基于混合方法的 SSL VPN 加密流量识别研究

王 琳[1] 封化民[1,2] 刘 飚[2] 崔明辉[1] 赵 会[2] 孙曦音[1]

[1](西安电子科技大学 陕西 西安 710071)

[2](北京电子科技学院 北京 100070)

摘 要 SSL 是一种保证网络通信安全的协议,在流量传输中得到广泛使用。根据其应用的不同方式,可以分为普通的 SSL 加密流量和 SSL VPN 流量。许多不法分子常常将一些恶意流量隐藏在 SSL VPN 中进行传输。因此,SSL VPN 流量的识别对于网络监管来说十分重要。提出一种混合方法,将指纹识别与机器学习方法相结合,实现 SSL VPN 流量的识别。该方法基于时间相关的流特征,利用基于 GA(Genetic Algorithms)的改进 RF(Random Forest)算法,实现了 92.2% 的识别准确率。实验结果表明,该方法能有效识别出网络中的 SSL VPN 流量。

关键词 SSL VPN 恶意流量 流量识别 指纹识别 遗传算法 随机森林

中图分类号 TP393 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.02.055

SSL VPN ENCRYPTED TRAFFIC IDENTIFICATION BASED ON HYBRID METHOD

Wang Lin[1] Feng Huamin[1,2] Liu Biao[2] Cui Minghui[1] Zhao Hui[2] Sun Xiyin[1]

[1](*Xidian University*, *Xi'an 710071*, *Shaanxi*, *China*)

[2](*Beijing Electronic Science and Technology Institution*, *Beijing 100070*, *China*)

Abstract SSL is a protocol that guarantees the security of network communication. It is widely used in traffic transmission. According to its different ways of application, it can be divided into SSL encrypted traffic and SSL VPN traffic. Many criminals often hide some malicious traffic in SSL VPN for transmission. Therefore, the identification of SSL VPN traffic is very important for network supervision. We proposed a hybrid method which combined fingerprint recognition with machine learning method so as to realize the identification of SSL VPN traffic. On the basis of time-related traffic features, the method adopted the improved RF (Random Forest) algorithm based on GA (Genetic Algorithm) to achieve 92.2% recognition accuracy. The experimental result shows that the proposed method can effectively identify the SSL VPN traffic in the network.

Keywords SSL VPN Malicious traffic Traffic identification Fingerprint recognition Genetic algorithm Random forest

0 引 言

随着信息技术的飞速发展,互联网规模快速扩张,各种类型的网络服务不断增多,信息的安全性问题也受到越来越多的关注。为了保障数据传输的安全性,越来越多的流量都在加密后进行传输,这在保障信息安全的同时,也给流量的审计带来了挑战。

流量识别是网络监管的重要组成部分,也是近些年来的热门研究方向。自“棱镜”监控项目曝光以后,全球的加密网络流量急剧增加。安全套接层 SSL 协议在流量加密传输的过程中得到广泛使用,SSL 加密在打击非法信息窃取和黑客攻击,保护敏感数据等方面起着至关重要的作用。SSL VPN(基于安全套接层的虚拟专用网)也因其安全性和易用性在安全传输中得到广泛使用,但这也使得一些恶意流量有了可乘之机。

目前对于 SSL VPN 流量识别的研究还比较少,本文提出一种混合方法,通过两步来实现 SSL VPN 流量

收稿日期:2018 - 07 - 24。国家重点研发计划项目(2018YFB0803601)。**王琳**,硕士生,主研领域:机器学习,网络安全。**封化民**,教授。**刘飚**,讲师。**崔明辉**,硕士。**赵会**,硕士。**孙曦音**,硕士。

的识别。首先利用指纹识别的方法从网络流量中提取出 SSL 流,然后利用统计学习方法,根据时间相关的流特征,使用改进的 RF 算法,识别出其中的 SSL VPN 流量。

对于 SSL 流量的识别,目前常用的方法有基于机器学习的方法和基于指纹识别的方法,它们有其各自的优缺点。由于混合方法分两步实现,考虑到识别效率和准确率的因素,本文提出一种改进的"指纹识别"方法,通过匹配分组特定位置的字段信息来判断当前数据包是否使用 SSL 协议进行加密,进而判断当前流是否为 SSL 流。对于第二阶段的 SSL VPN 流量识别,本文提出两种基于遗传算法的改进 RF 算法,并与 RF 算法及其他机器学习方法进行了比较。实验结果表明,改进的 RF 算法有更好的识别效果。

1 相关研究

Paxson 等[1]在 20 世纪 90 年代初就开始了基于分组大小和基于流的流量分类研究,其中一些统计特征,如:分组长度、分组到达时间间隔和流持续时间等,被认为对协议识别有较好的效果。后来,文献[2]仅使用流量的前几个分组统计数据,提高了流量识别的效率。此外,为了提高大规模、高速网络的分类效率,文献[3]提出了基于签名的流量识别方法。此方法缩短了流量分类的时间,但无法检测出未知的或手动创建的签名。

