计算机科学
COMPUTER SCIENCE

基于启发式搜索特征选择的加密流量恶意行为检测技术

俞赛赛, 王小娟, 章倩倩

引用本文

俞赛赛, 王小娟, 章倩倩.基于启发式搜索特征选择的加密流量恶意行为检测技术[J]. 计算机科学, 2022, 49(11A): 210800237-6.

YU Sai-sai, WANG Xiao-juan, ZHANG Qian-qian. Detection of Malicious Behavior in Encrypted Traffic Based on Heuristic Search Feature Selection [J]. Computer Science, 2022, 49(11A): 210800237-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800237

基于启发式搜索特征选择的加密流量恶意行为检测技术

俞赛赛[1]　王小娟[2]　章倩倩[3]
1 中国电子科技集团共识第三十研究所　成都 610096
2 北京邮电大学电子工程学院　北京 100089
3 海军士官学校图书馆　安徽 蚌埠 233040
(734641272@qq.com)

摘　要　随着加密流量在网络中的占比越来越大,隐藏在加密流量中的恶意行为也越来越多,网络安全威胁形势越来越严峻。具有某些恶意行为的加密流量包含有多种流量特征,其特征之间本身也存在一定的冗余性。冗余的特征会增加检测时间,降低模型检测的效率。文中依据启发式搜索策略原理对加密流量包含的多种不同的特征进行筛选,找出具有代表性的特征组合。首先根据随机森林算法对特征重要度进行排序,筛选出对分类结果影响较大的特征,然后利用 Pearson 相关系数计算所有特征之间的相似度,筛选出彼此之间较为独立的特征组合。在数据集 CTU-13 上的实验结果表明,通过筛选出具有代表性的特征组合,在不降低检测准确率的情况下,减少了检测时间,提高了对加密流量恶意行为的检测效率。

关键词:加密流量;恶意行为;启发式搜索策略;特征选择
中图法分类号　TP309

Detection of Malicious Behavior in Encrypted Traffic Based on Heuristic Search Feature Selection

YU Sai-sai[1],WANG Xiao-juan[2] and ZHANG Qian-qian[3]
1 Consensus 30 Research Institute of China Electronics Technology Group,Chengdu 610096,China
2 School of Electronic Engineering,Beijing University of Posts and Telecommunications,Beijing 100089,China
3 Naval Academy Library,Bengbu,Anhui 233040,China

Abstract　With the proportion of encrypted traffic in the network increasing,there are more and more malicious behaviors hidden in the encrypted traffic,which makes the situation of network security more and more serious. Encrypted traffic with some malicious behavior contains a variety of traffic characteristics,among which there is some redundancy. Redundant features will increase the detection time and reduce the efficiency of model detection. Based on the principle of heuristic search strategy,this paper selects many different features of encrypted traffic and finds out the representative combination of features. Firstly,the feature importance is sorted according to the random forest algorithm,and the features that have a great impact on the classification results are selected. Then,the similarity between all features is calculated by Pearson correlation coefficient,and the relatively independent feature combinations are selected. Experimental results on the data set CTU-13 show that,by screening representative feature combinations,detection time is reduced and the detection efficiency of encrypted traffic malicious behavior can be improved without decreasing the detection accuracy.

Keywords　Encrypted traffic,Malicious behavior,Heuristic search strategy,Feature selection

随着互联网应用越来越深入人们的生活,我国网民数量呈递增趋势快速增长。互联网的高速发展,不仅给我们的生活带来了便利,而且还给许多技术创新领域提供了诸多机会。作为互联网中信息传输、交互的有效载体,网络流量包含大量有价值的信息数据。为了保障信息数据能够安全稳定地在网络中传输,大部分网络流量数据在传输过程中都采用了加密技术。为了获取非法利益,不法分子将恶意流量也采用加密手段混入到网络流量中,以防止被安全人员检测出来。因此,互联网走向全面加密的时代已经成为一种趋势,如何通过加密流量的特征进一步分析检测出其是否具有恶意行为,成为了当前网络安全研究的热点话题[1]。

