基于深度森林的网络匿名流量检测方法研究与应用

李成豪1，魏松杰1

（1. 南京理工大学计算机科学与工程学院，江苏省南京市 210014）

**摘要：**流量分类一直是许多工作研究的主题，但是互联网服务的快速发展和加密的普遍使用使其成为一个公开的挑战。加密对于保护互联网用户的隐私至关重要，这是近年来出现的各种隐私增强工具中使用的一项关键技术。Tor匿名通信系统是其中最受欢迎的一种，基于匿名通信系统Tor构建的暗网是现今规模最大的暗网实体，其常被犯罪分子用来从事各类违法活动，为犯罪分子提供了隐匿空间，给网络监管带来了十分严峻的挑战。因此对Tor流量进行高效识别对网络监管和打击犯罪有着重要意义。本文根据Tor匿名通信流量特点设计了一组用于Tor流量行为检测的网络流特征，针对在原有深度森林模型内存需求和时间开销大的不足基础之上设计了一种改进的深度森林模型用于Tor匿名网络流量的识别。实验结果表明，与已有识别方法相比，提出的模型准确率可达99.86%，并且在时间和内存的花费有着明显的降低。

**关键词 :** 流量分类 ; Tor ; 暗网 ; 匿名流量 ; 深度森林

**Research and application of Network Anonymous Traffic Detection Method based on Deep Forest**

LI Cheng-hao1, WEI Song-jie1

(1. *Department of Science and Engineering,* *Nanjing University of Science and Technology*, *Nanjing, Jiangsu Province* 210014, *China*)

**Abstract：**Traffic classification has been the subject of much work, but the rapid development of Internet services and the pervasive use of encryption make it an open challenge. Encryption is critical to protecting the privacy of internet users, a key technology used in various privacy-enhancing tools that have emerged in recent years. The Tor anonymous communication system is one of the most popular ones. The dark web based on the anonymous communication system Tor is the largest dark web entity today. It is often used by criminals to engage in various illegal activities, providing criminals with a hiding place. space, which brings very serious challenges to network supervision. Therefore, the efficient identification of Tor traffic is of great significance for network supervision and crime fighting. According to the characteristics of Tor anonymous communication traffic, this paper designs a set of network flow characteristics for Tor traffic behavior detection, and designs an improved deep forest model based on the shortcomings of the original deep forest model's memory requirements and high time overhead. Identification of Tor anonymous network traffic. The experimental results show that, compared with the existing recognition methods, the accuracy of the proposed model can reach 99.86%, and the cost of time and memory is significantly reduced.

**Key words：**traffic classification; Tor; darknet ; anonymous traffic; deep forest

**0 引言**

随着互联网的飞速发展，互联网的规模在不断的扩大，逐渐渗透到社会、经济、政治的方方面面，而互联网中个人的安全和隐私保护问题日益严峻。匿名网络设计应运而生，其初衷虽然是为了保护用户的隐私，但往往被不法分子滥用，以逃避网络追踪，进而实施恶意网络活动。由于匿名通信网络使用加密、多节点转发和填充等多种流量混淆方法隐藏了原始流量的特点，这使得传统流量识别技术已经无法准确识别匿名流量。个人用户、企业和政府每天都会使用各种网络应用，每天都会产生数以亿计的网络流量，恶意流量很容易将自己隐藏在海量的流量中，进而发动攻击，给网络安全带来极大的威胁。因此研究如何高效准确的识别海量流量中具有恶意的网络行为，对于网络安全发展具有极其重要的科学研究意义。

Tor作为目前使用最广泛的匿名通信系统之一[1]。它在2004年推出了一项称为隐藏服务的技术，实现了客户端与服务端的双端匿名。通过使用隐藏服务技术，任何服务运营商都可以在提供网络服务的同时隐藏其服务器的物理地址。所以Tor拥有的用户最多，分布范围最广，平均每天的用户数量超350万人[2]。朴茨茅斯大学的Owen Gareth等研究人员在为期六个月的研究中观察到80%以上的隐藏服务的访问请求都指向已知的著名虐童网站。此外，枪支与毒品交易也是暗网市场不可或缺的部分。因此Tor检测一直是学术界研究的主题之一，但其中许多研究侧重于降低Tor的匿名性或提高其性能。对Tor流量的分析与识别的研究少之又少。因此对Tor匿名网络流量进行分析与识别具有很强的代表性。

由于Tor采用了TLS加密技术，检测者无法提取Tor流量内部的有效特征载荷，加密网络流量给网络安全防御带来了巨大的挑战，在不加解密的基础上识别Tor匿名流量具有十分重要的研究意义。Tor 虽然通过加密和封装等方式将非正常流量隐藏于普通流量，但是仍然可以基于Tor网络流特征的微小差异对其进行识别。随着机器学习的快速发展，越来越多的研究人员将机器学习技术应用于对Tor匿名流量的识别中，并取得了一定的研究成果，但也存在模型复杂度高、训练时间长、需要庞大的训练数据集作支撑、模型泛化能力差等方面缺陷。