对于 SSL 加密流量,文献[4]在 2010 年对 HTTPS 加密流量进行了研究,使用 SVM 算法来区分 WebMail 流量和其他 HTTPS 流量,但没有更进一步地细分其他类型的流量。

文献[5]在 2010 年提出一种混合框架,结合签名和统计的方法,来实现 SSL 流量的分类。首先使用签名匹配的方法来识别出 SSL 流量,然后使用统计的方法(朴素贝叶斯方法)来判断 SSL 流的具体应用协议。

目前的流量识别研究中,大多是对某种协议流量的识别,而对于 VPN 流量识别的研究很少。文献[6]在 2016 年首次提出仅利用时间相关的流特征识别 VPN 流量,比较了 KNN 和 C4.5 两种算法的识别效果,并公开了实验所使用的数据集。

西佛罗里达大学的 Bagui 等[6]在文献[7]发布的数据集的基础上做了进一步的研究,比较了 Logistic 回归、SVM、朴素贝叶斯、KNN、随机森林(RF)和梯度提升树(GBT)六种方法的识别效果,并对算法参数进行了适当优化,对 VPN 流量实现了 90% 以上的识别率。

2 SSL 加密流量识别

2.1 SSL 握手协议

SSL 位于传输层和应用层之间,是一种在主机之间提供安全通信的协议,保证了数据的可靠性传输[7]。SSL 协议由握手协议、记录协议、更改密文协议和警报协议组成,如图 1 所示。

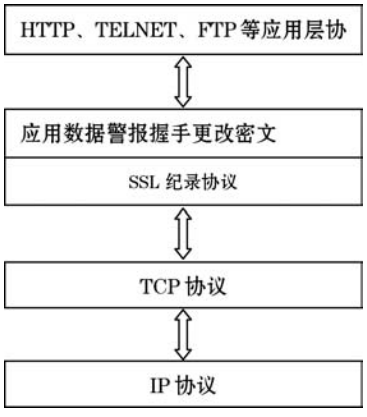

图 1 SSL 协议位置与组成

握手协议是 SSL 协议中十分重要的协议,通信双方通过握手来提供身份验证。客户端和服务器在此过程中确认所使用的密钥和算法,此外,还需协商双方的信息摘要算法、数据压缩算法等数据加密传输时需要用到的信息。握手协议完成后,双方开始数据的加密传输。握手协议的通信流程如图 2 所示。

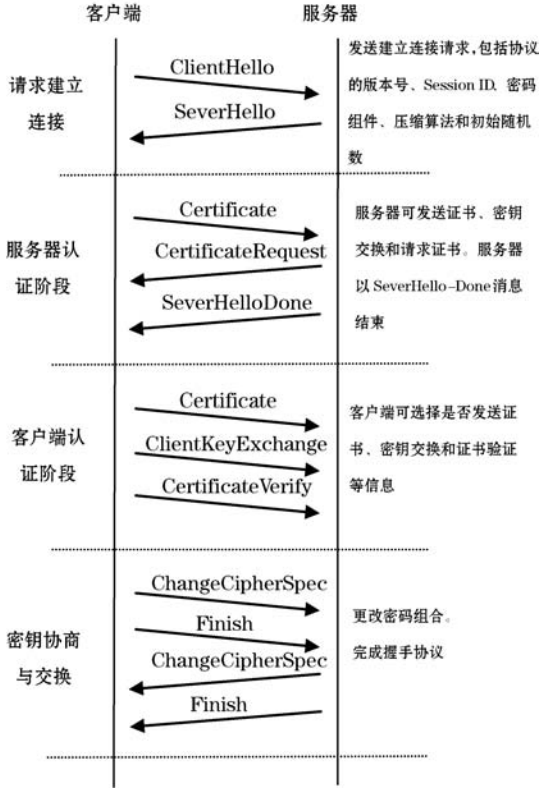

图 2 SSL 握手协议通信流程

(1) ClientHello 消息。客户端发送此消息到服务器,包括客户端所支持的算法和初始随机数。

(2) ServerHello 消息。服务器从可选的加密算法中选择一种作为通信使用的加密算法。

(3) ServerCertificate 消息。ServerCertificate 包含服务器标识和用于生成加密参数的随机数。

(4) CertificateRequest 消息。可选消息,当服务器需要验证客户端的身份时可发送此消息。

(5) ServerHelloDone 消息。这条消息在上面几条消息完成后发送,表明服务器在等候客户端的回应。客户端收到这条消息后便开始验证数字证书的合法性。

(6) ClientKeyExchange 消息。此消息负责传输客户端生成的预主密钥。

(7) ChangeCipherSpec 消息。通知服务器开始使用协商好的各项参数。

(8) Finish 消息。这是第一条安全处理过的消息,包含了之前所有握手消息的验证码。

2.2 SSL 握手协议报文结构

SSL 握手协议的各类消息都有其特定的报文结构特征,下面对其进行介绍,主要包括 ClientHello 报文、SeverHello 报文、Certificate 报文、SeverHelloDone 报文、ClientKeyExchange 报文、ChangeCipherSpec 报文。

