traffic's flag

序号	特征名称
19	带 SYN 的包数
20	带 PSH 的包数

2.4　数据包间隔特征

受害主机在接收到客户端发送的控制命令后，需要执行指定的命令，执行完成后再将结果返回给客户端，攻击者在收到数据后，也需要在分析结果后给出下一步攻击命令，这样就带来了较大的数据包处理间隔. 而正常通信流量的数据包间隔往往较小且更稳定，如图 3 所示，流下行包最大间隔差异，本文在此属性上派生出 14 个特征以描述会话流，如表 4 所示.

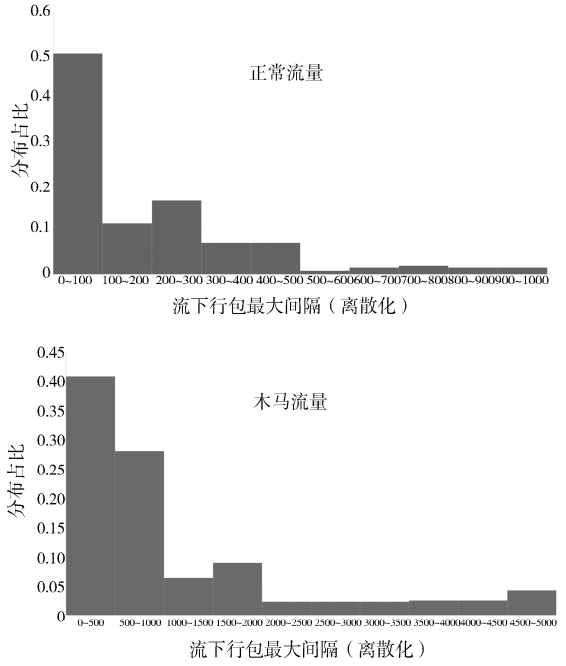

图 3　流下行包最大间隔差异

Fig. 3　The difference of max interval in down-stream

表 4　数据包间隔特征

Tab. 4　The features of packet's time interval

序号	特征名称
21	流发送包间隔的平均值
22	流发送包间隔的标准差
23	流发送包的最大间隔
24	流发送包的最小间隔
25	流上行包最小间隔
26	流上行包最大间隔
27	流上行包平均间隔
28	流发送上行包间隔的标准差
29	流发送上行包间隔的总和
30	流下行包最小间隔
31	流下行包最大间隔
32	流下行包平均间隔
33	流发送下行包间隔的标准差
34	流发送下行包间隔的总和

2.5　会话流基本特征

由于木马攻击活动具有持续性，因此其部分通信连接会保存较长的时间，而正常应用程序出于减小服务器负载的目的会在完成信息传输后断开连接，释放资源，因此大部分正常连接持续时间都短于木马流. 同时为了衡量数据流在时间维度上的差异，本文增加了 9 个会话流基本特征，如表 5 所示.

表 5 会话流基本特征

Tab. 5　The basic features of traffic

序号	特征名称
35	流持续时间
36	平均每秒流的数据包数
37	平均每秒流的字节数
38	平均每秒上行数据包数
39	平均每秒下行数据包数
40	数据包最小长度
41	数据包最大长度
42	数据包长标准差
43	数据包平均长度

3　子集规模自适应特征选择算法

在模式识别中，特征选择作为一种降维方法一直是研究的热点[12-16]，考虑到特征对模型预测能力的影响以及特征间的相关性，通过某种方法从原始特征集合中选择更优的特征子集后，能够在后续

机器学习模型中得到更好的预测效果，同时降低在大规模数据下的计算代价.

按照搜索策略来划分特征选择方法，可以分为采用全局最优搜索的特征选择算法、采用随机搜索策略的特征选择算法和采用序列搜索策略的特征选择算法三类. 其中采用全局最优搜索可以找到最优子集，但计算代价也是最大的，目前使用较广泛的是后两者[17-19]. 若按照特征子集评价标准来划分特征选择方法，主要分为 Filter（过滤法）和 Wrapper（包装法）. 其中，Filter 方法独立于后续机器学习算法的结果，通过某些统计指标来衡量选择的优劣，使用较广泛的指标有特征间距离、特征信息熵等；而 Wrapper 方法将后续采用的机器学习算法的结果作为指标来衡量特征选择的优劣，这种方法与算法结合得更加紧密，但也损失了特征选择的一般性.

本文采用序列搜索中的后向选择策略和 Filter 式的评价标准构造特征选择算法.

