

遵 义 师 范 学 院

本 科 毕 业 论 文（设计）

|  |  |
| --- | --- |
| 题 目： | **基于流量特征的远控木马检测方法分析与实现** |

|  |  |
| --- | --- |
| 学 院： | 信息工程学院 |
| 专 业： | 网络工程 |
| 年 级： | 2021级 |
| 姓 名： | 胡涛 |
| 学 号： | 2140909050230 |
| 校内指导教师： | 皮奔（助教） |
| 校外指导教师： | 无 |

**摘要**

远控木马作为一种恶意程序，可对受感染的计算机系统实施远程控制，进而窃取敏感信息、破坏系统数据，严重威胁信息安全。因此如何实现对远控木马的快速精准检测，成为当前信息安全领域研究的热门问题。现有研究者提出了许多远控木马流量检测方法。但这些方法大多需要对较长时间段的通信流量进行分析，不可避免地导致检测延迟，难以满足及时检测远控木马攻击的需求。为了能够尽早检测出远控木马流量，本文系统分析了远控木马与正常软件在前5个下行负载包流量范围内的通信行为差异。提取具有代表性的流量特征，并构建了木马通信会话快速检测模型。实验结果显示，本文所提出的方法取得了95.1%的检测准确率。验证了该方法在快速检测远控木马方面的有效性。

**关键词：**远控木马，特征分析，机器学习

Abstract

Remote access Trojans, as a type of malicious software, can remotely control infected computer systems, thereby stealing sensitive information and destroying system data, posing a severe threat to information security. Therefore, how to achieve rapid and accurate detection of remote access Trojans has become a hot research topic in the field of information security.Existing researchers have proposed many methods for detecting remote access Trojan traffic. However, most of these methods require the analysis of communication traffic over a relatively long period, inevitably resulting in detection delays and failing to meet the need for timely detection of remote access Trojan attacks.To detect remote access Trojan traffic at an early stage, this paper systematically analyzes the differences in communication behaviors between remote access Trojans and normal software within the scope of the first five downstream payload packets. Representative traffic features are extracted, and a rapid detection model for Trojan communication sessions is constructed.The experimental results show that the method proposed in this paper achieves a detection accuracy of 95.1%, which verifies the effectiveness of this method in the rapid detection of remote access Trojans.

**Key words**: Remote Access Trojan ,Feature analysis,Machine learning

目录

[1 引言 1](#_Toc23957)

[1.1 研究背景和意义 1](#_Toc27764)

[2 相关工作 3](#_Toc22597)

[2.1基于主机的检测方法 3](#_Toc15607)

[2.2 基于网络的检测方法 4](#_Toc539)

[2.2.1 基于流量负载的检测方法 4](#_Toc22000)

[2.2.2 基于网络行为的检测方法 4](#_Toc14816)

[3远控木马概述 7](#_Toc21558)

[3.1 木马病毒概述 7](#_Toc17780)

[3.2 远控木马攻击阶段 7](#_Toc14566)

[3.3 远控木马通信过程 8](#_Toc12605)

[4 相关理论与技术 11](#_Toc6053)

[4.1 机器学习概述 11](#_Toc24531)

[4.2 主要类型 11](#_Toc6156)

[4.2.1监督学习 11](#_Toc29765)

[4.2.2无监督学习 11](#_Toc27485)

[4.2.3强化学习 11](#_Toc21368)

[4.3 常见算法 12](#_Toc8432)

[4.3.1 随机森林 12](#_Toc7519)

[4.3.2 支持向量机 12](#_Toc22266)

[4.3.3 k近邻 12](#_Toc17702)

[5 实验及分析 13](#_Toc29338)

[5.1 实验环境 13](#_Toc26502)

[5.2 实验数据 13](#_Toc8536)

[5.3 评估标准 15](#_Toc14550)

[5.4 特征分析 16](#_Toc14189)

[5.4.1下行数据包数量 16](#_Toc28830)

[5.4.2下行数据包负载量 17](#_Toc22938)

[5.4.3最大通信时间间隔 18](#_Toc1709)

[5.4.4上下行数据包负载比 19](#_Toc5711)

[5.5 结果分析 20](#_Toc30216)

[5.5.1 实验结果 20](#_Toc29041)

[5.5.2 对比实验 20](#_Toc7212)

[6 总结与展望 23](#_Toc16138)

[6.1结论 23](#_Toc16244)

[6.2 展望 2](#_Toc19611)3

[参考文献 25](#_Toc23975)

[致谢 27](#_Toc5617)

1 引言

1.1 研究背景和意义

中国互联网络信息中心发布的第55次《中国互联网络发展状况统计报告》[1]显示，我国网民规模从1997年的62万人增长至2024年的11.08亿人，互联网普及率升至78.6%。随着互联网用户的剧增与数字基础设施的覆盖，网络攻击事件呈现爆发式增长，网络安全风险与日俱增。其中远程控制木马（以下简称：远控木马）作为一类高危害恶意软件，具有“被控端”和“控制端”两个组件，被植入到受害终端用于执行各类恶意操作的木马程序，称为“被控端”，被控端通常通过钓鱼邮件、恶意网站、软件捆绑等方式植入到受害主机，并通过进程注入、伪装等技术手段建立与控制端的隐蔽通信连接。与之对应的是“控制端”，是攻击者使用的程序，其主要功能在于与“被控端”建立通信，进而实现远程控制“被控端”。因远控木马具有窃取敏感信息、实施设备接管等攻击行为，且传播方式多种多样，并采多种技术手段来进行隐藏自身而难以被完全防范，对个人隐私保护、企业数据安全及关键信息基础设施构成了严重威胁。

如1-1所示，据《火绒安全2024年终端安全洞察报告》[2]显示，在2024年全网主动传播的恶意程序中，木马病毒占比达50.89%。

图1-1木马数量占比

中国网络空间安全协会联合360安全、深信服等公司发布的《2024年10月网络安全态势报告》[3]进一步显示：当月拦截恶意程序攻击达23.96亿次，其中远程控制木马以29.88%的占比位居榜首，远超挖矿程序（18.62%）、僵尸网络（15.37%）等其他攻击类型。