2.2.1 ClientHello 报文结构

根据 RFC6101[9],ClientHello 报文数据结构如表 1 所示。

表 1　ClientHello 报文结构

HandShake	Version		Length1		Type	Length2		
0x16	0x03				0x01			

(1) ClientHello 消息的 TCP 数据段第 1 个字节为 0x16,表示当前数据包是 HandShake 消息报文,属于 SSL 握手协议的一部分。

(2) 第 2、3 两个字节表示 SSL 协议版本号,其中第 2 字节的值均为 0x03,表 2 为 SSL 各个版本号对应的字节表示。

表 2　SSL 协议版本

SSLv3		TLS1.0		TLS1.1		TLS1.2	
0x03	0x00	0x03	0x01	0x03	0x02	0x03	0x03

(3) Length1 占两个字节,表示剩余数据包的长度。

(4) 第 6 个字节为 HandShake Type,0x01 表示当前数据包为 ClientHello 消息。

(5) Length2 占三个字节,表示其后剩余数据包的长度。

2.2.2 SeverHello 报文结构

SeverHello 报文数据结构如表 3 所示。

表 3　SeverHello 报文结构

HandShake	Version		Length1		Type	Length2		
0x16	0x03				0x02			

(1) SeverHello 报文第一个字节为 0x16,表示当前数据包为握手协议的一部分。

(2) 与 ClientHello 对应的 Type 值不同,SeverHello 的第 6 个字节为 0x02,表示当前数据包为 SeverHello 消息。

2.2.3 Certificate 报文结构

Certificate 数据报文结构如表 4 所示。

表 4　Certificate 报文结构

HandShake	Version		Length1		Type	Length2		
0x16	0x03				0x0b			

(1) Certificate 报文第一个字节为 0x16,表示当前数据包为握手协议的一部分。

(2) Certificate 的第 6 个字节为 0x0b,表示当前数据包为 Certificate 消息。

2.2.4 SeverHelloDone 报文结构

SeverHelloDone 数据报文结构如表 5 所示。

表 5　SeverHelloDone 报文结构

HandShake	Version		Length1		Type	Length2		
0x16	0x03				0x0e			

(1) SeverHelloDone 报文第一个字节为 0x16,表示当前数据包为握手协议的一部分。

(2) ServerHelloDone 的第 6 个字节为 0x0e,表示当前数据包为 SeverHelloDone 消息。

2.2.5 ClientKeyExchange 报文结构

ClientKeyExchange 数据报文结构如表 6 所示。

表 6　ClientKeyExchange 报文结构

HandShake	Version		Length1		Type	Length2		
0x16	0x03				0x10			

(1) ClientKeyExchange 报文第一个字节为 0x16,表示当前数据包为握手协议的一部分。

(2) ClientKeyExchange 的第 6 个字节为 0x10,对应 ClientKeyExchange 消息报文。

2.2.6 ChangeCipherSpec 报文结构

ChangeCipherSpec 数据报文结构如表 7 所示。

表 7 ChangeCipherSpec 报文结构

HandShake	Version		Length1		Type
0x14	0x03		0x00	0x01	0x01

（1）由 SSL 协议格式可知 ChangeCipherSpec 报文第一个字节为 0x14，表示这是 ChangeCipherSpec 协议的一部分。

（2）第 6 个字节为 0x01，对应当前数据包为 Change-CipherSpecMassge 消息。

2.3 SSL 流识别

2.2 节介绍了 SSL 握手协议的通信过程以及报文结构。由于 SSL 握手协议是通过明文传输的，所以，可以通过解析捕获的 PCAP 文件获取数据包的头部信息，与上面几种不同类型消息的报文结构进行比对，从而可以识别出当前数据包是否为 SSL 握手协议的特定消息类型。对于一个完整的 SSL 会话，其通信过程一定包含以下几种类型的消息：ClientHello、SeverHello、SeverHelloDone、ClientKeyExchange、ChangeCipher-Spec。若某个数据流中没有检测到以上消息，则可以将其判断为非 SSL 流。若数据流检测到其中部分消息，则存在两种可能：一是 SSL 握手过程不完整，导致 SSL 连接建立失败，因此将这类数据流判断为非 SSL 流；二是这条数据流本身是 SSL 流，但抓包时由于网络延迟等各种原因，存在丢包的可能。若数据流中没有全部包含以上 5 种类型的消息，则将此流判断为非 SSL 流。反之，则判断为一个 SSL 流。具体识别流程如图 3 所示。