3.1　特征重要性及联合相关性度量

本文在后向选择策略的基础上，定义特征重要性评价系数以及特征的联合相关性评价系数. 基于这两系数，本节提出一种特征子集自适应选择算法（Adaptive Feature Subset Selection Algorithm，AFSA），AFSA 算法通过每一轮迭代计算特征间的组合效应，选出最优特征，且能自适应地确定特征数量.

3.1.1　*改进的重要性及联合相关性评价系数*　特征的重要性评价系数指通过该特征识别出某类 C 的能力强弱，重要性评价系数越大，说明通过该特征能够更好地区分类 C 与其他类. 根据香农信息熵理论，若某特征 f 在类 C 上的取值范围较集中，表示其不确定性较小，在类 C 上具有较强代表性，同时，若特征 f 在类 C_1 和 C_2 上的取值分布范围重合区间较小，表示该特征在此两类上分布差异较大，通过特征 f 能够很好地区分类 C_1 和 C_2. 特征重要性评价系数结合了特征 f 的取值集中程度和在不同类上的分布差异.

特征的联合相关性评价系数则用来衡量特征 f 与剩余特征集合的相关性关系，本文采用标准化互信息来计算两两特征间的相关性，若特征 f 与剩余特征相关性较高，且在去除该特征后剩余特征集合内相关性较低，则表明该特征给特征集合带来了较大的冗余信息. 基于以上分析，本节给出以下的定义.

假定有木马流量数据集 S，包含 M 条数据，每条数据由 N 个特征值和一个类别标签构成，广义上有两种类别：木马流量和正常流量，但正常流量间具有差异性，因此本文先对正常流量通过 K-Means 聚类后，根据结果更新正常流量这一类别，同时本文采用 Z-score 方法对数据进行标准化以消除不同量纲的影响.

定义 1　特征集中度 P_{im}，表示特征 f_i 在 C_m 类上的分布集中度.

$$P_{im}=1/(Z_{max}-Z_{min})V_s \tag{1}$$

其中，$Z_{\max}$、$Z_{\min}$ 为标准化后特征最大、最小值；V_s 表示特征取值的离散系数.

定义 2　特征值分布差异 D_{imn}.

$$D_{imn}=n_{im}\cdot n_{in}/n_{imn}{}^2 \tag{2}$$

从图 1～图 3 可以看出，同一特征在两类上取值分布具有差异，其中，n_{imn} 表示两类在同一特征上取值重合区间内样本数；n_{im}、n_{in} 分别表示两类的样本总数.

定义 3　特征重要性评价系数 I_i.

$$I_i = P_{\text{i\_trojan}}\cdot\sum_{k=1}^{M-1} D_{i\_trojan\_k}/(L-1) \tag{3}$$

特征重要性评价系数衡量了特征 f 在木马类别上取值集中程度及与其他类的分布差异，该值越大，表示特征在选择时权重越大.

定义 4　特征联合相关性评价系数 E_i.

该评价系数的思想来源于图像关联分析中的关联信息熵[20]，是一种度量信息冗余的指标，文献[13]引入该思想到特征选择中，相较于文献[13]中提出的关联信息熵公式，本文采用特征间标准化互信息作为矩阵元素，更好地度量特征集整体的相关性. 设有原始木马流量特征集合 $F=\{f_1, f_2, f_3, \ldots, f_N\}$，从中选择特征 f_k 后剩余特征子集 F/f_k，基于特征间的相关关系，构造以下相关性模型 H_k，形式为

$$H_k=\begin{bmatrix} NMI_{k1} & NMI_{12} & \cdots & NMI_{1N} \\ NMI_{21} & NMI_{k2} & \cdots & NMI_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ NMI_{N1} & NMI_{N2} & \cdots & NMI_{kN} \end{bmatrix} \tag{4}$$

例如 $F=\{f_1, f_2, f_3, f_4, f_5\}$ 时 f_2 的相关性模型 H_2 的形式如下.

$$H_2=\begin{bmatrix} NMI_{21} & NMI_{13} & NMI_{14} & NMI_{15} \\ NMI_{31} & NMI_{23} & NMI_{34} & NMI_{35} \\ NMI_{31} & NMI_{43} & NMI_{24} & NMI_{45} \\ NMI_{51} & NMI_{53} & NMI_{54} & NMI_{25} \end{bmatrix} \tag{5}$$