通过阅读大量文献发现在已有的研究工作中，研究人员使用Netmate、pacp2flow、Tranalyzer等工具进行流量特征的提取，但这些工具多数都是基于数据明文字段进行计算，对于强加密的Tor匿名流量并没有很好的效果。通过研究Tor匿名通信系统本身的工作特性，发现以其通信频率、时延等为切入点计算时间相关特征具有很强的区分性[13]，因此本文通过CIC Flow Meter处理Tor数据集产生的83中流量特征进行评估，最终选择重要性前35名特征作为Tor流量检测的时间相关特征。为解决传统机器学习模型中存在的缺陷，本文通过构建改进深度森林模型最终取得了较高的精确率和召回率。

**1 相关工作**

迄今为止，国内外学者从不同角度对网络流量分类方法进行了研究。早期通过Tor匿名通信系统的端口号、IP通信地址等进行识别。基于端口识别的分类方法在早期可以通过检查分组的传输层端口号[4]，然后根据IANA（Internet Assigned Numbers Authority）定制的知名端口号与注册端口号列表将分组与应用对应起来。随着网络应用的不断更新发展Tor流量也对其数据包进行了加密处理无法检测其端口号，进而导致基于端口的Tor流量方法检测已不再适用，Moore等人通过实验发现，此方法的准确率不足70%。目前根据原理和方法将混淆流量检测方法分为深度包和基于机器学习的流量检测技术。

深度包检测技术（Deep Packet Inspection, DPI）的方法出现在上个世纪90年代[5]。这种方法不仅仅检测包的端口，同时也对包头和包的数据进行检测。基于深度包检测的方法优势主要有两个：一是非常高的准确率；二是可以在早期就检测出异常行为。但是基于深度包检测的方法存在着计算复杂度高、无法处理加密信息、指纹获取困难、无法检测位置网络数据等缺点。LIU[1]等人指出,利用深度包检测技术对加密流量进行检测,只有30%~70%的准确率。（Liu X B, Yang J H, Xie G G, et al. Automated mining of packet signatures for traffic identification at application layer with apriori algorithm[J]. J Commun, 2009, 29(12): 51-59.）

随着人工智能的不断发展，机器学习广泛应用于流量识别与分类任务并得了良好的成果。机器学习方法和入侵检测系统相结合，充分发挥机器学习方法强大的数据学习能力和判别能力，为网络特征流量检测增添了新的技术手段。主流的机器学习分类方法有支持向量机[6]、决策树[7]、贝叶斯方法[8]和随机森林[9]等。李晓明等人对比了贝叶斯网络、支持向量机和朴素贝叶斯算法在网络匿名流量识别上的精确率和召回率。（李晓明, 任慧, 颜金尧. 基于机器学习的网络流量分类算法分析研究[J]. 中国传媒大学学报: 自然科学版, 2017, 24(2): 9-14.）机器学习方法应用在网络特征流量检测系统的核心思想是：根据网络流量信息（网络服务类型、IP地址、协议类型、连接状态等）和主机信息（文件访问权限、主干资源访问、应用程序进程等）来建立相应的机器学习模型，进而对正常的流量和异常流量进行分类。

深度学习最大的贡献是表征学习，即通过端到端的训练，发现更好的特征[10]。表征学习的目的是自动学习有用的数据特征，对复杂的原始数据化繁为简，把原始数据中的无效信息剔除，把有效信息更有效地进行提炼。

综合以上研究现状，本文使用的深度森林[11]是一种基于决策树的集成学习方法，有着很强的表征学习能力。深度森林包含级联森林和多粒度扫描两个主要组件。级联结构使得深度森林可以进行表征学习。在级联森林中，每层都是由多个决策树森林的集成组成，也就是说，每层都是一个集成的集成。级联的每层接收上一层处理得到的输出信息作为其增广的输入信息。对于特征之间存在关系信息的数据，深度森林通过一种称为多粒度扫描的技术进一步增强其表征学习的能力。深度森林不像深度学习，不需要复杂的超参数和庞大的数据集依然可以得到很高准确率。但是深度森林由于所有的样本都需要通过深度森林的所有层级，这就导致时间复杂度随层数线性变化。其次滑动窗口使每一个样本产生成百上千个新样本，造成了巨大的计算开销。为了解决深度森林存在的问题，本文在深度森林的基本框架中引入了置信筛分机制[12]用以降低深度森林的内存需求和时间开销。