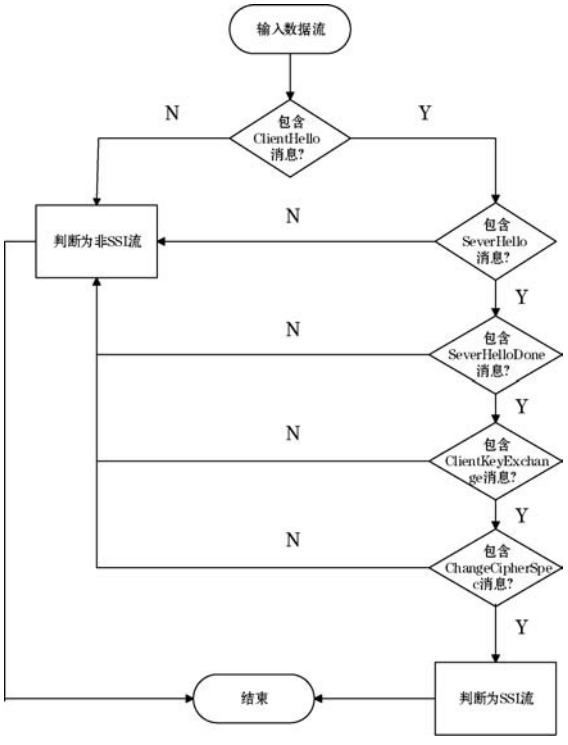

图 3 SSL 流识别流程图

在实际的实验环境中，通常会确定一个截断时间，把这个时间段内相同 ip 和端口号的数据包识别成一个流，而将超过截断时间的数据包分配到下一个流中。因此，可能出现如下情况，即一个流本身为 SSL 流，但并不是从其握手阶段开始截取的，而是从数据加密传输阶段开始截取的。这时，用上面的方法没有检测到 SSL 握手协议的消息，会产生漏识别的情况。为了解决这一问题，这里提出一种改进的指纹识别方法。

2.4 改进的指纹识别方法

使用 SSL 加密的数据包根据其消息类型的不同有不同的消息格式，但其前五个字节的格式是固定的，分别表示通信的阶段（握手（Handshake）、开始加密传输（ChangeCipherSpec）还是正常通信（Application）等）、SSL 协议版本号和剩余包长度，如表 8 所示。

表 8 SSL 协议前五字节格式

Content Type	Version		Length	
Byte1	Major	Minor	Byte4	Byte5

其中 ContentType 和 Version 的字节表示与其对应的类型见表 9 和表 10。

表 9 ContentType 对应信息

Hex	Dec	Type
0x14	20	ChangeCipherSpec
0x15	21	Alert
0x16	22	Handshake

表 10 Version Type 对应信息

Major Version	Minor Version	Version Type
3	0	SSLv3
3	1	TLS 1.0
3	2	TLS 1.1
3	3	TLS 1.2

改进的指纹识别方法通过 tcp 数据段的头几个字节信息识别当前数据包是否使用 SSL 协议加密，包括 ChangeCipherSpec、Alert、Handshake、Application 几种类型的消息。这种方法扩大了 SSL 流识别的范围，不仅能够识别 SSL 握手阶段的流，同时也能识别数据传输阶段的 SSL 流。具体的识别流程如图 4 所示。