H_k为一个 $N-1$ 阶方阵,矩阵元素 NMI_{ij} 为两个特征间的标准化互信息:

$$NMI(X;Y)=\frac{2\times I(X;Y)}{H(X)+H(Y)} \tag{6}$$

其中,$I(X;Y)$为 X 和 Y 的互信息;$H(X)$和 $H(Y)$为 X 和 Y 的熵,根据性质知 $0\leqslant NMI_{ij}\leqslant 1$,$NMI_{ij}=NMI_{ji}$,那么 H_k为实对称方阵. 对称方阵进行特征分解得到的特征值表示在各个特征向量上矩阵的信息量,而每个特征对相关性影响可以用其特征值表示,假定 H_k存在 K 个正特征值 e_k,定义特征联合相关性评价系数为

$$E_i=-\sum_{k=1}^{K}\frac{e_k}{N-1}\log_{N-1}\left(\frac{e_k}{N-1}\right) \tag{7}$$

当特征 f_k与其他特征完全相关,且特征子集间相互无关时,矩阵 H_k成为单位矩阵 I,单位矩阵的特征值均为 1,根据式(7)可以计算出 E_i为 1,这时将特征 f_k视为带来较大不确定性的特征,在后续选择中权重较低,若特征 f_k与其他特征不相关,此时 E_i为 0,将该特征视为带来较小不确定性的特征,后续选择中权重更高,因此该系数满足特征选择的要求.

3.2　子集规模自适应后向特征选择算法

通过 3.1 节定义的两个评价系数,本文设计了基于序列后向选择的子集规模自适应特征选择算法,特征选择中如何确定移除的特征数量是一个研究热点,而人工设定数量的方式不够灵活,本文算法通过以下策略对子集规模进行控制,如算法 1 所示.

算法 1　特征子集自适应后向选择算法-AFSA

输入　原始特征集合 F,数据集,类别 C.

输出　终选特征子集 S.

1) 遍历 F,计算特征 f 重要性评价系数 I_f;

2) 计算重要性评价系数均值 I_e,将低于均值的特征放到预移除特征集合 F_d中,剩余特征为集合 F_r,$F=F_d+F_r$;

3) 计算 F 的特征间标准化互信息 NMI_{ij};

4) 遍历 F_r,计算每个特征相对于 F_r的联合相关性评价系数 E_{ri},同时计算 F_r联合相关性评价系数均值和重要性评价系数均值的比值 R_{ri}作为参照值,以 F_r中特征的系数比值最小值作为适应值;

5) 遍历 F_d,计算每个特征相对于F_d+F_r的联合相关性评价系数 E_{di},计算联合相关性评价系数均值和重要性评价系数均值的比 R_{di}后做升序排序;

6) 若 F_d中末尾特征 f_{last}的 R_d大于参照值 R_{ri},则在 F_d中移除特征 f_{last},否则结束,若第一轮比较时无可移除特征,那么令 R_{ri}为步骤 4)中的适应值;

7) 若 F_d为空,算法结束,否则回到步骤 5);

8) 结束后输出特征选择结果 F_d+F_r.

由于上述步骤 6)第一次移除时,可能出现无法移除特征的情况,本文的目标是尽可能移除较差作用特征,因此算法考虑对参照值 R_{ri}作一定范围调整,即以 F_r中特征的联合相关性评价系数和重要性评价系数比的最小值作为参照值 R_{ri},若仍然无可移除特征,算法终止,表明原始特征集合较为优异.

3.3　算法复杂度分析

尽管特征选择在整个检测系统只需进行一次,但算法的计算代价也需要尽可能的低. 按照 3.1 节中所述,设有 N 维特征,M 个类别,k 条样本数据,3.2 节算法中计算特征重要性评价系数代价为 $O(NMk)$,两两特征计算 NMI 的计算代价为$O(k^2)$,最坏情况下迭代次数为 F_d,此时总的相关性评价系数计算代价为 $O(N^3\times N)$,由于 $N\ll k$,那么算法时间复杂度为 $O(k^2)$,相较于经典的 mRMR 算法[21]的 $O(N^2k^2)$,本算法计算代价更低.

4　实验测试及分析

为了验证本文提出方法的有效性,本文设计了两组对比实验:(1) 将本文初选特征集和终选特征集与文献[7]中 16 个特征基于相同分类器做实验对比,验证特征提取和特征选择的有效性;(2) 与常用基于信息熵的特征选择算法作对比,验证本文特征选择算法的改进效果. 实验均使用相同的训练集和测试集,采用朴素贝叶斯分类算法和支持向量机分类算法. 这两种算法在相关研究[6-7,10-11]中多被采用,且属于分类算法中原理差异较大的代表性算法,能够衡量特征集合的效果. 为了得到更为准确的检测效果,本文采用 10 折交叉验证方法来计算评估指标.