从上述报告中可知，远控木马仍是当前网络空间安全领域面临的主要威胁。当前远控木马主要采用基于TCP协议的反弹端口架构作为主要通信方式，本文选取反弹端口型远控木马作为主要分析对象。该类远控木马的典型行为模式表现为：被控端在受感染主机被执行后主动向控制端发起TCP连接请求，在成功建立通信连接后，获取主机信息回传至攻击者，为其后续渗透攻击提供情报支持。远控木马在被控端激活后的潜伏期内往往表现出低流量特征。因此，如何通过远控木马运行初期的少量流量及时且高精度地检测远控木马是当前远控木马检测研究中的难点问题，该问题的解决对远控木马的及时检测、避免受害主机遭到破坏具有重大意义。

2 相关工作

自1986年首个PC-Write木马出现之后，远控木马就由基本文件操作演变成带有多功能模块的威胁载体，发展到2025年，其功能不仅包含传统的文件管理、远程控制，更集成了挖矿等新型攻击。远控木马的发展也促进了检测方法的深入研究，本章根据远控木马在静态文件特征，动态行为以及通信流量与正常应用软件存在的差异，把远控木马检测方法分成基于网络的检测方法和基于主机的检测方法，其中，基于主机的检测方法通过分析程序代码特点（比如恶意字符串，异常API调用等）以及运行期间的可疑行为（比如敏感注册表项更改，异常进程创建等）实现木马程序识别。而基于网络流量的检测方法则通过分析远控木马与合法应用程序的通信流量行为特征差异或流量负载构建检测模型。

2.1基于主机的检测方法

基于主机的检测方法主要通过两类技术路径识别远控木马。一是针对木马程序源代码等静态文件展开逆向分析。二是创建虚拟执行环境，诸如虚拟机，云沙箱，以观察它们的运行时的行为特点，此类技术常常会收集并分析例如主机资源占有率、系统文件被篡改变更历史、进程注入操作、线程调用形式、注册表重要数值更改，特定API函数调用等数据。

Ahmadi等人[4]提出了基于灰度图像的可视化检测方法，通过将程序运行时的API调用序列转化为灰度图像，结合图像识别算法实现对木马程序的智能检测。肖茂等人[5]通过对部分PE文件使用彩色标签框来标记，进一步凸显了恶意软件图像区段的分布信息。韩潇宁等人[6]对木马程序进行反编译，在程序中提取了五个突出特征来识别远控木马。BehradFar等人[7]通过分析受害者主机中的静态DLL文件，动态API序列与文件特征，使用特征算法取得的最佳特征以检测识别远控木马与正常应用程序的区别。商海波等人[8]则采用行为特征分析法，监测应用程序的安装阶段、进程启动阶段及网络通信阶段的行为特征，包括系统文件夹创建、自启动服务注册、注册表项修改等关键操作，以此判别程序的恶意属性。Baek等人[9]提出了基于一种基于动静态相结合的方法以检测恶意软件，在进行静态分析后提取程序的操作码，使用双向长短期记忆模型学习，检测良性软件。在随后的动态分析中，使用虚拟环境对良性软件动态行为进行记录，通过监控系统行为变化来收集行为日志和进程内存信息，最终通过训练模型完成恶意软件检测。谷勇浩等人[10]对PE文件进行分析，从文件头部提取出静态特征，再使用沙箱环境提取动态特征，构建多重异质图，通过异质图对恶意软件进行网络关系的建模。然而，该类检测技术存在显著局限性。基于行为特征的检测方法需实时监控系统文件操作、注册表变更等底层系统行为，检测组件需深度集成于目标主机，不可避免地会占用一定系统资源，对主机性能产生负面影响。此外，随着远控木马技术的演进，部分高级变种已具备动态行为规避能力，能够通过随机化API调用序列、伪装正常进程等手段逃避基于行为特征的方法检测。

2.2 基于网络的检测方法

基于网络的检测方法可归为两类，其一为基于流量负载的检测方法，即基于流量数据包负载来识别攻击行为。其二为基于网络行为的检测方法，即基于通信行为分析来识别木马流量。

2.2.1 基于流量负载的检测方法

基于流量负载的检测方法通过检测流量负载来判定是否遭受远控木马攻击，例如基于恶意指令序列，异常协议字段等建立规则库进行匹配；或者基于深度学习的自主表征学习模式，采用卷积神经网络等从原始字节流中提取潜在特征。

Xie等人[11]运用CNN与长短期记忆神经网络（LSTM）从木马通信会话中同时提取时间信息与空间信息来检测HTTP木马流量；Jia等人[12]运用ResNet-LSTM自动提取原始流量中特征来检测木马流量。Chen等人[13]通过提取流量字节序列生成灰度图像提出了一种基于时间卷积网络的网络流量入侵检测方法；皮奔等人[14]通过提取流量字节序列生成Markov矩阵，使用轻量卷积神经网络模型自动提取正常流量与木马流量的潜在差异化特征，实现木马流量检测。基于流量负载规则库的检测方法具备检测速度较快、检测准确率较高的优势。然而，其缺点也较为明显，即检测过程中高度依赖特征的完善程度，对于特征库之外的远控木马，无法实现有效匹配与检测。而基于流量负载的深度学习检测方法在面对加密通信木马时存在一定局限性，木马采用的加密手段多样，且相关样本数量有限，导致检测准确率受到一定影响。

2.2.2 基于网络行为的检测方法

基于网络行为的远控木马检测方法的实现主要通过深入分析远控木马与正常应用流量的差异化通信行为特征构建检测模型进行检测，不涉及负载内容分析。

Pallaprolu等人[15]对会话中所有流量数据包提取特征向量，使用集成学习投票选取每一个分类器的辨别结果来实现更高准确率。Li S等人[16]通过对Ghost、Grat2010等远控木马进行分析，提取其在通信过程中产生的流量，并运用网络行为特征进行检测。李巍等人[17]通过对远控木马网络通信行为展开分析，将远控木马的通信过程划分为三个阶段，即连接建立阶段、命令控制阶段和连接保持阶段，分阶段提取统计特征，然后采用C4.5算法进行检测。宋紫华等人[18]提取远控木马建立连接后的五个流量数据包中提取特征，建立检测模型快速识别木马流量。Jiang等人[19]将满足从TCP三次握手时期的SYN数据包开始，相邻数据包之间时间间隔小于t秒这两个条件的数据包定义为早期阶段，从该阶段的数据包中提取特征来建立远控木马检测模型。总结看来，基于网络行为的检测方法通常需要采集较多的远控木马流量才能获取高的检测精度，因此如何及时高精度检测出远控木马流量仍然是值得研究的问题。