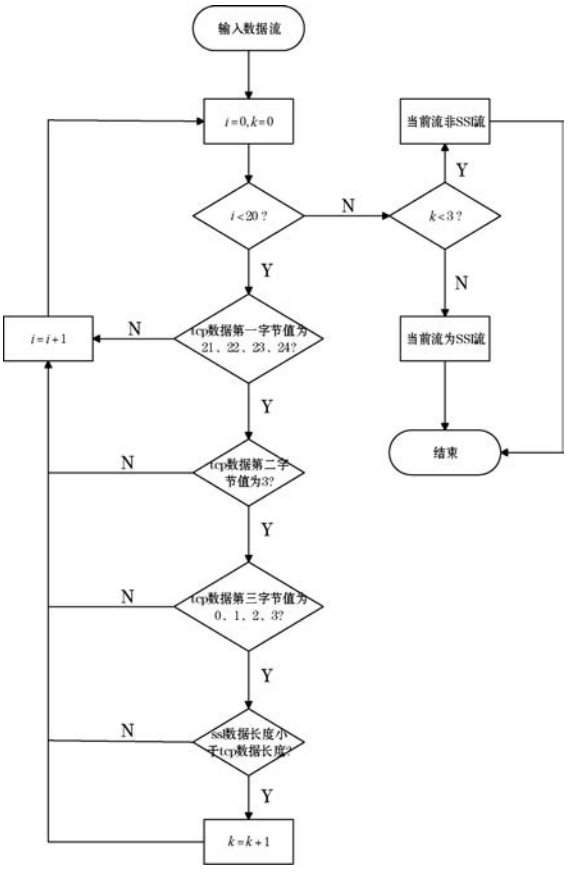

图 4　改进的 SSL 流识别流程图

3　改进的 RF 算法实现 VPN 流量识别

3.1　RF 算法介绍

与普通的机器学习方法相比,集成学习在实践中通常表现出更好的效果。集成学习利用多个学习器的协作来完成学习任务,有时又被称为多分类器系统、基于委员会的学习等[10]。

根据个体学习器的生成方式,目前的集成学习方法大致可分为两类:一是序列化方法,个体学习器之间存在强依赖关系,且必须以串行的方式产生;二是并行学习方法,个体学习器间不存在强依赖关系并且可以同时生成;其代表分别是 Boosting 和随机森林算法 RF。

随机森林算法是由 LeoBreiman(2001)提出的,它利用 Bootstrap 方法,从原始训练样本集 N 中有放回地重复随机抽取样本,生成 k 个新的训练样本集合(这 k 个新的训练样本集合可能存在部分重合),然后利用这些新的训练样本集构建随机森林。RF 算法本质上是对决策树算法的改进,将多个决策树结合到一起来做决策。由于森林中的每棵树都由不同的训练样本集构建,因此也有着不同的分布。

随机森林算法的构建过程如下:

(1) 从原始训练集出发,使用 Bootstraping 方法进行有放回的随机抽样,选出 m 个样本,重复进行 n 次采样,生成 n 个训练集。

(2) 对于 n 个训练集,分别训练生成 n 个决策树模型。

(3) 对于单个决策树模型,假设训练样本特征的个数为 N,那么每次分裂时根据信息增益/信息增益比/基尼指数选择最好的特征进行划分。

(4) 每棵树都按同样的方式分裂,直到该节点的所有训练样例不再可分。同时,在构建决策树的过程中不需要剪枝。

(5) 将构建好的多棵决策树组成随机森林,用投票的方式决定分类结果。

相比于传统的决策树算法,随机森林算法通过生成多棵树来避免过拟合的问题,为了使预测效果更好,应尽可能使生成的每棵树之间保持足够的差异性。因此,选择合适的参数十分重要,其中最为重要的两个参数分别是 n_estimators 和 max_features,它们分别代表弱学习器的最大迭代次数和 RF 划分时考虑的最大特征数。

本文的原始特征集合使用 Lashkari[6] 在其论文中提出的时间相关的流特征,一共 23 个特征,具体信息如表 11 所示。

表 11　识别 SSL VPN 流量使用的流特征

特征	描述
Duration	流持续时间
FIAT	正向分组到达时间间隔(均值、最大最小值、方差)
BIAT	反向分组到达时间间隔(均值、最大最小值、方差)
Flow-IAT	整个流的分组到达时间间隔(均值、最大最小值、方差)
Active	从活跃流变为静默流的时间(均值、最大最小值、方差)
Idle	从静默流变为活跃流的时间(均值、最大最小值、方差)
FB-psec	流每秒传输的字节数
FP-psec	流每秒传输的分组的数量

3.2　遗传算法介绍

遗传算法 GA 是根据遗传学知识和生物进化理论提出的一种模型,它通过多次迭代来实现特定目标的优化。遗传算法利用遗传、突变、杂交和自然选择的概念实现启发式搜索,最终得到最优解或准最优解,其流

