**学号：2015200780**

**北京化工大学**

**硕士研究生学位论文开题报告**

**论文题目：网络流量分类技术的研究和应用**

**学院名称：信息科学与技术学院**

**专 业 ：计算机科学与技术**

**研究生姓名 ：杨文俊**

**导师姓名：赵英（教授）**

**开题日期：2016年9月26日**

|  |  |  |
| --- | --- | --- |
| **考核**  **成绩** |  | |
| **审核**  **小组**  **成员**  **以及**  **职称** | **姓 名** | **职 称** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

目录

[1 课题来源、项目名称 1](#_Toc399227274)

[2 文献综述部分 1](#_Toc399227275)

[2.1 本课题相关领域的历史、现状和前沿发展情况 1](#_Toc399227276)

[2.1.1 网络流量分类历史发展 1](#_Toc399227277)

[2.1.2 主要网络流量分类方法与评述 2](#_Toc399227278)

[2.1.3 目前存在的障碍和未来发展方向 4](#_Toc399227279)

[2.2 前人的研究成果 5](#_Toc399227280)

[2.2.1 流量分类的层次划分 5](#_Toc399227281)

[2.2.2 分类方法性能评估策略 5](#_Toc399227282)

[2.2.3 网络流量分类研究成果 6](#_Toc399227283)

[2.3 本课题的创新之处 7](#_Toc399227284)

[2.3.1 利用相关信息进行流量分类 7](#_Toc399227285)

[2.3.2 多种分类技术集成 7](#_Toc399227286)

[2.4 已查阅的文献目录 7](#_Toc399227287)

[3 研究计划部分 9](#_Toc399227288)

[3.1 论文选题的立论、目的和意义 9](#_Toc399227289)

[3.1.1 本文的立论 9](#_Toc399227290)

[3.1.2 本文的目的 9](#_Toc399227291)

[3.1.3 本文的意义 9](#_Toc399227292)

[3.2 本课题主要研究内容 9](#_Toc399227293)

[3.2.1 流的相关信息 9](#_Toc399227294)

[3.2.2 分类技术集成 10](#_Toc399227295)

[3.2.3 测试数据 10](#_Toc399227296)

[3.3 研究方案 11](#_Toc399227297)

[3.3.1 技术方案 11](#_Toc399227298)

[3.3.2 实施方案所需条件 11](#_Toc399227299)

[4 本课题难点 12](#_Toc399227300)

[5 预期的研究成果及创新点 12](#_Toc399227301)

[6 工作计划进度及经费预算 12](#_Toc399227302)

**开题报告的内容**

# **课题来源、项目名称**

项目名称：网络流量异常检测技术的研究和应用

# **文献综述部分**

## 本课题相关领域的历史、现状和前沿发展情况

### 网络流量异常检测历史

随着信息和通讯技术的快速发展，网络已成为人们日常生活中不可或缺的信息传播媒介。然而，通信网络本身的开放性以及其承载协议的开放性在为人们带来便利的同时，也给网络安全带来了极大的隐患。为保证网络的安全、高效运行，减少各种异常事件对通信网络正常业务的危害，构建可信的通信网络环境，就必须实时、准确地对通信网络的运行情况进行分析和监测，发现网络运行中存在的安全隐患，以便及时制定相应的网络策略对潜在问题进行响应。网络流量异常

检测作为确保网络安全的基础工作，已成为目前全世界学术界和工业界共同关注的前沿研究课题之一。 流量异常检测起源于 20 世纪 90 年代初期。1990 年，卡迈基梅隆大学的 R. A. Maxion 在文献[1]中首次提出对网络流量的异常检测。他将网络中的流量行为分为“正常”和“异常”两类：符合预期的网络流量行为定义为正常行为，而与“正常”网络流量行为偏离较大的行为定义为异常行为。基于正常行为和异常行为的定义，他提出了一种网络异常行为的诊断方法，通过分析网络流量数据的构成，构建流量正常行为的模型，从而对网络中流量异常行为进行实时检测。与入侵检测中常见的误用检测相比，流量异常检测的优势在于不需要对网络流量异常进行定义，并且能够检测出未知的流量异常行为。

