基于深度森林的网络匿名流量检测方法研究与应用标题

李成豪1，魏松杰2

（1. 南京理工大学计算机科学与工程学院，江苏省南京市 210014；2. 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：**流量分类成为许多研究工作的主题，但是互联网服务的快速发展和加密技术的普及使得它成为一个开放的挑战。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流行的一种匿名流量通信工具之一，它通过加密发送方和接收方之间的通信，并通过分布式服务器网络路由来实现发送方和接收方的解耦。Tor匿名通信网络的发展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。本文针对Tor匿名流量特点在原有深度森林的不足之上设计了一种基于改进深度森林模型用于Tor匿名网络流量行的识别。实验结果表明，与已有识别方法相比，提出的识别模型在准确率、时间、内存等方面有着明显的提升。

**关键词 :** 流量分类 ; Tor ; 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-jie2

(1. *Department of Science and Engineering,* *Nanjing University of Science and Technology*, *Nanjing, Jiangsu Province* 210014, *China*; 2. *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 research work, but the rapid development of Internet services and the proliferation of encryption techniques 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 anonymous traffic communication tools. It realizes the decoupling of the sender and the receiver by encrypting the communication between the sender and the receiver and routing through a distributed server network. The development of Tor anonymous communication network provides a hiding space for criminals and brings a very serious challenge to network supervision. In view of the characteristics of Tor anonymous traffic and the shortcomings of the original deep forest, this paper designs an improved deep forest model for the identification of Tor anonymous network traffic lines. The experimental results show that compared with the existing recognition methods, the proposed recognition model has obvious improvements in terms of accuracy, time and memory.

**Key words：**traffic classification; Tor; 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。互联网逐渐渗透到社会、经济、政治的方方面面，互联网中人们的安全和隐私保护问题日益严峻。匿名网络设计的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，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。所以研究如何准确高效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，它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已经超过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己知的著名虐童网站。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。

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。

本文根据Tor工作的特点，采取时间相关特征[3]对其进行设计了一组用于Tor流量行为检测的网络流量特征，通过CIC Flow Meter对Tor匿名接入流量的pcap文件进行处理，得到关于时间相关性流量特征的数据集。并提出了基于深度森林的Tor匿名流量检测模型，可以在少量的数据集以及较短的时间和较少内存消耗取得较高的准确率。

# 1 相关工作

目前为止，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。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。

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。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。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。

通过阅读大量文献我们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，我们发现以其通信频率、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。通过构建深度森林模型最终取得了较高的精确率和召回率。

# 2 匿名流量检测系统

## 2.1 匿名流量检测流程

本实验首先对数据集进行了特征提取、数据预处理等操作，然后使用改进的深度森林模型从混合流量中检测出 Tor 流量如图3.1所示。

## 2.2 Tor匿名流量特性分析

从Tor原始流文件中提取的数据集中包含大量的空值高纬稀疏数据，严重影响对Tor流量预测的准确性。本文通过分析Tor流量的特性，通过筛选用较少的特征便可以精准描述Tor流量行为。

用户使用Tor客户端发送请求，在Tor匿名通信系统中默认需要经过是三个节点的转发才能到达目标服务器，而且三个节点一般部署在不同国家的不同地区，因此Tor流量的转发时延相对于正常流量会大，可以使用数据包时延、流持续时间等特征描述转发时延差异。

通过研究发现Tor一般将应用程序数据包封装成固定的长度，所以Tor数据包长度的变化范围是有限的，Tor和NonTor流量在数据包长度特征上存在明显差异，可以用此特征来描述数据包大小变化的情况和差异性。

通过研究发现用户在建立Tor连接之后，如果在一段时间内应用程序不进行任何操作，网络处于空闲状态；如果应用程序进行相关操作，将会产生持续的网络访问，网络处于忙碌状态，Tor匿名通信系统无法隐藏这些网络流特征，因此可以使用流变成活动状态之前的空闲时间和流变成活动状态的空闲时间的特征描述Tor这一特点。

综上所述， 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Tor匿名网络具有不可避免的时延、流持续时间、数据包长度分布等特性仍然存在，根据对这些特征的处理，依然可以区分Tor和NonTor流量。