程图如图 5 所示。

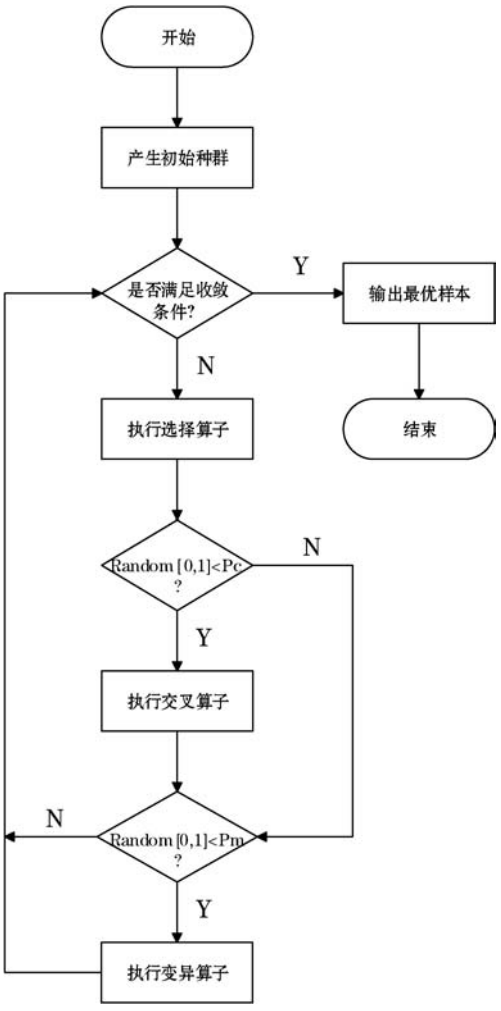

图 5 遗传算法流程图

3.3 基于参数优化的改进 RF 算法

前面讲到,RF 算法中有两个需要人为设定的重要参数,分别是 max_features 和 n_estimators,其中 max_features 表示随机森林允许单个决策树使用特征的最大数量。max_features 对于算法的性能有着重要的影响,当增加 max_features 时,通常可以提高模型的性能,因为我们在每个节点上都有更多的选择。然而,这并不完全正确,因为它降低了单个树的多样性,而这正是随机森林独特优势。但是,增加 max_features 会降低算法的速度。因此,有必要平衡和选择最佳 max_features。n_estimators 表示建立子树的数量,理论上来说,更多的子树可以使模型性能更好,使预测的结果更加稳定。但受限于机器的处理性能,较大的 n_estimators 会使代码变慢,大大降低模型的效率。因此,需要选择一个合适的 n_estimators 值,使得模型在满足性能要求的同时也具有较好的预测效果。

本文提出一种基于遗传算法改进的 RF 算法,通过遗传算法来实现 RF 算法中上面两个参数的优化。对于模型效果的评估,本文将流量识别的准确率与召回率作为评价标准,并将其作为遗传算法的适应度函数。

本文的特征集合里共有 23 个特征,为了比较不同 max_features 值的效果,需要选出当前条件下的最优特征子集。由于特征子集的选取存在的可能性较多,采用枚举的方式显然是不可取的。本文利用遗传算法,采用迭代的方式选出最佳的特征子集。将 23 个特征分别编号为 1 ~23,染色体的长度为 max_features 的值,染色体每个位点上的取值规定为 1 ~23 之间的一个常数,且不能重复出现。适应度函数设置为模型的 F 值,对于某一个特征子集模型的 F 值计算,为了保证结果的准确性,这里计算 10 次取平均值作为最后的 F 值结果。表 12 为分类结果混淆矩阵,F 值的计算公式如式(1) -式(3)所示:

$$查准率:P=\frac{TP}{TP+FP} \tag{1}$$

$$查全率:R=\frac{TP}{TP+FN} \tag{2}$$

$$F\ 值:F=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{样例总数+TP-TN} \tag{3}$$

表 12 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

算法构建如图 6 所示。

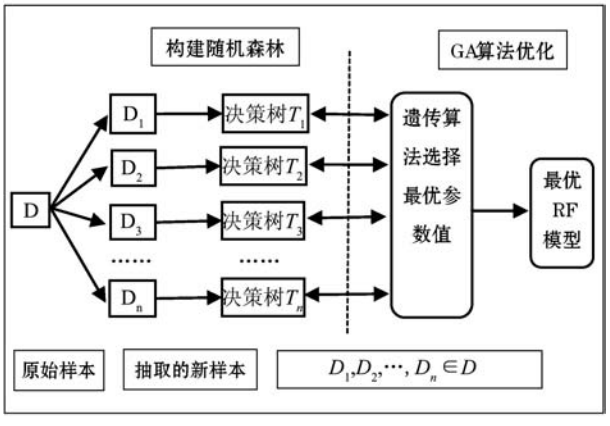

图 6 PGA-RF 算法构建

3.4 基于子分类器优化的改进 RF 算法

考虑到参数优化对分类器性能的影响有限,本文同时采用选择性集成方法与参数优化方法进行比较。选择性集成的概念由文献[11]提出,它利用某种策略从已构建的个体学习器集合中选出一部分构成新的集

成。研究表明选择性集通过把预测性能不好的个体学习器去除,只保留少量优质的个体学习器,能够提高集成预测性能。此外,选择性集成还可以提高集成的泛化能力。

随机森林由决策树构成,因此每棵决策树的好坏在很大程度上影响着模型的效果。随机森林最大的特点是通过增加样本扰动和属性扰动使个体学习器之间产生差异性,进而使集成的性能增加[12]。选择优质的个体学习器使得子分类器有更高的准确率,但这也可能导致个体学习器之间的差异性减少,从而使得集成模型的效果下降。为了解决这一问题,本文在生成决策树时限制可用特征的数量,从而增大每棵树之间的差异性。同时,通过遗传算法的多次迭代,可以保证筛选得到的集成模型的子分类器之间有较好的差异性。