网络流量异常检测具有两方面的意义：第一，流量异常检测能够发掘出可能对网络本身或其承载的数据业务造成危害的异常行为，从而使得网络管理员能够及时采取相应措施，降低这些异常行为所造成的危害，保证网络的安全运行；第二，流量异常检测能够发现网络中流量行为的新趋势，从而使得网络运营商能够根据流量行为的特点，调整网络规划，优化网络资源的配置，保证网络的高效运行。

近年来，随着互联网的迅猛发展，越来越多的新型网络应用逐渐兴起，网络规模不断扩大，网络组成也越来越复杂。作为增强网络可控性的基础技术之一，网络流量分类对帮助互联网服务提供商了解网络运行状态、优化网络运营与管理具有重要的意义。借助网络流量分类，网络管理者可以实时将网络中所有流量按不同应用类型进行划分与分析，为部署服务质量控制（QoS） 机制提供依据，并针对不同类型的应用提供不同的服务质量等级，从而避免减轻网络拥塞，确保关键业务服务质量，维持网络高效通畅运行。同时，依靠流量分类，网络服务提供商可以预测网络业务的发展趋势，合理的规划网络基础体系结构，使用户得到更好的上网体验。另外，在网络安全方面，流量分类是入侵检测系统（ intrusion detection system， IDS） 的核心部分，可发现网络中的突发流量（ 如蠕虫传播、大规模分布式拒绝服务攻击等） 与未知协议流量，从而及时采取防御遏制措施。[1][2]

网络流量分类在过去的二十年间主要分为三个历史阶段。[3]第一阶段，使用传输层（TCP、UDP）端口推断应用类型。该方法是基于RFC1340所提出的端口注册机制，即IANA对公用端口[0-1023]和注册端口[1024-49151]进行管理与分配，通过协议所用的端口可以快速判断应用类型。第二阶段，基于有效载荷（payload-based）的分类方法。第三阶段，基于机器学习及其它技术的分类方法。

下图展现了在流量分类历史上的一些主要事件，其中横轴表示时间，纵轴表示被IEEEXplore检索的文章数。

### 网络流量分类方法与评述

目前，对于网络流量进行分类的研究主要包括四类：基于端口号的分类方法、基于有效载荷的分类方法、基于主机行为的分类方法，以及基于机器学习的分类方法。其中，每一类方法又有其不同的实现方法。

#### 基于端口号的分类方法

传统的流分类方法依赖于对TCP或UDP数据包中端口号的分析，将熟知的端口（IANA指定）进行映射来识别不同的应用类型。位于网络中的分类器只需要找到一次TCP连接中的SYN包，并从这个SYN包中找到目的端口号即可。UDP也使用类似的方法（尽管不像TCP一样具有建立连接和连接状态维护的过程）。

这种方法的实现原理简单，适用于高速网络上的实时流量分类[4]。然而，它也面临着一系列问题： （1）大量的新应用没有IANA注册端口，而使用随机或用户定义的端口；（2）应用设计者或用户使用其它端口隐藏自身流量，规避过滤器和防火墙；（3）IPv4地址用完导致网络负载和端口地址复用。[5]如服务器使用同一IP地址的不同端口提供不同服务。正是这些不可规避的因素导致此种分类方法的精确度大大降低。

#### 基于有效载荷的流量分类方法

为了更加有效地避免基于端口的分类方法的不足，一种新的网络流量分类的方法被提出，这种方法通过分析包的有效载荷对网络流量类型进行识别，该方法也被称为“深层包检测（DPI）”。该方法具有较高的准确性[6]，并且被应用于一些商业软件产品和开源项目中，如部署在Linux核心防火墙。在此方法中，对数据包的有效载荷进行分析，以确定是否含有给定协议特有的已知模式、关键字和正则表达式。网站[7]给出了全面的已知模式的列表。此外，在进行系统入侵检测时，使用DPI方法识别网络异常是必要的预备步骤。

虽然该方法具有很高的分类准确率，但分析代价太大，因为要进行大量的包存取操作，并且在现代架构中存储器的读写速度一直是计算效率提高的瓶颈。此外，DPI方法的另一大缺点是关键字或模式通常要人工发掘，十分繁琐且准确率不高。而且十分重要的一点是这种方法无法应用于私有协议或加密流量，而且直接分析应用层的内容会带来隐私侵犯和安全性等问题。[5]