一般情况下,对于未加密的流量而言,主要采用传统的流量检测技术(如深度报文检测技术等)检测其是否具有恶意性。但随着 TLS,SSL 加密协议被广泛应用,加密后的流量数据内容已无法被检测工具准确识别。基于机器学习的加密流量检测方法主要通过提取具有恶意行为的加密流量的某些流量特征,通过构建机器学习算法检测模型,不断对模型进行迭代、优化,可以实现利用流量特征准确识别、判断出隐藏在正常网络流量中具有恶意行为的加密流量[2]。

本文主要针对加密流量恶意行为特征进行研究分析,通过采用启发式搜索策略,对已有的流量特征进行筛选,找到一组具有代表性的特征组合。在不降低检测准确率的前提下,降低模型检测所耗费的时间,提高模型的检测效率。

1　相关研究

通过对加密流量进行分析可以对其中的恶意行为进行

通信作者:王小娟(wj2718@163.com)

检测。目前已经有很多针对网络中的加密流量的分类与分析的研究，包括端到端的加密流量分类[3]以及实时分类[4]等。这些研究往往用到机器学习[5]、深度学习[6]等方法。在对加密流量进行分析后要对其中的恶意行为进行检测同样也有很多方法，包括基于逆向工程[7]、审计日志[8]、接口响应[9]等。但这些方法往往直接使用数据的所有特征，或者进行一定程度的数据预处理，没有考虑到特征的重要度以及特征间的相似度，导致特征冗余，效率降低。为了应对这一挑战，我们考虑用启发式搜索特征选择来进行加密流量恶意行为检测。

启发式搜索又称为信息搜索，主要利用问题已拥有的启发信息引导搜索的方向，通过不断减小搜索范围，逐步实现降低待解决问题的复杂度，从而达到获取相对最优解决方案的目的。虽然采用启发式搜索策略可以降低待解决问题的复杂性，但是也存在着一定的弊端。首先，启发式搜索策略主要依据以往经验知识进行搜索判断，具有一定的局限性；其次，虽然引入的启发式信息越强，就越有可能大大降低搜索的工作量，但仍然无法保证能够获取到最佳的解决方案[10-12]。

针对加密流量恶意行为特征之间存在的冗余性，本文将利用启发式搜索策略对其进行搜索判断。首先，对所有特征进行特征重要度评估，按照计算后得到的特征重要度值的大小进行排序，筛选出对分类结果影响较大的特征；其次，利用Pearson相关系数计算所有特征之间的相似度，筛选出彼此之间较为独立的特征组合。

1.1 特征重要度

特征重要度是用来衡量特征在分类过程中贡献作用大小的一个指标。特征重要度值越大，说明该特征在分类过程中做出的贡献就越大，越具有代表性。本文将采用随机森林算法对特征的重要度进行评估，下面将对特征重要度评估原理过程进行详细介绍。

(1)根据平均不纯度的减少情况进行评估

这种评估方法主要使用基尼指数，具体计算过程如下：

假设有 m 个特征 $X_1, X_2, X_3, \cdots, X_m$，用 VIM 表示特征重要度，用 GI 表示基尼指数 Gini。首先，计算第 j 个特征在随机森林所有决策树中节点分裂不纯度的平均改变量，即为每个特征 X_j 的基尼指数值 Gini。Gini 的具体计算式如下：

$$GI_m = \sum_{k=1}^{|k|} \sum_{k' \neq k} p_{mk} p_{mk'} \tag{1}$$

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。具体含义为：随机从节点 m 中抽取两个样本，计算其类别标记不一致的概率。特征 X_j 在节点 m 处的重要程度，即为节点 m 分枝前后的 Gini 指数变化量。具体计算式如下：

$$VIM_{jm}^{\text{Gini}} = GI_m - GI_l - GI_r \tag{2}$$

其中，GI_l 和 GI_r 分别表示节点 m 分枝后两个新节点的 $Gini$ 指数。假设特征 X_j 在决策树 i 中出现的节点在集合 M 中，那么 X_j 在第 i 棵树的重要度为：

$$VIM_{ij}^{\text{Gini}} = \sum_{m \in M} VIM_{jm}^{\text{Gini}} \tag{3}$$

假设随机森林中共有 n 棵树，那么：

$$VIM_j^{\text{Gini}} = \sum_{i=1}^{n} VIM_{ij}^{\text{Gini}} \tag{4}$$

最后，把所求得的重要度进行归一化处理，即：

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{n} VIM_i} \tag{5}$$

(2)根据平均准确率的减少进行计算

这种计算方法首先对所有特征进行加噪处理，记录处理前后检测准确率的变化情况，通过判断其变化大小来确定特征的重要程度。如果变化较大，说明该特征重要程度较大；如果变化较小，说明该特征重要程度较小。整个过程采用带外数据错误率作为评价指标进行衡量[13]。具体研究分析过程如下：