4.1　实验环境与数据样本

本文在四川大学某实验室局域网出口搭建了木马流量检测系统,测试局域网共有主机 35 台,其中 30 台为正常使用机器,用于生成正常流量,5 台为目标机器用于生成木马流量,在局域网外设置一

台控制主机，用于控制木马，通过设置端口白名单的方式来保证流量的纯净，网络拓扑如图 4 所示. 实验收集了恶意软件社区（VirusShare、Github、MalShare）中上传的木马样本，选择后带有控制端的可用木马共 42 款.

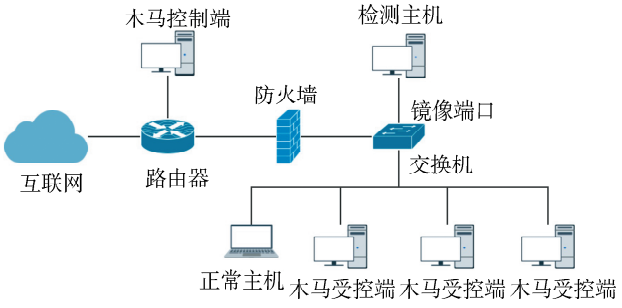

图 4　木马流量检测系统网络环境

Fig. 4　Network environment of Trojan detection system

在持续一周的流量采集中，共捕捉到正常流量 32 GB、木马流量 5 GB，在经过流量清洗后，共得到正常会话流 26 778 条，木马流量 4 261 条.

4.2　实验评估指标

取木马流量为 Positive，正常流量为 Negative. 本文使用精确率、召回率和 F_1 值三个指标来评价检测效果，定义如下.

精确率：$\mathrm{Prec}=TP/(FP+TP)$　　(8)

召回率：$\mathrm{Recall}=TP/(TP+FN)$　　(9)

F_1 值：$F_1=\dfrac{2\times\mathrm{Prec}\times\mathrm{Recall}}{\mathrm{Prec}+\mathrm{Recall}}$　　(10)

4.3　实验及结果分析

用于对比的特征选择算法为快速相关性过滤[17]（FCBF）和信息增益法（IG），均为基于信息熵的特征选择方法. 其中 IG 算法以特征的信息增益为指标，计算各个特征的信息增益并作排序，移除信息增益较低的特征，为了更准确地比较，其移除的数量设置与 AFSA 相同. FCBF 算法步骤如算法 2 所示.

算法 2　快速相关性过滤算法－FCBF

输入　特征集合 F，数据集，阈值 T，类别 C.

输出　特征子集 S.

1）遍历 F，计算特征 f_i 与类别的标准化互信息 SU_{ic}；

2）保留 SU_{ic} 大于阈值 T 的特征并排序；

3）以剩余特征中 SU_i 值最大者为主特征，计算其他特征 f_j 与它的标准化互信息 SU_{ij}；

4）将 SU_{ij} 与 f_j 的 SU_{jc} 值比较，若大于 SU_{jc} 则移除特征 f_j；

5）回到步骤 3），在剩余特征中继续选择主特征，直到剩余特征数为 1，输出子集.

实验后各算法移除的特征如表 6 所示.

表 6　三种特征选择算法移除的特征

Tab. 6　The list of removed features by three kinds of algorithm

AFSA	FCBF	IG
包最大长度	包最大长度	每秒流的数据包数
数据包最大长度	数据包最大长度	总下行包数量
总上行包长度	总上行包长度	总上行包长度
下行包长度标准差	下行包长度标准差	下行包长度标准差
流上行包最小间隔	发送包间隔均值	数据包平均长度
每秒流的数据包数	发送包最大间隔	发送包最大间隔
总下行包数量	——	上行包间隔总和
上行包最大长度	——	下行包间隔总和

特征选择有效性验证结果见表 7 和表 8，相对于文献[7]的特征集，本文初选特征集使用朴素贝叶斯分类时的精确率和召回率提升分别为 0.31%、12.24%，使用 SVM 时的提升分别为 0.55%、5.2%. 通过本文特征选择算法得到的终选特征集，使用朴素贝叶斯分类时的精确率提升为 0.88%，召回率提升为 2.12%，使用 SVM 时的精确率、召回率提升分别为 1.25%、1.4%.