#### 基于主机行为的流量分类方法

为了弥补基于端口号和有效负载的流分类方法存在的缺陷，研究者提出一种基于主机行为的流分类方法，该方法通过分析主机在传输层的行为模式来进行流量分类，主要具有以下三个特点：（1）无需解读数据包的负载，因而不会涉及隐私侵犯的问题；（2）不需要知道与端口号相关的信息，因而不会被其误导；（3）只需要在路由器上就能够获取到的NetFlow信息，因而不需要额外的设备开销。正是由于这些特点，并且该方法可以和关联分析相结合，故可以应用于网络异常检测。[8]

虽然这种基于主机行为的流分类方法在一定程度上改善了基于端口和负载方法存在的问题，但其自身也存在一定的限制：（1）它无法识别一些特定应用的子类型，例如，它可以识别出P2P类型的流量，但却无法进一步识别是哪种P2P应用产生的流量；（2）该方法依赖于数据包首部中各个域之间的关系，因此当传输层首部被加密时，该方法无法使用；（3）当使用网络地址转换（NAT）时，只能通过服务器使用的不同端口号来区分，对分类准确率具有一定的影响。[9]

#### 基于机器学习的流量分类方法

为了克服上述方法的不足，近年来许多研究者开始利用机器学习方法解决流量分类问题。机器学习方法不依赖匹配协议端口或解析协议内容识别网络应用，而是利用流量在传输过程中表现出来的“网络流”（flow）的各种统计特征区别网络应用，方法本身不受动态端口、载荷加密甚至网络地址转换的影响。[10] 机器学习方法分为两大类，即基于有监督的算法和基于无监督的算法。

有监督机器学习分类是基于已标注类型的样本集进行机器学习并建立分类规则，将未知样本分类为已知的类型。有监督机器学习方法一般检测率高，但要求样本数据事先正确标记类别，无法对未知应用类型的流量进行分类；为生成具有良好的泛化性能的检测模型，往往需要利用大规模标注过的训练数据提高学习算法结果的准确度，但是标记必须由人手工完成。[11]

无监督机器学习方法根据流量统计特征的相似性进行聚合分簇，然后建立各个簇与类的映射关系。无监督机器学习具有能够自动发现新应用的特点，适用于网络流量不断变化的情形，如新应用的出现快于对应用特征的采集和识别。但是其检测精度与分类速度明显低于有监督的分类方法。[12]

基于机器学习的流量分类技术经过近年来的不断发展，已经取得了长足的进步。但基于机器学习的流量分类技术的实际应用仍受到很多问题的困扰，其主要挑战包括：（1）样本标注的效率瓶颈。（2）样本分布不均，实际网络流量中，web、mail、p2p 三类网络应用占绝大部分。（3）不能很好地进行实时分类。

### 目前存在的障碍和未来发展方向

#### 可用的数据和标准

对于流量分类来说最明显的问题就是网络研究中数据获得的问题。平衡个人隐私、数据安全、重要基础设施的保护和科学研究的关系对于法律、政策制定者和研究人员来说是一项长久的挑战。

目前有一些应对方案，但是并不能完全解决。（1）使用陈旧数据使隐私问题最小化。这样受限于样本集，对现实网络分类准确率一定不高。（2）把分类算法交给数据管理者进行测试，返回结果。数据管理者出于各种考虑，并不太愿意如此做。（3）对流量进行预处理，替换掉其中隐私信息。但是现有工具不能很好的识别、标记隐私信息。[13]

#### 网络的发展对流量分类的影响

并不存在完美的分类技术，并且随着网络的发展，原先的使用端口识别的方法准确性大大降低，目前仍有一些发展趋势使流量分类变得更加困难。（1）协议封装，如HTTP隧道技术。这使得分类器需要更为完整的负载检测或更为复杂的协议分析机制，才能准确识别流量。（2）流量加密，大大增加了特征值的提取的难度。（3）支持多种服务的应用，由于不同服务的评判标准不同（服务质量、安全政策），所以对此类流量不仅仅是识别流量所对应的应用，而且要识别应用所对应的服务（信号，视频流、聊天、数据转换）。