首先利用带外数据计算随机森林中每一棵决策树的带外误差，将其标记为 X_1；然后对带外数据的某个特征随机进行加噪处理，即随机改变该特征的值，再次计算每棵决策树的带外误差，将其标记为 X_2。如果随机森林中共有 N 棵决策树，那么该特征的重要程度即为 $\Sigma(X_2 - X_1)/N$，根据对该特征进行加噪处理前后的带外误差变化情况，判断该特征对分类结果的影响程度，即为该特征的重要程度。

1.2 Pearson 相关系数

Pearson 相关系数是由皮尔逊依据 19 世纪 80 年代的弗朗西斯·高尔顿所提出的“相似而又稍有不同”的想法演变而来。Pearson 相关系数可以用来衡量两个变量之间的相关性大小，一般情况下，其取值范围会介于 -1 到 1 之间。Pearson 相关系数的绝对值越接近于 1，这两个变量的相关性就越强；Pearson 相关系数的绝对值越接近于 0，这两个变量的相关性就越弱[14]。具体的研究分析过程如下：

假设存在两个变量，分别为 X 和 Y，那么这两个变量之间的相关系数可以根据以下公式计算：

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\Sigma((X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} \tag{6}$$

其中，cov 代表方差，E 代表数学期望。

假设存在两个数值属性，分别为 A 和 B，如果要计算两个属性之间的相互关系数，来估计两个数值属性的相关度 r，那么可以根据以下公式进行计算。

$$r = \frac{\sum_{i=1}^{n} (a_i - A)(b_i - B)}{n \sigma_A \sigma_B} \tag{7}$$

其中，n 代表元组的个数，a_i 和 b_i 分别代表元组 i 在 A 和 B 上的值，$\bar{A}$ 和 $\bar{B}$ 分别代表 A 和 B 的均值，σ_A 和 σ_B 分别代表 A 和 B 的标准差。

如果数值属性 A 和 B 之间的相关性 r 大于 0，说明 A 和 B 是正相关的，那么 A 的值将会随着 B 的值增大而增大。r 值越大，说明 A 和 B 之间的相关性越强，那么其中的一个数值属性蕴含着另一个数值属性的可能性就越大；如果数值属性 A 和 B 之间的相关性 r 等于 0，说明 A 和 B 是相互独立的，它们之间不存在相关性；如果数值属性 A 和 B 之间的相关性 r 小于 0，说明 A 和 B 是负相关的，那么 A 的值将会随着 B 的值减小而增大。r 值越小，说明 A 和 B 之间的相关性越强，那么其中的一个数值属性蕴含着另一个数值属性的可能性就越大。因此，如果 r 值的绝对值越高，那么数值属性 A 或 B 就可以作为冗余属性而被删除掉[15-16]。

2 算法设计

2.1 流量特征

在网络流量传输过程中，数据包的大小以及传输速率等都会对流量提取工作产生一定的影响。本文将主要从时间和空间两个方面，通过分析、研究网络流量数据包在传输过程中

已有的部分特点，总结归纳出加密流量所具有的部分特征，具体从数据包本身包含的四元组数量、传输的持续时间及相应标准差、发送字节数与接收字节数的数量及周期之比等方面进行整理[17]。流量特征的具体内容如表1所列。

表1　已提取的12种特征情况

Table 1　12 extracted characteristic conditions

序号	特征描述
$f1$	数据包内平均四元组数量
$f2$	平均持续时间
$f3$	持续时间下的标准差
$f4$	持续时间下的标准偏差
$f5$	发送的字节数量
$f6$	接收的字节数量
$f7$	发送的字节数量与接受的字节数量之比
$f8$	传输中已建立的连接数与总连接数之比
$f9$	传入的报文数量
$f10$	发出的报文数量
$f11$	连接周期的平均值
$f12$	连接周期的标准差