3远控木马概述

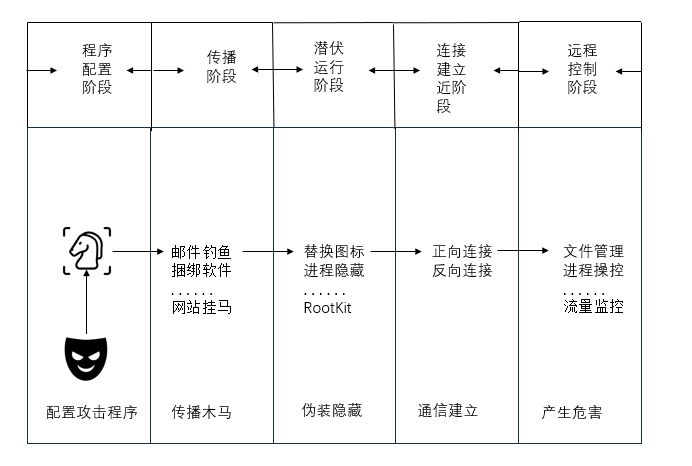
3.1 木马病毒概述

木马病毒是一种恶意软件，按传播方式、功能、攻击目标等来分类，有远控木马、键盘记录型木马、银行木马、下载型木马、破坏性型木马、勒索型木马、后门型木马、间谍软件、Rootkit木马等分类。远控木马因其执行隐蔽性，在受害主机中难以被发现。其被控端运行于被感染主机中，一旦建立通信连接，控制端可根据黑客意图随时发起攻击，进而实现窃取受害主机的隐私信息和重要文件，实施监控、资料修改、挖矿等恶意操作。

远控木马严重影响网络空间安全。由于其执行隐蔽性，攻击者侵入之后能够盗取用户的社交账号，银行卡账号及其密码，游戏账号，重要文件等机密资料，导致关键数据泄漏。且被控设备会被攻击者完全掌控，可实施例如非法开启摄像头和麦克风、修改文件系统、毁坏系统文件等极具危险性的行为。此外，攻击者借助被感染的设备创建僵尸网络，可对特定目标实施DDoS攻击。

3.2 远控木马攻击阶段

如图 3- 1所示，远控木马实施攻击主要包含以下五个关键阶段。

图 3-1远控木马攻击流程

（1）程序配置阶段

攻击者首先完成被控端程序构建，同步实施代码混淆、加壳保护、反调试机制等保护措施，完成通信协议配置及自启动模块等初始化操作。并可针对目标环境的安全防护措施进行适配性优化逃避查杀。

1. 传播阶段

被控端通过电子邮件钓鱼、捆绑软件、移动存储介质传播、恶意网站挂马及特定系统漏洞利用等途径植入受害主机。

（3）潜伏运行阶段

被控端触发执行后会隐藏自身，例如伪装成系统文件或者常见软件图标，利用Rootkit技术来隐藏自身进程与模块信息，借助DNS隧道、HTTP(S)代理等加密通信手段减少流量特征的暴露几率。

（4）连接建立阶段

远控木马主要存在两种连接模式，正向连接是由控制端主动访问被控端以建立通信，很可能受到防火墙限制。反向连接是由被控端主动向控制端发起连接，适合于NAT环境下的通信。

（5）远程控制阶段

攻击者使用控制端对被感染主机进行控制。实现基础功能包括文件的上传下载、文件查看修改、进程操控、挖矿、僵尸网络组建等攻击行为。随着远控木马的不断发展，部分远控木马集成了反调试、反查杀、反虚拟机等多种对抗机制，可根据检测环境动态调整攻击策略以逃避检测。

3.3 远控木马通信过程

由于大部分远控木马基于TCP协议通信，因此本文以基于TCP协议通信的木马作为研究对象。远控木马的连接过程通常可划分为三个阶段：连接建立阶段、命令交互阶段与保持存活阶段[20]。如图3-2所示。

在连接建立阶段，反弹端口型木马被控端成功入侵到目标主机并被触发执行后，会主动尝试连接控制端。一旦成功建立连接，被控端会自动向控制端反馈目标主机的上线信息。如操作系统类型，IP地址等。

在命令交互阶段，攻击者使用控制端下发相关攻击指令，被控端接收到指令后，就会执行对应攻击操作，一般的攻击指令涵盖远程桌面管理、文件管理、注册表管理、摄像头监控等。

在保持连接阶段，如果攻击者没有立即察觉到被控端上线，亦或是未针对被控端上线予以回应，则进入保持连接阶段。在此期间，远控木马的被控端和控制端会利用请求响应等多种机制来维持连接状态，由于交互数据包在并不涉及实质性的信息交互，其通信负载相对较少。

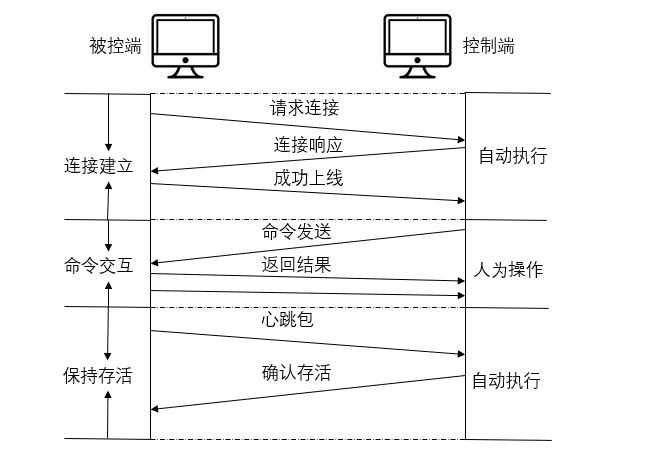


图 3- 2 远控木马通信流程

4 相关理论与技术

4.1 机器学习概述

机器学习是多学科交叉领域，广泛应用于图像识别、自然语言处理、推荐系统等领域。机器学习通过自主挖掘数据潜在规律，完成相应预测或决策任务。监督学习依赖于标注数据来构建分类或回归模型，而无监督学习则是更加侧重从未标注数据中发现聚类与潜在结构，强化学习通过环境交互与奖励机制优化策略。