#### 可扩展性

网络流量分类的另一大挑战就是实时在线分析，由于在每一维度上网络基础设施大量增长（如带宽），并且要权衡准确性、性能和开销，这使得在线分析变得困难。为应对在线分析的复杂性，早期的研究均是对系统进行了一定的简化处理。如DPI分类器，不是减少了每个流负载的检测长度，就是简化了模式匹配方法。[14]一些研究人员表明：在特征个数受限情况下，每个流中包含4-5个包就能达到最大的准确度。[15] 在[16]中，作者分析了一个在线分类器中特征的计算复杂性，其中最大的复杂度是O(n\*log(n))，源端口、目的端口、初始窗口byte数的复杂度是O(1)，大多数复杂度是O(n)。

架构设计也会影响性能，随着并行计算的发展，对算法和系统进行多核并行运算优化也是未来研究方向之一。[17][18]

#### 分类技术集成

由于某些方法对于特定的类具有更好的分类性能，一个可以结合不同方法的系统相对于单一分类器就可能获得更好的准确性。这种思想已经广泛地应用于网络异常和入侵检测系统当中，但是流量分类领域鲜有涉及。可以对同一流量使用不同的分类方法，然后通过某种方法（随机选择、最大似然、Dempster–Shafer方法）综合各自结果。虽然这种结合不同方法的分类器增加了计算的复杂性，但是相对于单一分类器，它可以降低获取同等准确度分类所需的流的数量，从而可以减少分类时间。[19]

## 前人的研究成果

### 流量分类的层次划分

目前，对网络流量分类的研究很广泛，使用的方法也很多，但主要是基于以下三个层面的：

（1）Packet-level的流量分类：主要关注数据包（packet）的特征及其到达过程，如数据包大小分布、数据包到达时间间隔的分布等。

（2）Flow-level的流量分类：主要关注流（flow）的特征及其到达过程，可以为一个TCP连接或者一个UDP流。其中，流通常指一个由源IP地址、源端口、目的IP地址、目的端口、应用协议组成的五元组。

（3）Stream-level的流量分类：主要关注主机对及它们之间的应用流量，通常指一个由源IP地址、目的IP地址、应用协议组成的三元组，适用于在一个更粗粒度上研究骨干网的长期流量统计特性。

在上述三个层面的流量分类中，使用最广泛的是Flow-level的流量分类。这种以流为单位分析网络中传输数据的方法，是分组交换网络发展的必然需求。[1]

### 分类方法性能评估策略

模型评估是指评价分类模型在未知样本集上处理分类问题的能力，其关键指标是对未知样本的预测准确率。通常用于衡量分类准确率的评估标准，主要包括以下三个方面：

TP（true positive）：类型i中的样本被分类模型正确预测的样本数，记为TPi。

FN（false negative）：类型i中的样本被分类模型预测为其它类型的样本数，记为FNi。

FP（false positive）：不属于类型i的样本被分类模型预测为类型i 的样本数，记为FPi 。

基于以上概念，下面给出评价分类模型准确性的4个常用指标: 类准确率（ recall） 、类可信度（ precision）、整体准确率（ overall recall）以及综合评价指标（F-mesure）的描述，计算方法如公式（1）～（ 4） 所示：



（1）



（2）



（3）



（4）

在这4个评测指标中，分类模型的整体准确率和综合评价指标应用最广。整体准确率它反映了分类模型正确预测样本数占总样本数的比例，这个指标用于测量分类器在全部样本数据上的准确性。综合评价指标用于评估分类器对每个具体类的准确性。[10]