2.2　特征选择的启发式搜索策略

在机器学习算法模型检测过程中，特征的选择尤为重要。特征重要度越大，分类检测的结果越准确，反之误报率越大；特征数量越多，对于提高分类检测结果的准确率越有利。然而特征之间也存在一定的冗余与耦合，具有相似作用的特征数量越多，反而会对模型检测整体的准确率产生一定的影响，同时也会增加检测时间，降低模型检测的效率。因此，本文将主要以寻找最具有代表性的特征组合为目的，在不降低模型检测整体准确率的前提下，找到能够满足该条件的特征组合，以降低模型检测所耗费的时间，提高模型的检测效率[1,18]。

特征选择的启发式搜索策略主要是在已有流量特征的基础上，利用启发式搜索策略，完成对流量特征的筛选、重组等，最终得到一组相对最优的特征组合。特征选择的启发式搜索策略具体可以分为3个步骤：

首先，利用随机森林算法训练所有的特征数据，计算每种特征的重要度，完成第一次特征筛选过程；

其次，利用Pearson相关系数理论计算特征之间的相关性大小，根据得到的相关性值判断、筛选特征之间关联性最小的特征组合；

最后，将两次筛选出的结果进行整理，结合使用启发式搜索策略，对特征进行重新组合，选取特征重要度值较大、彼此之间相关性较小的特征，并依据该启发式搜索策略完成对所有特征的重新排序。具体实施流程如图1所示。

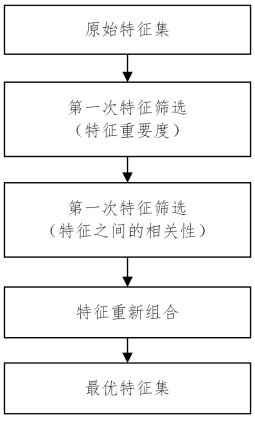


图1　启发式搜索策略流程图

Fig. 1　Flow chart of heuristic search strategy

2.2.1　特征重要度评估

特征重要度评估主要计算所有特征的重要度值，即计算所有特征在分类过程中所做出的贡献大小。然后依据每个特征所做出的贡献值由大到小对所有特征进行排名，完成第一次特征筛选。本文主要使用随机森林算法对所有特征进行重要度评估，特征重要度评估流程图如图2所示。

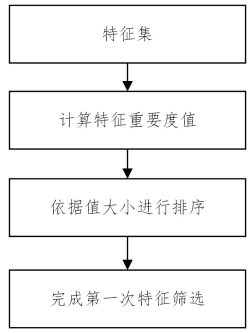


图2　特征重要度评估流程图

Fig. 2　Flow chart of feature importance evaluation

2.2.2　Pearson相关性系数评估

Pearson相关性系数评估主要计算所有特征之间的相关性大小值，即用于反映特征之间耦合性大小。依据计算得到特征之间的相关性大小值，筛选出彼此之间耦合性较小的特征组合，完成第二次特征筛选[19-22]。本文主要使用Pearson相关系数理论对所有特征进行相似度计算，Pearson相关性系数评估流程图如图3所示。

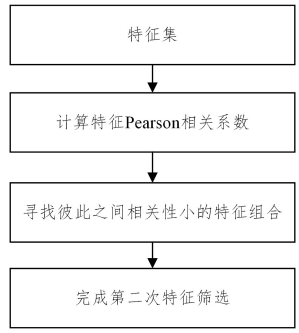


图3　特征重要度评估流程图

Fig. 3　Flow chart of feature importance evaluation

2.3　加密流量恶意行为检测

随机森林算法是一种被广泛应用的传统机器学习算法，目前已经被广泛应用于多个不同领域。随机森林算法属于集成学习内容，是将多棵决策树聚集形成的一种集成算法。随机森林算法既可以用于回归问题，又可以用于解决分类问题，是一种被高频率使用的机器学习算法。

随机森林算法属于监督学习算法，主要利用其内部多棵决策树完成对输入数据的训练，通过对所有决策树预测结果进行分析、整理，将投票数最多的类别作为最终输出结果。随机森林算法实现原理简单，在解决二分类问题上具有高精度的优势。随机森林算法分类器构建过程示意图如图4所示，在对原始样本数据进行训练时，如果能够构建足够多的决策树分别对样本数据进行检测，那么就会在一定程度上降低随机森林算法分类器出现过拟合的风险，提高分类器的检测精度[23-26]。