机器学习存在三个基本要素，分别是数据、模型和算法，其中数据是机器学习的基础，是用于模型训练并进行优化改进的重要依据，与传统依靠经验的决策方式不同，机器学习采用数据驱动决策方式。模型通过使用数据来执行决策输出的假定函数，模型可以是计算型的，也可以是规则型的；算法是学习模型对于所需要处理数据的具体计算方法，通常会涉及到数学与最优化问题，其作用是从假设空间中挑选出最优模型从而得到最优结果。

4.2 主要类型

机器学习主要分为三种类型：监督学习、无监督学习、强化学习。

4.2.1监督学习

监督学习（Supervised Learning）作为机器学习领域的关键组成部分，使用带有标签的数据集对模型进行训练，旨在构建输入特征与目标标签之间的映射关系，实现对未知数据的预测或分类。其流程主要包含以下四个阶段：数据准备、模型训练、参数调优及预测应用。监督学习的常用算法既包括线性回归、逻辑回归、决策树、支持向量机等传统方法，也涵盖基于神经网络的深度学习模型。

4.2.2无监督学习

无监督学习（Unsupervised Learning）的关键核心在于从不带标签的数据中自动发现隐藏的映射规律，常常被用于聚类、降维、数据关联分析、生成建模等一系列任务。无监督学习算法的优势在于降低对带标签数据的依赖性，适用于标注成本高或标签定义模糊的场景。

4.2.3强化学习

强化学习（Reinforcement Learning）专注于动态决策分析，关注智能体如何在动态环境中采取一系列行动，以最大化累积奖励。智能体通过与环境进行交互，从环境中获得反馈，并据此不断调整自己的行为策略。

4.3 常见算法

4.3.1 随机森林

随机森林算法（Random Forest，RF）是一种基于集成学习（Ensemble Learning）的机器学习方法，其核心机制是通过构建多棵决策树，并汇总多棵决策树的预测结果，选择最优结果来提升模型的准确性。其核心机制包含双重随机性：第一重随机性是指通过样本数据的随机性为每棵决策树生成具有差异化的训练子集，即从原始数据集中有放回地随机抽取样本，以此来确保单棵树的学习基尼多样性；第二重随机性则是指在每一棵树的节点划分过程中引入特征随机性，仅从随机选取的特征子集中搜索最优划分条件，进一步降低树间相关性。

4.3.2 支持向量机

支持向量机算法（Support Vector Machine，SVM）是一种经典的监督学习算法，主要被用于分类和回归两大任务。其最关键的核心思想则是寻找一个能够最大化分类间隔的超平面来划分所有数据，这个超平面到各类样本的最小距离达到最大，而离超平面最近的样本点被称为支持向量。

4.3.3 k近邻

k近邻算法（K-Nearest Neighbor，KNN）是机器学习之中一种基于实例的监督学习算法，k近邻算法的关键核心逻辑是通过样本数据间的相似性进行预测判断。对于新数据点，算法计算出新数据点与训练集中所有数据样本的距离，并选择最近的k个邻居，根据选择出的这K个邻居的多数类别（分类任务）或平均值（回归任务）得出结果。该算法更适用于规模较小、维度更低的场景，例如简单分类或局部特征明显的模式识别任务。

5 实验及分析

5.1 实验环境

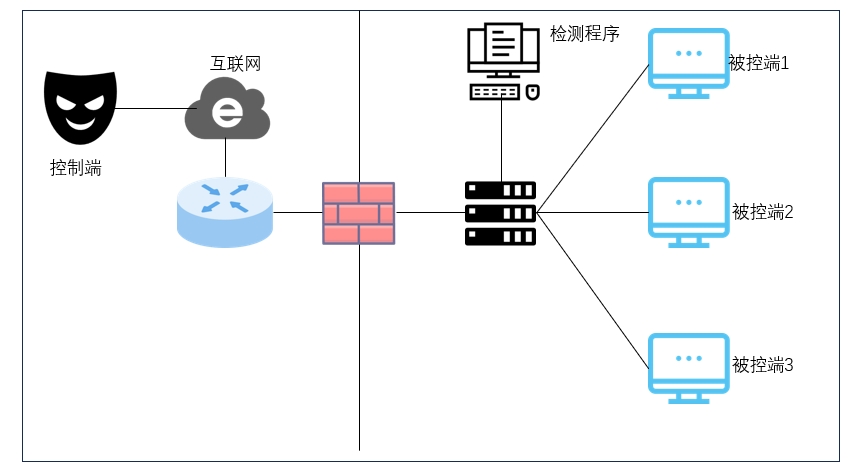
为保证实验安全进行，本文所有被控端均运行于windows 7 32位、windows 7 64位、windows 8、windows server 2019与windows 10专业版虚拟机中，对不同操作系统进行攻击。将检测程序搭建在具有16GB内存，AMD Ryzen 7 5700处理器的主机中。并使用具有IP公网的Windows操作系统云服务器作为控制端运行环境。通过Wireshark采集通信流量。利用Python3.7，Scikit-learn（sklearn）库实现机器学习检测模型构建。本文实验网络拓扑如图5-1所示。

图 5-1实验拓扑图

5.2 实验数据

本文实验所用木马样本如详见表5-1所示。样本序号1-35所产生的流量数据按照4:1的比例划分为训练集和验证集。远控木马样本序号36-44生成的流量单独提取出来作为测试集，以检验模型在未知数据上的检测效果。在流量采集阶段，按照图 5- 1所示的实验拓扑图，对每个远控木马运行时产生的流量进行了持续约5分钟的捕获，共捕获2982条远控木马流量。

表 5- 1 远控木马样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **木马名称** | **序号** | **木马名称** | **序号** | **木马名称** |
| 1 | BlackWorm | 16 | BlackNix | 31 | L6-RAT |
| 2 | Cloud Net | 17 | Proton | 32 | MLRAT |