基于子分类器优化的改进 RF 算法主要流程如下,首先用训练集构建一定数量的决策树,组成原始的决策树集合。然后根据选择性集成的思路,从原始的决策树集合中筛选出性能较优的决策树,构成新的决策树集合。最后利用遗传算法迭代多次得到最优的随机森林模型,构建方法如图 7 所示。

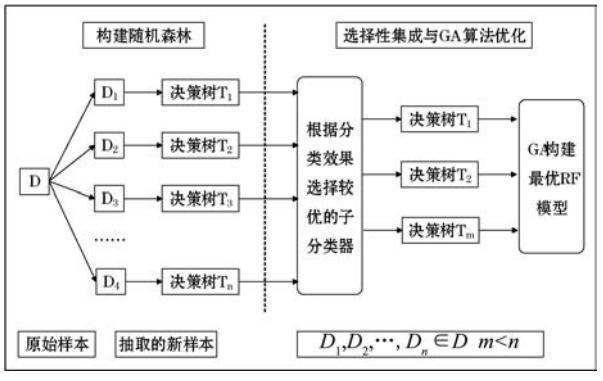

图 7　CGA-RF 算法构建

4　实验结果及分析

4.1　数据集

本文采用的数据集是 Lashkari 等[6]在 2016 年发布的 VPN-nonVPN 数据集,该实验室官网对数据集有详细介绍并且提供下载。为了产生一个有代表性的真实流量数据集,他们定义了一系列任务来保证数据的多样性和质量。他们创建了 Alice 和 Bob 两个账号来使用 Skype、Facebook 等服务,以产生相关流量。这些流量根据应用类型的不同共分为 7 个类别,每个类别又包括普通会话流量和 VPN 会话流量,因此共有 14 个类别的流量。

数据集的流量是通过 wireshark 和 tcpdump 捕获的,一共 28 GB 数据。对于 VPN 流量,实验使用 OpenVPN 进行 SSL VPN 连接。同时,对于文件传输流量,实验使用 Filezilla 作为客户端来产生 SFTP 和 FTPS 流量。不同类别的流量生成方式如表 13 所示。

表 13　数据集中各类流量的生成信息

流量类别	生成流量的相关应用
Browsing	Firefox 和 Chrome
Email	SMPTS, POP3S 和 IMAPS
Chat	ICQ, AIM, Skype, Facebook 和 Hangouts
Streaming	Vimeo 和 Youtube
File Transfer	Skype, 使用 Filezilla 产生的 FTPS 和 SFTP
VoIP	Facebook, Skype 和 Hangouts
P2P	uTorrent 和 Transmission

4.2　SSL 流识别结果

对于 SSL 流的识别,本文使用改进的指纹识别方法除 ICQ 等少数流量无法识别外,其余应用的 SSL 流识别率均达到 99% 以上,结果如图 8 所示。

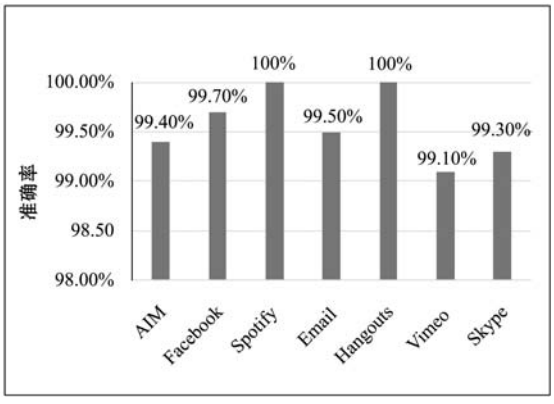

图 8　SSL 流识别结果

4.3　改进 RF 算法识别 VPN 流量实验结果

本文对比了 KNN(K 近邻)、LR(逻辑回归)、RF(随机森林)、PGA-RF(基于参数优化的改进 RF 算法)、CGA-RF(基于子分类器优化的改进 RF 算法)5 种算法的识别效果,分别对比了 Accuracy(准确率)、Precision(精确率)、Sensitivity(召回率)、Specificity(特异性)四个指标,结果如表 15 所示。

表 15　SSL VPN 流量识别结果

模型	准确率	精确率	召回率	特异性
KNN	0.835	0.839	0.825	0.839
LR	0.686	0.704	0.667	0.681
RF	0.903	0.909	0.892	0.907
PGA-RF	0.916	0.921	0.911	0.923
CGA-RF	0.922	0.926	0.919	0.921