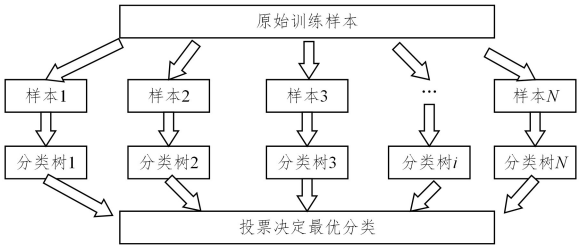


图 4 随机森林构建过程示意图

Fig. 4 Schematic diagram of random forest construction process

针对加密流量中隐藏的多种具有不同特点的恶意行为，依据每种恶意行为所具有的不同特点，本文将构建随机森林算法模型，将提取到的加密流量恶意行为相关的特征向量值输入到检测模型中进行检测，最终依据得到的准确率、误报率、F1-score、运行时间等技术指标对分类结果进行判断。随机森林算法模型构建流程图如图 5 所示。

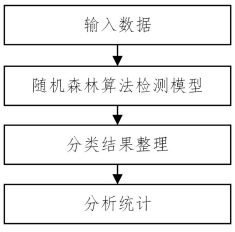


图 5 随机森林算法模型构建流程图

Fig. 5 Flow chart of random forest algorithm model construction

3 实验

3.1 实验数据

实验数据来自捷克技术大学提供的网络开源数据集 CTU-13，该流量数据集包含了正常加密流量数据和恶意加密流量数据。其中，正常加密流量数据是在没有执行恶意软件的正常主机上捕获的，而恶意加密流量数据是在 Windows 虚拟机中执行不同种类的恶意软件时捕获的，主要以僵尸网络为主，可以再具体划分为 13 种不同类型的恶意加密流量数据。

本文具体用到的数据集是将原始 CTU-13 流量数据集进行预处理之后，生成的以四元组作为唯一标识的特征向量集。该特征向量集包含了 12 种流量特征，分为正常加密流量数据、恶意加密流量数据两种类型[27]。数据集内包含的四元组数量具体情况如表 2 所列。

表 2 启发式特征选择方案实验数据

Table 2 Experimental data of heuristic feature selection scheme

类别	数量
四元组数量	24166
具有恶意行为的加密流量四元组总数	10538
正常加密流量四元组总数	13628

3.2 特征重要度评估

特征重要度评估实验主要利用随机森林算法对每个特征的重要度进行计算，根据计算结果由大到小进行排序。特征重要度评估结果如表 3 所列。

表 3 特征重要度评估结果

Table 3 Evaluation results of feature importance

序号	特征编号	特征重要度	序号	特征编号	特征重要度
1	*f*7	0.1867230	7	*f*3	0.0345650
2	*f*10	0.0649120	8	*f*1	0.0285750
3	*f*2	0.0578830	9	*f*6	0.0216230
4	*f*9	0.0453178	10	*f*11	0.0114720
5	*f*8	0.0406320	11	*f*12	0.0083750
6	*f*5	0.035602	12	*f*4	0.0067370

根据特征重要度评估产生的结果可知，特征重要度值越大，说明该特征在分类过程中起到的作用就越大，对分类结果的影响也越大；特征重要度值越小，说明该特征在分类过程中起到的作用越小，对分类结果的影响也越小[28-29]。

3.3 Pearson 相关系数评估

具有恶意行为的加密流量特征之间，可能会存在一定的关联性与耦合性。本文利用 Pearson 相关系数原理，通过对每个流量特征之间进行相似度计算，结果如表 4 所列。找出特征之间关联性小、彼此独立的特征组合，完成第二次特征筛选。根据特征相似度计算结果可知，特征之间的相关系数值决定了特征之间的相关性大小。一般情况下，如果相关系数的绝对值越接近于 1，那么这两个特征的相关性就越强；如果相关系数的绝对值越接近于 0，那么这两个特征的相关性就越弱。通过特征相似度计算的结果可以判断所有特征之间的相关性大小，找出彼此之间独立的特征组合。