续表 5- 1 远控木马样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 木马名称 | 序号 | 木马名称 | 序号 | 木马名称 |
| 3 | DRAT | 18 | Quasar | 33 | NanoCore |
| 4 | Revenge | 19 | Remcos | 34 | NovaLite |
| 5 | Slayer | 20 | Babylon | 35 | Orion |
| 6 | VanToM | 21 | Bozok | 36 | Paradox |
| 7 | MQ5 | 22 | BX | 37 | SpyNet |
| 8 | NjRAT | 23 | Comet | 38 | Pcshare |
| 9 | Offence | 24 | ctOs | 39 | Poison |
| 10 | OZNOE | 25 | CyberGate | 40 | Xtreme |
| 11 | SpyGate | 26 | Daleth | 41 | Mega |
| 12 | ucuL | 27 | DarkComet | 42 | Pentagon |
| 13 | Xena | 28 | Darktrack | 43 | VorteX |
| 14 | VorteX | 29 | Greame | 44 | NingaliNET |
| 15 | Bifrost | 30 | KilerRAT |  |  |

本文实验所用正常软件样本如表5-2所示。本文采集正常应用流量数据包含浏览器、电子邮件、视频软件、云服务器、游戏等范围。正常软件样本序号1-35运行时产生的流量分成4:1的训练和验证集。序号36-45运行时产生的流量作为测试集使用。捕获正常软件产生流量约5分钟时间，共捕获3363条流量。表 5- 3给出了实验中使用的流量详细数量。

表 5- 2良性应用样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 应用名称 | 序号 | 应用名称 | 序号 | 应用名称 |
| 1 | 360 | 16 | 爱奇艺 | 31 | 金山词霸 |
| 2 | 360浏览器 | 17 | 百度网盘 | 32 | 金山文档 |
| 3 | 360云盘 | 18 | 虎牙直播 | 33 | 酷狗音乐 |
| 4 | Cbox | 19 | 金山词霸 | 34 | 酷我音乐 |
| 5 | Chrome | 20 | 金山文档 | 35 | 鲁大师 |

续表 5- 2良性应用样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 应用名称 | 序号 | 应用名称 | 序号 | 应用名称 |
| 6 | Edge | 21 | 酷狗音乐 | 36 | 芒果TV |
| 7 | Opera | 22 | 酷我音乐 | 37 | 有道词典 |
| 8 | QQ拼音 | 23 | 鲁大师 | 38 | 微信 |
| 9 | QQ音乐 | 24 | 芒果TV | 39 | 喜马拉雅 |
| 10 | QQ游戏 | 25 | 美图秀秀 | 40 | 向日葵 |
| 11 | Tim | 26 | 搜狗浏览器 | 41 | 小鸟壁纸 |
| 12 | UC浏览器 | 27 | 搜狗输入法 | 42 | 印象笔记 |
| 13 | WPS | 28 | 搜狐影音 | 43 | 优酷 |
| 14 | YY语音 | 29 | 腾讯QQ | 44 | 有道云笔记 |
| 15 | 阿里邮箱 | 30 | 腾讯会议 | 45 | 讯飞输入法 |

表 5- 3 详细数据统计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 流量划分  流量类别 | 训练集 | 验证集 | 测试集 | 总计 |
| 良性流量 | 2410 | 602 | 351 | 3363 |
| 木马流量 | 2124 | 531 | 273 | 2982 |
| 总计 | 4534 | 1133 | 624 | 6291 |

5.3 评估标准

三个常见评价指标Accuracy、FNR、FPR被用于对本文实验效果进行评估，具体计算方法如式5-1、式5-2、式5-3所示。

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

以上指标通过TP、TN、FP、FN计算得到。其中TP表示模型预测为木马流量，实际也为木马流量的数量。TN表示模型预测为正常流量，实际也为正常流量的数量。FP表示模型预测为木马流量，实际为正常流量的数量。FN表示模型预测为正常流量，实际为木马流量的样本数量。FNR也称为漏检率，指的是实际为木马流量，被模型错误地预测为正常流量的比例。FPR又称误报率，是指实际为正常流量，被模型错误地预测为木马流量的比例。

5.4 特征分析

为实现对远控木马的快速检测，本文对远控木马样本在连接建立阶段的下行携带负载的数据包数量（以下简称“下行负载包”）进行统计分析。当连接建立阶段完成并进入命令控制阶段初期时的下行负载包数量约为5，下行负载包携带了攻击指令，因此下行负载包越小，主机被攻击程度就越小。因此本文以从远控木马发起连接到下行负载包数量为5时的这一阶段流量为分析对象，提取木马流量与正常流量的通信行为差异特征，实现远控木马在命令控制阶段初期流量的快速识别。