从实验结果可以看出,LR 的准确率为 68.6%,KNN 的准确率达到 83.5%,而 RF 的准确率达到 90.3%,可见集成算法的效果要明显好于普通的机器学习算法。同时,基于参数优化的改进 RF 算法准确率为 91.6%,基于子分类器优化的改进 RF 算法准确率为 92.2%,其效果都要优于直接使用 RF 算法的效果,且 CGA-RF 的效果要优于 PGA-RF,可见选择性集成对模型效果有所提升,实验结果对比如图 9 所示。

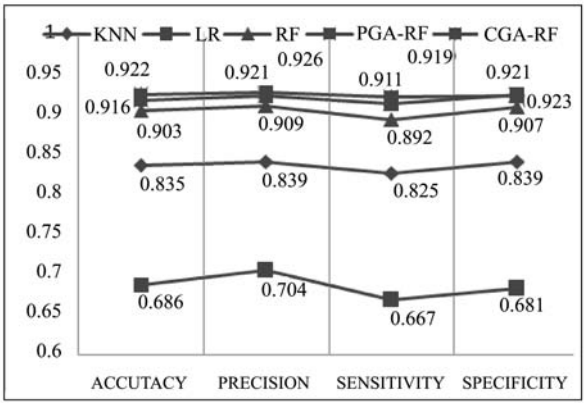

图 9 SSL VPN 识别结果

5 结 语

由于网络中的 SSL 加密流量越来越多,对于加密流量的监管也变得越来越重要。本文提出的混合方法,将指纹识别与机器学习方法相结合,实现了 SSL VPN 加密流量的识别。对于 SSL 流的识别,本文提出的改进的指纹识别方法对流漏识别的情况有所改善,使得识别效果更好。对于 SSL VPN 的识别,本文提出的改进 RF 算法也对准确率有所提升。实验结果表明,当前方法能够到达 92.2% 的准确率,实现了 SSL VPN 加密流量的有效识别。

参 考 文 献

[1] Paxson V. Empirically derived analytic models ofwide-area TCP connections[J]. IEEE/ACMTransactions on Networking,1994,2(4):316-336.

[2] Gabriel Gómez Sena, Belzarena P. Early trafficclassification using support vector machines[C]//International Latin American NETWORKING Conference, Lanc 2009, Pelotas, Brazil, September. DBLP,2009:60-66.

[3] Dainotti A, Pescapé A, Rossi P S, et al. Internet traffic modeling by means of Hidden Markov Models[J]. Computer Networks, 2008, 52(14):2645-2662.

[4] Schatzmann D, Spyropoulos T, Dimitropoulos X. Digging into HTTPS: flow-based classification ofwebmail traffic[C]// ACM SIGCOMM Conference onInternet Measurement. ACM, 2010:322-327.

[5] Sun G L,Xue Y, Dong Y, et al. An novel hybrid method for effectively classifying encrypted traffic[C]//Global Telecommunications Conference. IEEE, 2010:1-5.

[6] Lashkari A H, Draper-Gil G, Mamun M S I, et al. Characterization of encrypted and VPN traffic using time-related features[C]//The International Conference on Information Systems Security and Privacy,2016:94-98.

[7] Bagui S, Fang X, Kalaimannan E, et al. Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features[J] Journal of Cyber Security Technology,2017,1(2):108-126.

[8] 苏皞昕, 施勇, 薛质. 基于 SSL 流量的指纹识别[J]. 信息安全与技术, 2015(11):58-60.

[9] Internet Engineering Task Force(IETF). The Secure Socket Layer(SSL) Protocol Version 3.0[OL]. 2011. http://tools.ietf.org/html/rfc6101.

[10] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.

[11] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002,137(1/2):239-263.

[12] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(8):832-844.

(上接第 227 页)

[8] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2013, 29(9): 30-34.

[9] 李跃鹏, 金翠, 及俊川. 基于 word2vec 的关键词提取算法[J]. 科研信息化技术与应用, 2015,6(4): 54-59.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. eprint arXiv: 1301.3781, 2013.

[11] Flick C. ROUGE: A package for automatic evaluation of summaries[C]//The Workshop on Text Summarization Branches Out. 2004: 10.

[12] 吴晓倩, 胡学钢. 基于 N-最短路径的中文分词技术研究[J]. 安徽理工大学学报(自科版), 2014(1): 72-75.

(上接第 283 页)

[12] Fan Z, Su L, Liu X, et al. Multi-label Chinese question classification based on word2vec[C]//International Conference on Systems and Informatics. IEEE, 2018:546-550.

[13] 史兆鹏,邹徐熹,向润昭. 基于依存句法分析的多核心词义消歧[J]. 计算机工程,2017,43(9):210-213.

[14] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[15] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[16] 周练. Word2vec 的工作原理及应用探究[J]. 图书情报导刊, 2015(2):145-148.