## 2.3 Tor匿名流量特性分析

为了降低原始数据集中特征的冗余度，提高检测的准确率，本文在Tor和NonTor匿名流量数据集上对通过CIC Flow Meter产生的83种流量特征进行了评估。

本文尝试通过提取出一些流超时时间的数据集的基于时间的特征，我们将Tor数据集中的所有流标记为Tor。将其和相同流超时时间的加密NonTor数据集进行合并用作输入。对每个流超时时间数据集（2s，4s，8s，16s，32s，64s，128s）应用多种不同的特征选择方法，最终我们选择使用Infogain+Ranker特征选择算法在处理16s数据集取得的特征获得了最佳的实验结果如表1所示。

**表1 特征选择结果**

|  |  |
| --- | --- |
| 特征 | 相关度 |
| Flow Bytes/s | 0.8778 |
| Flow Duration | 0.8745 |
| Fwd IAT Mean | 0.7135 |
| Bwd IAT Mean | 0.7134 |
| Flow IAT Std | 0.7034 |
| Fwd IAT Min | 0.6961 |
| Flow IAT Mean | 0.6921 |
| Bwd IAT Min | 0.6874 |
| Idle Max | 0.6868 |
| Idle Min | 0.6755 |
| Idle Mean | 0.6726 |
| Flow Bytes/s | 0.6692 |
| **…** | **…** |

如上表所示，本文最终选择了特征相关度前35名的特征作为进行Tor匿名流量检测的相关性特征。它们具有Tor匿名流量检测的最大信息增益和卡方统计量。数据集提取的特征属性及其描述如下表2所示

**表2 特征描述**

|  |  |
| --- | --- |
| 特征属性 | 描述 |
| Fwd IAT mean/max/min/std | 正向流中包到达间隔的 mean/max/min/std |
| Fwd Packet Length mean/max/min/std | 正向流中包长度的 mean/max/min/std |
| Bwd IAT mean/max/min/std | 反向流中包到达间隔的 mean/max/min/std |
| Bwd Packet Length mean/max/min/std | 反向流中包长度的 mean/max/min/std |
| Flow IAT mean/max/min/std | 包到达间隔时间的 mean/max/min/std |
| Active mean/max/min/std | 流变为空闲态之前活动时间的 mean/max/min/std |
| Idle mean/max/min/std | 流变为活动态之前空闲时间的 mean/max/min/std |
| Total Fwd Packet | 正向数据包总数 |
| Total Bwd packets | 反向数据包总数 |
| Total Length of Fwd Packet | 正向数据包总大小 |
| Total Length of Bwd Packet | 反向数据包总大小 |
| Flow Bytes/s | 每秒流字节数 |
| Flow Packets/s | 每秒流数据包数 |
| Flow Duration | 流持续时间 |

## 2.4 流量数据预处理方法

为了获得更加有效的Tor和NonTor流量数据集，本文对获得的数据集进行预处理。

### 2.4.1 消除重复数据

通过CIC Flow Meter设定流超时时间将不同类型的pcap文件保存到csv文件中并将其合并，通过研究发现无论是Tor数据集还是加密的NonTor数据集都存在大量的重复数据流，并发现有的重复率高达30%。这些重复数据不仅没有研究价值，反而容易使得实验结果虚高。

### 2.4.2 消除重复数据

通过无论是Tor数据集还是NonTor数据集其中特征的取值范围跨度较大，例如特征Flow IAI Max的取值范围为[1，15998544]，这将严重影响模型的收敛速度和分类性能。于是本文采用MIN\_MAX数据归一化方法，将所有特征范围归一化到[0，1]之内。

## 2.5 检测目标定义和效果评价方法

Tor流量行为检测的主要目标就是从混合流量中检测Tor匿名流量，属于二分类实验。在匿名流量检测实验中，数据集中样本往往都是不平衡的。若正负样本比例差别非常大的情况下仍依靠准确率评判是不科学的。例如正常样本有99000条数据，异常样本只有1000条数据。在识别过程中即使所有的数据全判断为正常样本准确率也可以高达99%，此时准确率指标不仅不能表示性能很好，相反此结果是非常差的。本文使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1-Score作为评价指标来综合判断Tor匿名流量检测的效果。