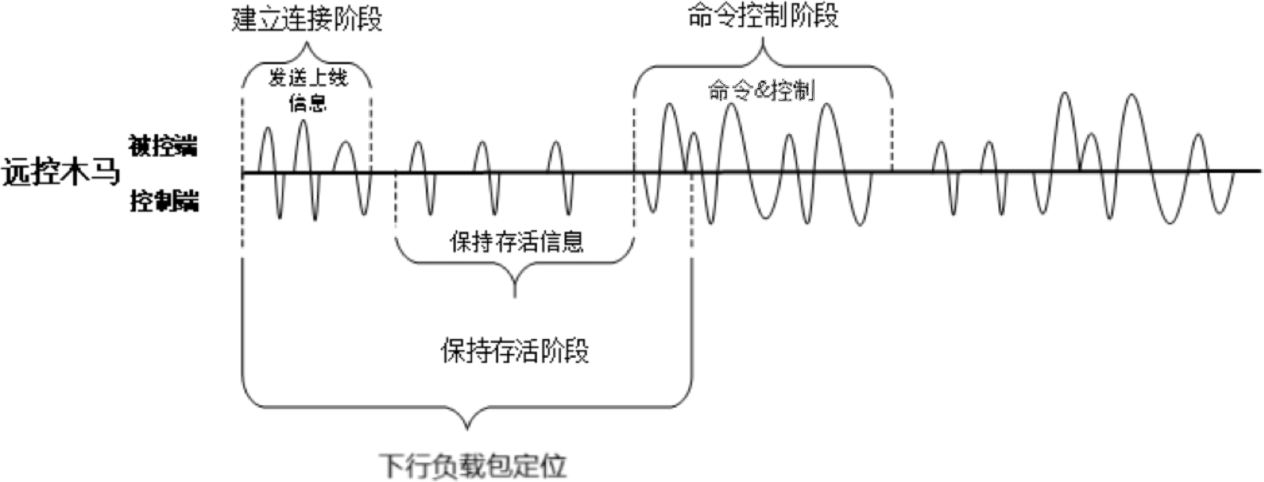
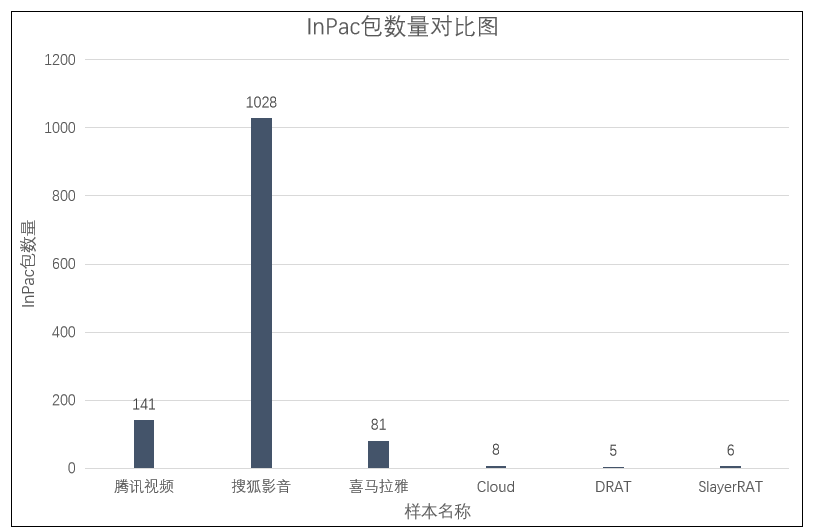


图5-2 下行负载包定位图

5.4.1下行数据包数量

远控木马在连接建立阶段由被控端主动发起连接请求，控制端仅需响应即可建立连接；命令交互阶段控制端向被控端下发简短命令，被控端执行后根据命令内容返回数据，此时下行流量主要是控制命令，往往较为精简；而保持存活阶段通过心跳包维持连接。远控木马运行时流量的行为特性使得控制端的下行流量主要集中在简短的命令传输，而大量数据上行（如敏感信息窃取）由被控端主动完成。而正常应用通常下行流量都相对较大，例如浏览视频、网页加载等操作。两者下行流量表现出明显区别，本文分别统计了三款远控木马与三款正常应用在本实验使用的不同操作系统环境中产生的流量下行数据包数量特征（InPac）平均值。如5-3所示，远控木马的行为特性致使其下行流量数据包数量往往低于正常应用。

图 5- 3 下行数据包数量特征统计图（单位：个）

5.4.2下行数据包负载量

在命令控制阶段，控制端向被控端发送攻击指令，大部分数据传输（如敏感信息窃取、屏幕截图、键盘记录等）均由被控端以较大负载的上行数据包完成，使得远控木马的下行数据包的负载较小。结合第5.4.1节分析可知，正常软件下行数据包负载包含了服务器响应的大量数据，例如浏览视频、网页加载等。因此两者在下行数据包负载量具有明显差异。本文分别统计了三款远控木马与三款正常应用在本实验使用的不同操作系统环境中产生的流量下行数据包负载量（InByte）平均值。如5-4所示，远控木马的行为特性致使其下行数据包负载量往往低于正常应用。

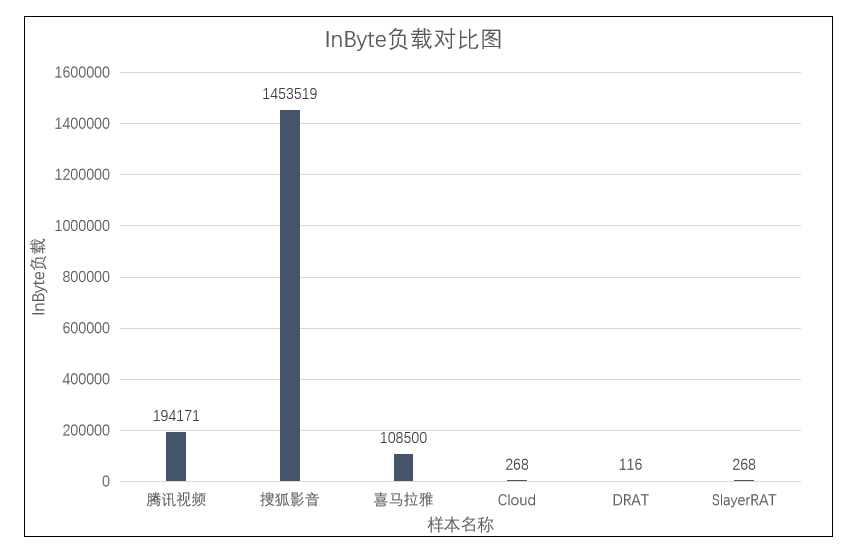


图 5- 4 下行数据包负载量特征统计图（单位：字节）

5.4.3最大通信时间间隔

基于远控木马的交互特性分析发现，其网络通信行为呈现出显著的非连续特征。由于远控木马需等待攻击者的人工指令介入或存在控制端响应延迟（如未及时捕获受害主机上线状态），其通信会话中常会出现数秒甚至数分钟的操作静默期，表现为有一段时间内无交互行为。相较而言，正常应用（视频流媒体、文件传输、即时通讯等）在执行功能时通常维持快速且稳定的数据交互。基于此，本文通过捕获通信会话内设定阈值之前的最大通信时间间隔作为判别特征。本文分别统计了三款远控木马与三款正常应用在本实验使用的不同操作系统环境中产生的流量最大通信时间间隔（max\_diff）。如图 5- 5所示，远控木马的行为特性致使其在通信初期最大通信时间间隔往往高于正常应用。