### 网络流量分类研究成果

在过去的20多年中，随着网络规模的急剧增大和网络应用的爆发式增长，不断有一些新的挑战出现。在此过程中网络流量分类也逐步发展，并取得了长足的进步。

科研人员对此进行了大量的工作，如 [20]对不同端点之间的连接图进行行为分析，表明P2P和C/S应用有着不同行为模式和连接图，并且证明此种方法可应用于主干网络。[21]利用启发式算法提高非监督流量聚类的性能，使用BoF（Bags of flows）作为相关信息的模型，大大提高了分类准确度。[22]使用网络流量相关信息进行流量分类，解决了小样本条件下分类准确率不高问题，使用相关信息对流量进行一次聚合，使具有很强相关性的流量同属于一个BoF中，并且在数学上论证了此方法的有效性，该方法应用于具有IP负载的wide和isp数据集上，只关注TCP流量分类。[23]引入C4.5 决策树方法来处理流量分类问题． 该方法利用训练数据集中的信息熵来构建分类模型，并通过对分类模型的简单查找来完成未知网络流样本的分类。与NB 方法不同，C4.5 决策树不依赖于网络流样本分布的先验概率，因此在网络流样本分布变化时依然具有较好的分类准确率; 具有较快的流量分类速度，在对待测网络流样本进行分类时，仅需进行特征值比较，计算量小，在处理大规模流量分类问题时具有明显的性能优势。[24] 引入有监督的朴素贝叶斯( Naive bayes，NB) 机器学习方法进行流量分类与应用识别。为解决样本特征并不全部满足正态分布的问题，Moore 等人采用特征选择方法对特征集合进行过滤，并使用核密度估计对朴素贝叶斯方法进行了改进，分类准确率得到提高达到95%以上。[25] 提出一种基于支持向量机( support vector machine，SVM) 的流量分类方法。该方法利用非线性变换和结构风险最小化( structural risk minimization，SRM) 原则将流量分类问题转化为二次寻优问题，具有良好的分类准确率和稳定性。

## 本课题的创新之处

### 利用相关信息进行流量分类

本课题的创新之处在于发现网络流量之间的相关信息，并且利用它进行分类。通常传统的分类方法通常把流（flows）视为独立的实例，并不考虑它们之间的相关信息。[22]使用一个三元组（目的IP，目的端口、协议）表示流之间的信息，并且通过实验表明通过此种方法在训练样本非常小的情况下可以显著提高分类性能。

探寻流之间的其它相关信息，并验证其正确性是本课题下一步所需要进行的工作。

### 多种分类技术集成

对于模式分类,往往存在这样一种现象,不同的分类方法常常得到不同的分类效果，目前还不存在一种方法能够对所有的分类问题都有良好的分类表现。 [26] 大量的研究成果表明，集成多个分类器的分类学习性能,形成一个综合系统，能显著的改善系统的整体性能，并在实际应用发挥着越来越重要的作用。

本课题下一步的工作就是寻找一种性能优良的结合方法和分类结果综合方法，使分类器具有更高的准确性和更强的健壮性。

## 已查阅的文献目录

[1] Roy A. Maxion. Anomaly detection for diagnosis[C]. in Proc. of International Symposium on

Fault-Tolerant Computing, Newcastle Upon Tyne, UK, 1990, 20-27

[1] 熊刚, 孟姣, 曹自刚，方滨兴. 网络流量分类研究进展与展望[J]. 集成技术, 2012, (1).

[2] 邓河. 基于机器学习方法的网络流量分类研究[D]. 湖南工业大学, 2009.

[3] Dainotti A, Pescape A, Claffy K C. Issues and future directions in traffic classification[J]. Network, IEEE, 2012, 26(1): 35-40.

[4] 彭芸,刘琼.Internet流分类方法的比较研究 [J].计算机科学,2007,34（8）: 58-61.

[5] Valenti, S., Rossi, D., Dainotti, A., Pescapè, A., Finamore, A., Mellia, M. Reviewing traffic classification[M]//Data Traffic Monitoring and Analysis. Springer Berlin Heidelberg, 2013: 123-147.

[6] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[M]//Passive and Active Network Measurement. Springer Berlin Heidelberg, 2005: 41-54.

[7] l7filter, Application layer packet classifier for Linux. http://l7-filter.clearfoundation.com/

[8] 周颖杰. 基于行为分析的通信网络流量异常检测与关联分析[D]. 电子科技大学, 2013.

[9] 彭雄威. 基于行为的网络流量分类方法研究[D]. 华中科技大学,2013.

[10] Kim, H., Claffy, K. C., Fomenkov, M., Barman, D., Faloutsos, M., & Lee, K. Internet traffic classification demystified: myths, caveats, and the best practices[C]//Proceedings of the 2008 ACM CoNEXT conference. ACM, 2008: 11.

[11] 胡婷, 王勇, 陶晓玲. 网络流量分类方法的比较研究[J]. 桂林电子科技大学学报, 2010, (3).

[12] 何震凯. 基于聚类分析的网络流量分类研究[D]. 湖南工业大学, 2009.

[13] Salgarelli L, Gringoli F, Karagiannis T. Comparing traffic classifiers[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(3): 65-68.