# 3 网络匿名流量检测模型设计

## 3.1深度森林简介

深度森林是一种新颖的决策树集成学习[14]方法，具有很高的预测精度。深度森林模型包含两个核心组件，即级联森林结构和多粒度扫描。级联结构使得深度森林可以进行表征学习。级联森林的每层由多个随机森林分类器和完全随机森林分类器组成，也就是说每层都是一个集成的集成。每层通过接受上层的结果来对其输入信息进行扩展。多粒度扫描使用滑动窗口扫描原始特征，并通过使用多个尺寸的滑动窗口，获得更多的变换的特征向量，进而增强级联森林，这使得深度森林模型在处理时序特征数据或空间特征数据的效果表现的十分好。

## 3.2改进深度森林模型设计

尽管Zhou等人[15]的实验结果表明进一步增大深度森林的模型将得到更好的效果。但如果在多粒度扫描中使用更多不同的滑动窗口产生多更的变换的特征向量，级联森林的每层随机森林中将使用更多的决策树，与此同时深度森林模型的训练和预测将需要更高的内存需求和时间开销。导致这种问题的主要原因有两方面。首先所有的样本都要经过级联森林的每一层，导致时间复杂度随着层数的增加而线性增加。其次在多粒度扫描中，使得每一个原始样本都会产生成百上千个变换的新样本，大幅度增加了训练样本，造成了巨大的计算开销。

### 3.2.1改进的级联森林

为解决所有样本都要经过级联森林的每一层而带来的时间和内存的开销，本文在级联森林结构引入了置信筛分机制。具体来说置信筛分是将级联的每一层的实例分为两个子集：一个是容易预测的；另一个是难以预测的。如果一个实例很容易预测就直接当成最终结果输出，只有当一个实例很难预测时，它才需要被传递到下一层。如图4.1改进级联森林所示.

改进深度森林的级联结构如上图所示。基于置信机制的深度森林与原始深度森林主要的不同点有两个方面。首先改进的深度森林在级联的每一层都有门结构，将样本集划分为容易预测的样本和难以预测的样本两个子集。如图所示具有高预测置信度的样本（Y）将直接输入到最后不需要经历到后续的所有层；只有低预测置信度的样本（N）被传递到下一层。其次，由于后续随之样本数量的不断减少，剩余样本的预测难度也越来越大。基于置信筛分机制的深度森林将逐层增加每个森林中决策树的数量用来增大森林模型复杂度来训练剩余的样本，而原始深度森林模型每层使用相同复杂度的深度森林模型。实验结果证明，随着层数的增加，从低到高增加级联中森林的复杂度可以使得模型具有更好的泛化误差上界。

对于一个三分类问题，一个样本的预测类别分布向量为，则该样本的预测置信度为0.7。如图4.2所示。

如何设置级联森林每一层的置信度阈值是一个关键的问题。本文使用的是一种简单的阈值设置准则，自动化的确定每一层所需要的置信度阈值，从而在保证深度森林的预测性能的同时，大幅度提升其计算效率。在级联的第t层，预测置信度阈值是基于当前层级的交叉验证错误率自动确定。令超参数表示高置信度样本所需要达到的交叉验证错误率。将所有的训练样本按照它们的预测置信度降序排序，其中表示样本的预测置信度。置信度阈值的设置方式如下：,（4.1）其中表示预测置信度最高的k个样本的交叉验证错误率。

### 3.2.2改进的级联森林

深度森林中的多粒度扫描过程会使得训练模型的内存和时间开销大大增加。为了解决这一问题，改进的多粒度扫描，本文采取采样多粒度扫描方法。在采样多粒度扫描中，首先通过滑动窗口生成新的样本，再从新的样本中随机采集一个子集用于后续的森林模型的训练。本文只考虑二分类问题如图4.3所示，随机采样不仅能够大幅度降低新样本的树木，还能够将多粒度扫描过程中所生成的转换特征的维度大幅度降低——从903维降低到90维。随着级联森林层数的增加，置信筛分将更难预测的样本保留下来，为此在每一层都新增采样多粒度扫描产生的转换特征增强其表示。