表 7　朴素贝叶斯分类时特征选择有效性验证结果

Tab. 7　Validation result of feature selection using Naïve Bayes Classification algorithm

特征集合	精确率/%	召回率/%	F_1
文献[7]特征	90.15	82.56	86.19
初选特征集	90.46	94.80	92.58
终选特征集	91.34	96.92	94.05

表 8　支持向量机分类时特征选择有效性验证结果

Tab. 8　Validation result of feature selection using SVM Classification algorithm

特征集合	精确率/%	召回率/%	F_1
文献[7]特征	97.23	89.24	93.06
初选特征集	97.78	94.44	96.08
终选特征集	99.03	95.84	97.41

总体在召回率上的提升高于精确率上的提升，

由于漏报的危害性大,即召回率的提升更为重要,在 SVM 分类算法上召回率提升总体小于在朴素贝叶斯分类上的提升,原因是本文特征选择中一部分影响是联合相关性系数带来的,而朴素贝叶斯对特征独立的强假设使得本方法带来的增益更高.

本文特征选择算法优异性验证结果如表 9 和表 10 所示,在召回率上本文终选特征集较优,使用朴素贝叶斯分类时达到了最高 96.92%,FCBF 所选的特征集合在精确率上稍高于 AFSA 算法,但其 F_1 值仍然低于本文终选特征集;且 SVM 分类时本文终选特征集达到最高 99.03%的精确率,简单使用信息增益的 IG 算法得到的特征子集检测效果最差,召回率与文献[7]特征集的结果接近.

表 9　朴素贝叶斯分类时本文算法优异性验证结果

Tab. 9　Validation result of our method's improvement using Naïve Bayes Classification algorithm

特征选择算法	精确率/%	召回率/%	F_1
AFSA	91.34	96.92	94.05
FCBF	92.87	93.16	93.01
IG	90.52	90.46	90.49

表 10　支持向量机分类时本文算法优异性验证结果

Tab. 10　Validation result of our method's improvement using SVM Classification algorithm

特征选择算法	精确率/%	召回率/%	F_1
AFSA	99.03	95.84	97.41
FCBF	98.78	94.20	96.44
IG	97.84	90.88	94.23

结合以上实验数据,对表 6 特征选择结果进一步分析. FCBF 算法移除的特征与本文 AFSA 算法移除的特征有一定的重合,在对算法的每一轮计算结果进行对比后发现,其未移除的特征中每秒流的数据包数、总下行包数量均被划分到保留特征中,即它们与类别的标准化互信息值较大,但在 AFSA 移除过程中,这两个特征分别在第 2 轮、第 5 轮被移除,它们的重要性评价系数差别并不明显,但联合相关性评价系数均较大,正是它们给总体特征集合带来较大冗余而被移除. 同时 FCBF 未做特征重要性度量,被其移除的发送包最大间隔和发送包间隔均值两个特征在 AFSA 中属于重要性评价系数较高而保留的特征. 该算法需要设置阈值也给特征选择带来更多的工作和不确定性.

IG 算法仅考虑单一特征与类别的相关程度,忽略了特征间的相关性,其移除的特征与前两者差异较大,其中部分特征的重要性评价系数较大,如发送包最大间隔、数据包平均长度,另外总上行包长度、下行包间隔总和两个特征在 AFSA 中计算的联合相关性评价系数较小,但在 IG 中表现为与类别关联较弱而移除,最终造成较差的实验结果. 本文算法通过重要性评价系数预先划分一次特征,接着通过每一轮迭代计算联合相关性评价系数来综合评价特征,充分考虑了特征与整体集合的相关性,得到更优的特征子集.

同时绘制出三种特征选择算法得到特征集合使用朴素贝叶斯分类结果的实验接收者操作特征曲线(Receiver Operating Characteristic, ROC),如图 5 所示.

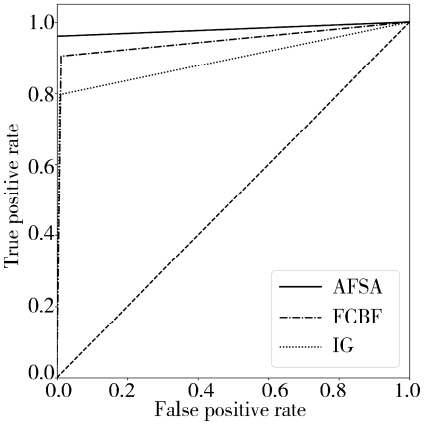

图 5　朴素贝叶斯分类下 ROC 曲线

Fig. 5　ROC of model's using NB classification

计算三条 ROC 曲线的 AUC(Area Under Curve)值,见表 11.