[14] 柏骏, 夏靖波, 吴吉祥, 任高明, 赵小欢. 实时网络流量分类研究综述[J]. 计算机科学 ISTIC PKU, 2013, 40(9).

[15] Bernaille L, Teixeira R, Salamatian K. Early application identification[C]//Proceedings of the 2006 ACM CoNEXT conference. ACM, 2006: 6.

[16] Li, W., Canini, M., Moore, A. W., & Bolla, R. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks, 2009, 53(6): 790-809.

[17] 任春梅. 网络流量分析关键技术研究[D]. 电子科技大学, 2013.

[18] 刘颖秋, 李巍, 李云春. 网络流量分类与应用识别的研究[J]. 计算机应用研究 ISTIC PKU, 2008, 25(5).

[19] Callado, A., Kelner, J., Sadok, D., Alberto Kamienski, C., & Fernandes, S.  Better network traffic identification through the independent combination of techniques[J]. Journal of Network and Computer Applications, 2010, 33(4): 433-446.

[20] Jin, Y., Duffield, N., Haffner, P., Sen, S., & Zhang, Z. L. Inferring applications at the network layer using collective traffic statistics[C]//Teletraffic Congress (ITC), 2010 22nd International. IEEE, 2010: 1-8.

[21] Wang, Y., Xiang, Y., Zhang, J., & Yu, S. A novel semi-supervised approach for network traffic clustering[C]//Network and System Security (NSS), 2011 5th International Conference on. IEEE, 2011: 169-175.

[22] Zhang, J., Xiang, Y., Wang, Y., Zhou, W., Xiang, Y., & Guan, Y. Network traffic classification using correlation information[J]. Parallel and Distributed Systems, IEEE Transactions on, 2013, 24(1): 104-117.

[23] 徐鹏, 林森. 基于C4．5决策树的流量分类方法[J]. 软件学报, 2009, (10):2692-2704.

[24] Moore A，Zuev D. Internet traffic classification using Bayesian analysis techniques［C］． In ACM International Conference on Measurement and Modeling of Computer Systems ( SIGMETRICS) 2005，Banff，Alberta，Canada，June， 2005．

[25] 徐鹏, 刘琼, 林森. 基于支持向量机的Internet流量分类研究[J]. 计算机研究与发展, 2009, (3).

[26] 黄东山. 特征选择及半监督分类方法研究[D]. 华中科技大学, 2011.

# 研究计划部分

## 论文选题的立论、目的和意义

### 本文的立论

近年来Internet规模呈现爆炸式膨胀，网络用户规模在一段时期内呈指数增长，网络的传输容量和速度以及业务类型得到快速增长，使得 Internet成为一个综合性服务平台。快速发展的网络对网络性能分析和管理带来挑战。

早在二十世纪末互联网就已经实现网络全球覆盖，其体系结构异常庞大，网络的动态性、异构性和复杂性使得网络运行机制和行为错综复杂，难以实施网络的全面控制和管理。为了更好理解网络内部运行机制以及关联因素，为网络规划与设计提供支持，保障网络QoS，提升用户体验度，研究人员从全方面多角度对Internet 开展了研究工作。网络流量作为整个网络复杂性动态性的载体，同时网络流量反应了网络的现状和发展趋势，因此网络流量就是研究分析网络体系架构、实施网络管理、为网络安全提供保障的切入点，开展网络流量分析是理解网络动态行为用以建模和性能分析的重要途径。网络流量是网络性能研究的重要基础，所有的网络行为特点都可以通过网络流量分析来获得。

### 本文的目的

本文的目的是设计一个网络流量分类系统，该系统集成不同的分类算法，各算法分别对网络流量进行分类，得到各自结果，随后系统根据各分类算法产生的结果进行综合，得到最终的分类结果。该系统所用的算法在分类时使用流的相关信息提高在小样本条件下分类的准确度。

### 本文的意义

网络流量分类具有以下重要意义：（1）为网络规划提供技术支持。（2）为网络的安全提供可靠依据。（3）为网络综合管理提供实践依据。（4）是新型网络应用与网络协议研究的需求。