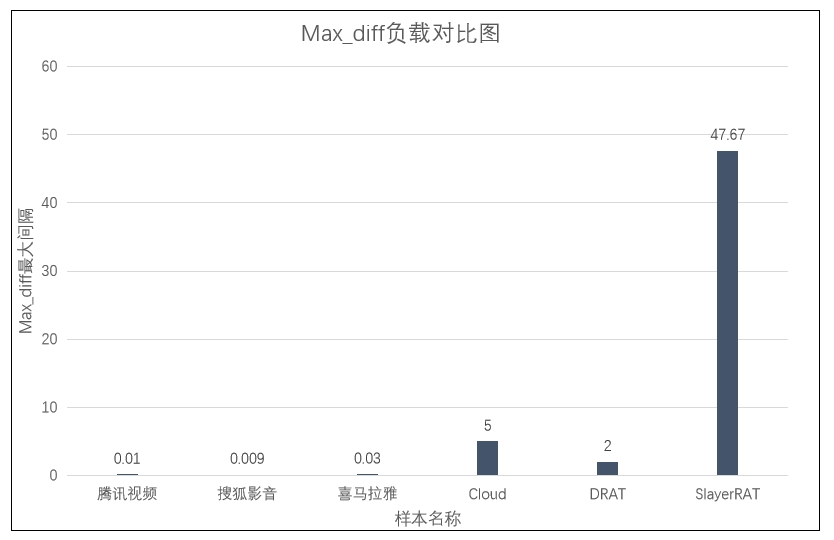


图 5- 5 最大通信时间间隔特征统计图（单位：秒）

5.4.4上下行数据包负载比

远程控制木马的通信模式在上下行负载分布上呈现明显的非对称性。木马上行流量通常承载大量窃取信息，如键盘记录、屏幕截取、文件窃取等；而下行流量以短指令为主（如心跳响应、命令触发等）。对于正常应用而言，其下行流量需持续传输高密度业务数据（如视频流、网页资源、下载文件等），而上行流量仅包含请求参数等。本文分别统计了三款远控木马与三款正常应用在本实验使用的不同操作系统环境中运行时产生的流量上下行数据包负载比（stdt）。如图 5- 6所示，远控木马的行为特性致使其上下行数据包负载比往往高于正常应用。

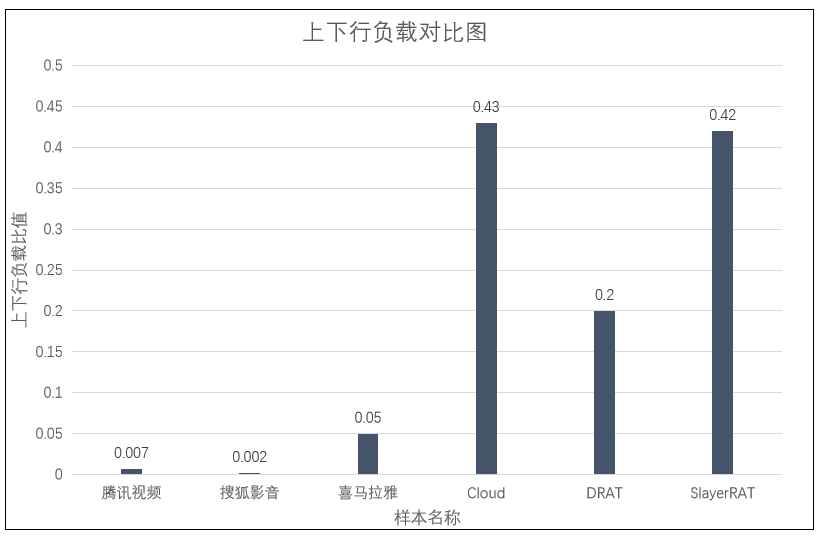


图 5- 6上下行数据包负载比

5.5 结果分析

5.5.1 实验结果

本文采用了5.3节中的评估指标对所提出的检测方法的有效性进行了验证。实验结果详见表5-4。在使用随机森林算法时，检测准确率达到了95.1%，同时漏检率（FNR）为9.5%，表明仅有9.5%的远控木马可能会逃脱检测。而误检率（FPR）为1.1%，表明仅有1.1%的正常网络流量会被误判为远控木马流量。这一结果表明本文所分析特征的有效性，能够快速准确地检测出远控木马流量，且影响正常网络活动极小。

表 5-4 实验结果

|  |  |
| --- | --- |
| **评估指标** | **指标值** |
| Accuracy | 0.951 |
| FNR | 0.095 |
| FPR | 0.011 |

5.5.2 对比实验

为进一步评估本文的检测精度，本文与文献[21]进行了对比实验。文献[21]从5分钟的通信流量中分析提取特征使用极端梯度提升（Extreme Gradient Boosting，XGBoost）算法分类，并优化了部分XGBoost参数。本文为测试在远控木马通信初期时的检测表现，测试了文献[21]中所提方法在下行负载包数量为5时的检测精度。如表5-5所示。

表5-5不同方法在测试集上的检测结果

|  |  |  |  |
| --- | --- | --- | --- |
| 评价指标  方法 | Accuracy | FNR | FPR |
| 文献[21] | 0.951 | 0.095 | 0.011 |
| 本文方法 | 0.925 | 0.137 | 0.025 |

如表5-5所示，文献[21]在下行负载包数量为5的实验条件下，检测准确率达到92.97%。但文献[21]的漏检率（FNR）和误检率（FPR）均高于本文所提出的检测方法。因为文献[21]聚焦于5分钟通信时间内的特征分析，在面对通信时间较短的情况下，所提取的特征差异性表现不够。而本文所提出的检测方法着重分析通信初期远控木马与正常软件之间的流量行为差异，能够有效地将远控木马流量从正常流量中区分出来。

6 总结与展望

6.1结论

随着互联网技术的快速发展和网络应用的全面普及，远控木马作为一种危害性极大的恶意软件，对网络安全构成了日益严峻的威胁。此类恶意程序可通过隐蔽手段渗透至目标计算机系统，在未经授权的情况下，借助互联网与攻击者控制端构建通信链路，进而实现对受感染系统的远程恶意操控。其通常被用于实施窃取用户敏感数据、破坏系统安全等违法犯罪行为，严重威胁个人隐私、企业数据安全及网络空间秩序。近年来，远控木马类型的攻击事件数量急剧上升，网络安全形势愈发严峻。本文通过对远控木马的通信行为的分析以及总结，发现由于远控木马的操控部分主要依赖人为决策所以可能存在攻击者响应延迟，远控木马在通信过程中会出现明显的操作停顿和流量交互中断现象。同时，远控木马在上下行流量数据包负载比方面也与正常软件存在较大区别，远控木马在其中表现为上行传输数据量大，下行主要为命令控制信息。综上所述，本文提取了最大通信时间间隔、上下行数据包负载比等关键特征，并且在下行负载包数量为5的情况下实现了对于远控木马95.1%的检测准确率。