表 11　AUC 值

Tab. 11　AUC value

特征选择算法	AFSA	FCBF	IG
AUC 值	0.980	0.946	0.890

图 5 及表 11 也证明本文特征选择算法相比对照算法的优异性.

5　结　论

现有基于通信流量的木马检测方法中存在所用特征的代表性不足、特征间信息冗余的问题,本文通过流量分析在一定规模的真实数据上充分提取木马会话特征,通过定义改进的特征重要性评价系数和联合相关性评价系数,基于此设计一种特征

子集自适应选择算法(AFSA). 实验结果表明，本文算法选择后特征集合能有效提升木马检测效果. 后续研究将集中于检测模型的选择与实时环境下系统的构建.

参考文献：

[1] 国家计算机网络应急技术处理协调中心. 2018 年中国互联网网络安全报告[EB/OL]. (2019-07-17)[2020-02-13]. https://www.cert.org.cn/publish/main/upload/File/2018annual.pdf.

[2] Qin J, Yan H J, Si Q, *et al*. A trojan horse detection technology based on behavior analysis[C]// Proceedings of the 6th International Conference on Wireless Communications Networking and Mobile Computing. Piscataway, NJ: IEEE, 2010.

[3] 肖锦琦，王俊峰. 基于模糊哈希特征表示的恶意软件聚类方法[J]. 四川大学学报：自然科学版，2018，55：469.

[4] 李巍，李丽辉，李佳，等. 远控型木马通信三阶段流量行为特征分析[J]. 信息网络安全，2015，15：10.

[5] Jiang D, Omote K. An approach to detect remote access Trojan in the early stage of communication [C]// Proceedings of the IEEE International Conference on Advanced Information Networking & Applications. Piscataway, NJ: IEEE, 2015.

[6] 胥攀，刘胜利，兰景宏，等. 基于多数据流分析的木马检测方法[J]. 计算机应用研究，2015，32：890.

[7] 兰景宏，刘胜利，吴双，等. 用于木马流量检测的集成分类模型[J]. 西安交通大学学报，2015，49：84.

[8] 张兆林，武东英，罗友强，等. 基于 Adaboost 算法的窃密木马检测模型研究[J]. 信息工程大学学报，2015，49：84.

[9] 汪洁，杨力立，杨珉. 基于集成分类器的恶意网络流量检测[J]. 通信学报，2018，39：155.

[10] 刘敬，谷利泽，钮心忻，等. 基于单分类支持向量机和主动学习的网络异常检测研究[J]. 通信学报，2012，36：136.

[11] 程光，陈玉祥. 基于支持向量机的加密流量识别方法[J]. 东南大学学报：自然科学版，2017，47：655.

[12] 申健，夏靖波，张晓燕，等. 基于分治排序策略的流量二次特征选择[J]. 电子学报，2017，45：128.

[13] 董红斌，滕旭阳，杨雪. 一种基于关联信息熵度量的特征选择方法[J]. 计算机研究与发展，2016，53：1684.

[14] Zheng K F, Wang X J. Feature selection method with joint maximal information entropy between features and class[J]. Pattern Recognition, 2018, 77: 20.

[15] 王颖，曹捷，邱志洋. 基于乌鸦搜索算法的新型特征选择算法[J]. 吉林大学学报：理学版，2019，57：869.

[16] 王华华，黄龙，周远文，等. 改进的 mRmR 特征选择方法在人体行为识别中的应用[J]. 重庆邮电大学学报：自然科学版，2019，31：261.

[17] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy[J]. J Mach Learn Res, 2004, 5: 1205.

[18] Zhang Y D, Wang S H, Phillips P, *et al*. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection[J]. Knowl-Based Syst, 2014, 64: 22.

[19] Diao R, Chao F, Peng T, *et al*. Feature selection inspired classifier ensemble reduction[J]. IEEE T Cybern, 2017, 44: 1259.

[20] Wang Q, Shen Y, Zhang Y, *et al*. Fast quantitative correlation analysis and information deviation analysis for evaluating the performances of image fusion techniques[J]. IEEE T Instru Meas, 2004, 53: 1441.

[21] Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE T Pattern Anal, 2005, 27: 1226.

引用本文格式：

中　文：张瑜，刘晓洁，李贝贝. 一种针对木马流量的特征选择方法[J]. 四川大学学报：自然科学版，2021，58：012004.

英　文：Zhang Y, Liu X J, Li B B. A feature selection method for remote access trojan's traffic [J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 012004.