表 4 特征相似度计算结果

Table 4 Calculation results of feature similarity

	*f*1	*f*2	*f*3	*f*4	*f*5	*f*6	*f*7	*f*8	*f*9	*f*10	*f*11	*f*12
*f*1	1.0000	0.5100	0.5000	0.0960	0.8000	0.6800	0.2800	0.0990	0.8000	0.8300	0.3000	0.2900
*f*2	0.5100	1.0000	0.5600	0.0110	0.5200	0.4900	0.2400	0.0920	0.5700	0.5600	0.0990	0.0730
*f*3	0.5000	0.5600	1.0000	0.1500	0.5000	0.4800	0.1500	0.0058	0.5400	0.5300	0.2500	0.2200
*f*4	0.0960	0.0110	0.1500	1.0000	0.1300	0.1000	0.0400	0.0200	0.0800	0.0900	0.5700	0.5400
*f*5	0.8000	0.5200	0.5000	0.1300	1.0000	0.6100	0.2700	0.1200	0.7800	0.8500	0.2900	0.2600
*f*6	0.6800	0.4900	0.4800	0.1000	0.6100	1.0000	0.3000	0.0310	0.8500	0.7900	0.1900	0.1600
*f*7	0.2800	0.2400	0.1500	0.0470	0.2700	0.3000	1.0000	0.9000	0.2900	0.3000	0.0710	0.0580
*f*8	0.0990	0.0920	0.0100	0.0200	0.1200	0.0300	0.9000	1.0000	0.0700	0.0900	0.0130	0.0160
*f*9	0.8000	0.5700	0.5400	0.0810	0.7800	0.8500	0.2900	0.0730	1.0000	0.9700	0.2600	0.2300
*f*10	0.8300	0.5600	0.5300	0.0880	0.8500	0.7900	0.3000	0.0960	0.9700	1.0000	0.2700	0.2400
*f*11	0.3000	0.0990	0.2500	0.5700	0.2900	0.1900	0.0700	0.0130	0.2600	0.2700	1.0000	0.9700
*f*12	0.2900	0.0730	0.2200	0.5400	0.2600	0.1600	0.0500	0.0160	0.2300	0.2400	0.9700	1.0000

3.4 特征组合筛选实验

在利用机器学习模型对恶意加密流量进行检测时，选择何种特征组合对分类检测结果有着重要的影响。本实验围绕已提取到的特征集，对每个特征展开重要度评估，然后计算所有特征之间的相关系数，结合启发式搜索策略，完成特征集的降序排列。最后按照排列顺序，分别选取前 1 个、前 2 个…前

12 个特征，生成 12 种不同的特征组合[30]。最终生成的 12 种不同的特征组合如表 5 所列。

表 5　12 种不同的特征组合

Table 5　12 different feature combinations

特征组合	包含的特征种类
1	*f*10
2	*f*10，*f*7
3	*f*10，*f*7，*f*9
4	*f*10，*f*7，*f*9，*f*3
5	*f*10，*f*7，*f*9，*f*3，*f*6
6	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12
7	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11
8	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11，*f*2
9	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11，*f*2，*f*8
10	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11，*f*2，*f*8，*f*5
11	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11，*f*2，*f*8，*f*5，*f*1
12	*f*10，*f*7，*f*9，*f*3，*f*6，*f*12，*f*11，*f*2，*f*8，*f*5，*f*1，*f*4

根据特征预处理实验中生成的 12 种恶意流量特征组合可知，本实验的主要内容为：先根据不同的特征组合构建相应的特征向量集，利用随机森林算法构建分类检测模型；然后将这些特征向量输入到检测模型中，完成检测并记录实验结果。根据实验结果，判断启发式特征选择算法是否有效，并找出具有代表性的特征组合。所有的实验结果如表 6 所列。

表 6　特征组合筛选实验结果

Table 6　Experimental results of feature combination screening

类别	准确率/%	误报率/%	精确率/%	召回率/%	F1-score/%	运行时间/s
组合 1	93.327	34.573	95.785	65.426	78.476	10.697
组合 2	94.336	30.353	96.522	67.895	81.336	16.783
组合 3	95.735	28.562	96.736	71.437	82.717	16.963
组合 4	95.976	26.255	98.568	73.744	84.473	22.508
组合 5	97.477	11.394	93.588	89.368	92.533	23.694
组合 6	98.436	10.332	97.373	89.576	93.768	22.535
组合 7	98.658	8.154	96.426	91.897	94.563	19.283
组合 8	98.326	7.803	96.337	92.594	94.672	19.587
组合 9	98.973	4.967	97.556	95.533	96.730	25.675
组合 10	99.356	3.656	97.810	96.573	96.943	25.973
组合 11	99.073	5.326	96.646	94.672	95.658	26.753
组合 12	99.325	4.875	97.437	96.790	96.798	26.856