6.2展望

随着计算机技术的不断革新，远控木马也在不断的进化。所以本文仍存在许多需要改进之处。一方面，尽管本文尽可能多的收集了多种类型的远控木马样本，但是远控木马种类繁杂且更新迭代十分迅速，有可能存在未知变种能够绕过当前的检测机制。另一方面，部分正常应用在运行过程中产生的网络流量与远控木马非常类似，不容易区别。在未来需要持续收集各类远控木马样本，深入学习相关知识，进一步完善检测模型，以提高检测的准确率和效率。

参考文献

1. 中国互联网络信息中心.(2025).中国互联网发展统计报告[年度报告].
2. Huorong Security. (2024). 火绒安全软件年度报告 [技术报告]. 火绒安全.
3. 中国网络空间安全协会. (2024). 2024网络安全年度报告 [年度报告].
4. Iman M R H, Ahmadi-Pour S, Drechsler R, et al. Processor Vulnerability Detection with the Aid of Assertions: RISC-V Case Study[C]//2024 IEEE Nordic Circuits and Systems Conference (NorCAS). IEEE, 2024: 1-7.
5. Xiao M, Guo C, Shen G, et al. Image-based malware classification using section distribution information[J]. Computers & Security, 2021, 110: 102420.
6. 韩潇宁.基于遗传算法的安卓恶意软件检测技术[D].大连理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.001597.
7. BehradFar M M, HaddadPajouh H, Dehghantanha A, et al. RAT hunter: Building robust models for detecting remote access trojans based on optimum hybrid features[J]. Handbook of Big Data Privacy, 2020: 371-383.
8. 商海波.木马的行为分析及新型反木马策略的研究[D].浙江工业大学,2006.
9. Seungyeon Baek, Jueun Jeon, Byeonghui Jeong, Young-Sik Jeong. Two-stage hybrid malware detection using deep learning[J]. Human-centric Computing and Information Sciences, 2021, 11(27): 10-22967.
10. 谷勇浩,王翼翡,刘威歆等.基于多重异质图的恶意软件相似性度量方法[J].软件学报,2023,34(07):3188-3205.DOI:10.13328/j.cnki.jos.006538.
11. Jiang Xie, Shuhao Li, Xiaochun Yun, Yongzheng Zhang, Peng Chang. HSTF-Model: An HTTP-based Trojan detection model via the Hierarchical Spatio-temporal Features of Traffics[J]. Computers & Security, 2020, 96: 101923.
12. Zijian Jia, Yepeng Yao, Qiuyun Wang, Xuren Wang, Baoxu Liu, Zhengwei Jiang. RAT Traffic Detection Based on Meta-learning[C]//International Conference on Computational Science. Springer,Cham,2021: 167-180.
13. Jinfu Chen, Shang Yin, Saihua Cai, Chi Zhang, Yemin Yin, Ling Zhou. An Efficient Network Intrusion Detection Model Based on Temporal Convolutional Networks[C]// 2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS). IEEE, 2021: 768-775.
14. 皮奔.基于卷积神经网络的远程控制木马检测方法研究[D].贵州大学,2023.DOI:10.27047/d.cnki.ggudu.2023.003600.
15. PALLAPROLU S C, NAMAYANJAJ M, JANEJAV P, et al. Label propagation in big data to detect remote access Trojans [C]//Proceedings of the 2016 IEEE International Conference on Big Data, Washington, Dec 5-8, 2016. Washington: IEEE Computer Society, 2016: 3539-3547.
16. Tang Y, Li S, Fang L, et al. Golden-chip-free hardware trojan detection through quiescent thermal maps[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019, 27(12): 2872-2883.
17. 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,(05):10-15.
18. 宋紫华,郭春,蒋朝惠.一种基于网络流量分析的快速木马检测方法[J].计算机与现代化, 2019(06):9-15.
19. Dan Jiang, Kazumasa Omote. An approach to detect Remote Access Trojan in the early stage of communication[C]//2015 IEEE 29th international conference on advanced information networking and applications. IEEE,2015: 706-713.
20. 王晨,郭春,申国伟,崔允贺. 利用序列分析的远控木马早期检测方法研究[J]. 计算机 科学与探索,2021.
21. Wei Jiang, Xianda Wu, Xiang Cui, Chaoge Liu. A Highly Efficient Remote Access Trojan Detection Method[J]. International journal of digital crime and forensics, 2019, 11(4):1-13.

致谢

至此本文即将结束之际，心中感慨颇多。首先在此诚挚的感谢我的指导老师皮奔。老师在学习以及此次论文的完成过程中给与了我巨大的支持与鼓励。在论文完成过程中遇到了许多问题，从一开始的论文题目选定到后来的开题报告，从艰难的实验分析到文章格式，老师都十分认真且及其耐心的对我进行指导，帮助我一一解决、克服这些困难。最终在老师的指导下完成了本文，在此再次诚挚的感谢我的指导老师。

在此我也要感谢给予我帮助的各位同学，在论文完成的过程之中他们提供了非常多的帮助、建议。他们十分热情的指出一系列可能存在的问题，并且不留余力的帮助我对这些问题进行分析与改进。在一次次与他们的深刻讨论中我收获了许多未曾想到的观点与角度令我受益匪浅。

**毕业论文（设计）原创性声明**

本人郑重声明：所呈交的本科毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：

日期： 年 月 日

**毕业论文（设计）使用授权声明**

本人完全了解遵义师范学院关于收集、保存、使用毕业论文（设计）的规定，即：

* 按照学校要求提交毕业论文（设计）的印刷本和电子版本；
* 学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；
* 允许论文被查阅和借阅；
* 学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。

作者签名： 指导教师签名：

日期： 年 月 日