基于置信机制的深度森林主要包含三个部分：置信筛分机制、可变模型复杂度机制和采样多粒度扫描。如果原始输入为400维属性特征，大小为100维的滑动窗口被用于采样多粒度扫描，采样规模为30。对于n个训练样本，该滑动窗口将产生包含个新的训练样本，每个样本有100维的新的训练集。使训练集得到两个不同类型的森林（随机森林和完全随机森林）。对于一个三分类问题，两个森林模型会产生180维的转换特征。该转换特征被用于后续级联森林的输入。直至达到预设的终止条件。

级联森林的每层将Tor流量数据集划分为两个子集。只将难以预测的样本子集传入下个层级，并随着样本的减少，增大森林的复杂度。给定一个测试样本，其特征表示随着层数的增加不断扩增。如果该样本的预测置信度超出了当前层级的置信度阈值，则根据当前层级的输出结果作为该样本的最终预测。如果该样本传递到下一层。在最坏情况下，该测试样本将遍历所有的层级在最终层获得其预测结果。预测结果为输出的类别分布向量中的最大类别。

# 4 实验

基金项目:国家自然科学基金[59637050]

作者简介：姓名（出生年—），性别，籍贯，职称，学位，主要研究方向为……。（多个作者间用分号隔开）

通信作者：姓名 邮箱

**参考文献**

参考文献不少于15条，并按照先后顺序在正文中标注，若引用中文参考文献需提供对应的英文参考文献且英文在前中文在后，必须按以下格式编写。

|  |  |
| --- | --- |
| 期刊 | [序号] 作者.论文题目[J].刊名，年，卷（期）：起页-止页.  例：WEN Weiping, GUO Ronghua, MENG Zheng,et al.Research and Implementation on Information Security Risk Assessment Key Technology[J].Netinfo Security，2015，15（2）：7-14.  文伟平，郭荣华，孟正，等. 信息安全风险评估关键技术研究与实现[J]. 信息网络安全，2015，15（2）：7-14. |
| 专著 | [序号] 作者.书名[M]. 出版地：出版社，出版年.  例：尼葛,洛蒂,苏娟,等.森林群落分类［M］.胡闵,译. 3版. 北京:科学出版社,1998. |
| 论文集 | [序号] KHAN M A, SHAH G A, AHSAN M, et al**（作者名）**. An Efficient and Reliable Clustering Algorithm for Wireless Sensor Actor Networks (WSANs)**（文章题目）**[C]∥IEEE**(组织者)**. 53rd IEEE International Midwest Symposium on Circuits and Systems**（会议名称）**, Auguest 1-4, 2010（**会议时间**）, Seattle, WA, USA（**会议地点**）. New York（**论文集出版地**）：IEEE（**论文集出版社**）, 2010（**出版时间**）：332-338（**页码**）. |
| 学位论文 | [序号] 作者. 论文题目[D]. 保存地：保存者，年份.  例：张志强. 随机扰动及其模型［D］.北京: 北京大学,2002. |
| 技术标准 | [序号] 标准编号 标准名称[S]. 出版地：出版者，出版年.  例：GB/T1569—2001 中国标准书号［S］.北京:标准出版社,1998. |
| 专利 | [序号] 专利所有者. 专利题名：专利国别，专利号[P]. 发布日期.  例：刘加林. 多功能一次性压舌板: 中国,92214986.2［P］.1993-04-14. |
| 技术报告 | [序号] 作者. 题目[R]. 地名：责任单位，报告代码及编号，年份. |
| 报纸文章 | [序号] 作者. 题目[N]. 报纸名，出版日期（版次）.  例：张田. 罪犯DNA为与生命伦理学［N］. 大众科技报,2000-11-12(7). |
| 电子文献 | [序号] 主要责任者. 电子文献题名[EB/OL]. 电子文献的出处或可获得地址，发表或更新日期(任选)/引用日期. |

**注意事项：**在英文参考文献中，责任者为西方人，姓在前名在后，姓全用大写，名可缩写，只用首字母，如FOURNEY M E；责任者为中国人的，姓在前，名在后，不可缩写，如XIAO Chuanqi。