如表 6 所列，随着特征数量不断增加，检测模型的准确率也在逐渐提高。当特征数量达到 6 个时，准确率基本保持在一定范围内。但在特征数量较少时，检测模型的误报率也较高，当特征数量达到 10 个以上时，误报率处于较低水平。结合检测时间的变化趋势可以发现，当特征数量为 7 个或 8 个时，模型的检测效果较好。利用 F1-score 对这两种情况进行对比分析后，可以发现特征组合 8 更优于特征组合 7。

因此，通过对所有特征筛选实验的结果进行分析总结，归纳分析出以下结论：

(1)随着特征数量的不断增加，恶意加密流量检测的准确率也在不断地提升；

(2)当选取的特征数量达到一定值时，模型检测的准确率会在一定范围内波动；

(3)随着选取的特征数量不断增加，模型检测所耗费的时间也在不断地增加，最终会在一定范围内进行波动。

根据上述对具有恶意行为的加密流量特征进行筛选以及实验分析产生的结果可知，隐藏在加密流量中的恶意行为在通信过程中所产生的流量特征，足以用于检测识别出该种恶意行为，并且通过利用启发式搜索特征选择算法，可以找出一组具有代表性的特征组合，在不降低整体检测准确率的前提下，缩短模型检测的运行时间，有效提高对加密流量中恶意行为的检测效率。

结束语　本文主要通过分析、提取具有恶意行为的加密流量的一些流量特征，利用启发式搜索特征选择算法，并构建相关的检测模型，从所有特征中找出分类效果较好的、彼此之间冗余性较小的特征组合，提高加密流量恶意行为的检测效率。

综上所述，本文通过实验证明，使用基于启发式搜索特征选择的加密流量恶意行为检测技术对恶意加密流量特征筛选之后，能够找到具有代表性的特征组合 8，并且该特征组合数量要少于全部特征的数量，同时检测准确率能达到 98%以上，检测所耗费的时间缩短了近 25%左右。

在未来的工作中，如何在不降低检测准确率的情况下进一步缩短检测时间，提高检测效率是一个值得研究的方向。同时，寻找其他算法得到更准确的重要度排序以及相似度计算也是一个值得注意的点。并且，未来需在更多的数据集上进行实验以验证所提方法的通用性。

参 考 文 献

[1] Cisco. 2018 Annual Cybersecurity Report：The evolution of malware and rise of artificial intelligence[R/OL]. (2018-02)[2019-07-22]. https://newsroom. cisco. com/c/r/newsroom/en/us/a/y2018/m02/cisco-2018-annual-cybersecurity-report-reveals-se-curity-leaders-rely-on-and-invest-in-automation-machine-learning-and-artificial-intelligence-to-defen. html.

[2] ZHEN C Z. Research on encrypted traffic type identification based on DPI and machine learning[J]. Information Communication，2018，31(4)：258-260.

[3] WANG W，ZHU M，WANG J，et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks[C]//2017 IEEE International Conference on Intelligence and Security Informatics(ISI). IEEE，2017：43-48.

[4] BAR-YANAI R，LANGBERG M，PELEG D，et al. Realtime classification for encrypted traffic[C] // International Symposium on Experimental Algorithms. Berlin：Springer，2010：373-385.

[5] MSADEK N，SOUA R，ENGEL T. Iot device fingerprinting：Machine learning based encrypted traffic analysis[C] // 2019 IEEE Wireless Communications and Networking Conference (WCNC). IEEE，2019：1-8.

[6] REZAEI S，LIU X. Deep learning for encrypted traffic classification：An overview[J]. IEEE Communications Magazine，2019，57(5)：76-81.

[7] CHENG L Y，YONG S，ZHI X. Android malicious behavior detection method based on reverse engineering[J]. Information Security and Confidentiality of Communications，2015(4)：83-87.