## 本课题主要研究内容

### 流的相关信息

（1）流的相关信息的发现

在网络流量分类研究中，通常把具有相同五元组（源IP地址、源端口、目的IP地址、目的端口、应用协议）的包的集合称为流（flow）。在以往的分类研究中，流只被视为独立的实例，它们之间的相关性往往被研究人员所忽视。

（2）使用流的相关信息的算法

寻找适当算法利用所发现的流的相关信息，例如朴素贝叶斯、最近邻居算法均可利用流的相关信息增加分类的准确性。

### 分类技术集成

（1）现有算法的集成

网络流量分类算法多种多样，如何选择相应的算法进行集成使系统在不同应用场景都具有良好的分类效果是一个亟待解决的问题。

（2）不同的分类算法结果综合

对于同一流量，不同的分类算法会产生不同的分类结果，如何对其产生的结果进行综合也是需要考虑的。

（3）计算、存储开销

由于使用多种算法对流量进行分类，所以计算、存储开销明显大于使用单一算法的分类器，那么对于计算开销、算法的选择和性能需求之间的平衡显得尤为重要。

### 测试数据

（1）特征的选择

目前很多特征选择的方法在选取特征时只考虑对单个特征的类别识别能力进行排序，然后选取得分最高的前k个特征。但由前k个强分类特征组成的特征子集并不一定具有很强的类别区分能力，因此这类特征选择方法不能直接从大量的特征中选出具有强类别识别能力的特征子集。在面对具体的高维特征选择问题时，现有的特征子集选择方法中，还不存在一种方法能够完全解决这类问题。特征选择方法通常为三类：包装法（wrapper）、过滤法（Filter）及嵌入法（Embedded）。

本课题将在综合考虑计算开销和分类精度的前提下，选取适当的特征提取方法。

（2）数据标记

对于数据分析问题,获取有用的监督信息（或称人工标记信息，有时也指类别标记）需要做大量繁琐的工作，耗费大量的时间、人力。

针对本课题具体内容，寻找到一种快速的机器标注方法，也是下个阶段所需要探索的。

## 研究方案

### 技术方案

#### 研究数据的获取

本课题对于数据的需求分为两个阶段：

第一阶段，研究主要集中在对流的相关信息的发掘和相应算法的分析、应用，此阶段只需要小规模数据对设想进行初步验证。故本阶段只需要使用Wireshark对本实验室的流量进行捕获。Wireshark可以捕获原始数据包，包括发送到正在运行的主机上的数据包和在其它主机在共享媒介上交换的数据包，完全满足此阶段对数据的要求。

第二阶段，研究主要集中在对假设方案的大规模验证上，对于通过第一阶段的假设，使用根大规模，更具有代表性的数据进行验证。此阶段的数据可以通过CAIDA（Center for Applied Internet Data Analysis）获得，CAIDA是最重要的互联网数据提供者之一，提供多种网络流量数据，可用于此阶段使用。

#### 软件框架设计方案

* 采用面向对象的软件设计方法，确定优化软件中的对象并设计所有公共接口
* 使用C++语言编写所有接口的代码

#### 算法设计测试方案

* 设计使流的相关信息的算法
* 使用C++实现算法
* 把不同算法组合为一个系统

### 实施方案所需条件

* 分好类的流量数据集
* 一台高性能计算机

# 本课题难点

* 流的相关信息发掘
* 不同算法结果的综合方法的选择
* 特征提取
* 系统的计算开销
* 样本数据集合的标注

# 预期的研究成果及创新点

* 发现新的流的相关信息
* 实现一个集成多种分类算法的系统，综合不同算法得到的结果，使该系统可适用于绝大多数网络环境，具有普遍性。

# 工作计划进度及经费预算

i. 2017年9月至2017年12月，进一步了解网络流量异常检测算法及算法设计。

ii. 2018年1月至2018年3月，设计算法、设计软件框架。

iii. 2018年4月至2018年7月，实现软件、模拟测试。

iv. 2018年8月至2018年12月，实际数据测试、改进软件功能，撰写。

v. 2019年1月至2019年6月，完成硕士学位论文。

|  |
| --- |
| **指导教师意见：**  **指导教师签名：**  **年 月 日** |
| **审核小组意见：**  **审核小组组长签字：**  **年 月 日** |
| **研究生根据审核小组意见对开题报告的改进措施：**  **年 月 日** |
| **备注：** |