[8] BERLIN K，SLATER D，SAXE J. Malicious behavior detection using windows audit logs [C] // Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. 2015：35-44.

[9] YANG M，WANG S，LING Z，et al. Detection of malicious behavior in android apps through API calls and permission uses analysis[J]. Concurrency and Computation：Practice and Experience，2017，29(19)：e4172. 1-e4172. 13.

[10] Aqniu. 一篇报告了解国内首个针对加密流量的检测引擎[EB/

OL]. (2019-3-15)[2019-7-22]. https://www.aqniu.com/tools tech/45207.html.

[11] BIN H, HONG Z Z, HONG Y L, et al. TLS malicious traffic detection based on the combined characteristics of message payload and flow fingerprint. [J/OL]. http://kns.cnki.net/kcms/detail/31.1289.TP.20191216.1035.003.html.

[12] LE T Y, MING H X, MIAO M. Analysis of the SSL protocol working process[J]. Cybersecurity skills Surgery and Application, 2017(7): 36-38.

[13] ANDERSON B, MCGREW D. Identifying encrypted malware traffic with contextual flow data[C] // Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. ACM, 2016: 35-46.

[14] FAN X Y. SSL/TLS protocol security research[D]. Nanjing: Southeast University, 2017.

[15] JING J, ZHI Z Y. Spark platform weighted hierarchical subspace randomized forest arithmetic research[J/OL]. [2022-02-25]. http://kns.cnki.net/kcms/detail/42.1671.TP.20191122.1607.022.html.

[16] WU Y L, KE Y T, CHEN Z, et al. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping[J]. Catena, 2022, 187: 104396.

[17] MALIK J, KAUSHAL R. CREDROID: Android malware detection by network traffic analysis[C] // Workshop on Privacy Aware Mobile Computing. New York: ACM, 2016.

[18] XU Y W. Research on HTTPS tunnel traffic detection technology based on fingerprint and statistical characteristics[D]. Xi'an: Xidian University, 2019.

[19] FENG D C, LIU Z T, WANG X D, et al. Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach[J]. Construction and Building Materials, 2022, 230: 117000.

[20] XUAN Z Z. Research on mobile traffic recognition and anomaly detection based on machine learning[D]Chengdu: University of Electronic Science and Technology of China, 2019.

[21] DREGER H, FELDMANN A. Dynamic application-layer protocol analysis for network intrusion detection[C] // Proceedings of the 15th USENIX Security Symposium. 2006.

[22] BASET S, SCHULZ RINNE H. An analysis of the Skype peer-to-peer internet telephony protocol[C] // 25th IEEE International Conference on Computer Communications, ser(INFOCOM2006). IEEE, 2006.

[23] LONG M R. Research and Implementation of Unknown and Encrypted Traffic Recognition Based on Convolutional Neural Network[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[24] GOODFELLOWI, BENGIO Y, COURVILLE A. Deep learning [M]. Massachusetts: MIT Press, 2016.

[25] PAN W, QIAO C X. Encrypted traffic identification method based on stacked autoencoder[J]. Computer Engineering, 2018, 44(11): 140-147.

[26] VOLKAN S, OMER K, MERIH G. A Bayesian network model for prediction and analysis of possible forest fire causes[J]. Forest Ecology and Management, 2022, 457: 117723.

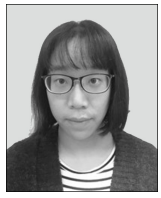
[27] JIE Q C, QIANG G. A feature selection method based on FG-Score[J]. Journal of Yibin University, 2018, 18(6): 4-8.

[28] SONG J G. Prediction of RNA spatial structure based on heuristic search strategy[D]. Tianjin: Tianjin Polytechnic University, 2019.

[29] KAI L. Research on adaptive feature selection and parameter optimization algorithm of stochastic forest[D]. Changchun: Changchun University of Technology, 2018.

[30] LI W X, GANG S, WEN Y X, et al. Correlation study of computer science and technology professional curriculum system based on Pearson coefficient[J]. Wireless Internet Technology, 2019, 16(21): 114-115.

YU Sai-sai, born in 1982, Ph.D, senior engineer. His main research interest includes cyber security and so on.



WANG Xiao-juan, Ph.D, associate professor. Her main research interests include cyber security, complex networks, deep learning and so on.