学号：2016200806

北京化工大学

硕士研究生学位论文开题报告

论文题目：基于 Spark 的网络流量分类研究和应用

学 院 名 称 ：信息科学与技术学院

专　　业 ：软件工程

研究生姓名 ： 刘兆禄

导 师 姓 名 ： 赵英（教授）

开 题 日 期 ： 2017 年 9 月 29 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

课题来源：自选题目

项目名称：基于 Spark 的网络流量分类研究和应用

2 文献综述部分

2.1 本课题相关领域的历史、现状和前沿发展情况

2.1.1 网络流量分类历史发展

近年来，随着“互联网+”、云平台、大数据等新技术高速发展，互联网现在已经成为经济发展和社会进步的不可或缺的推动力量。与此同时，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[1]。网络带给人们便利的同时，也给网络安全带来了极大的隐患。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[2]。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，它可以帮助了解网络的行为（NetWork Behavior，NB），为部署服务质量控制（Quality of Service，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，保障网络安全（Network Security）[3]。不仅如此，网络流量分类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[4]。

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，IDS 一般先构建正常的网络流量分类模型，再实时监控网络中的流量，通过偏差来判断新流量是否为异常流量。异常流量检测是一种通过保护系统信息收集和分析，从而发现异常的技术。发现和识别网络流量中的异常流量，给出异常流量警报[5]，而及时采取防御遏制措施。

网络流量特性的分析和研究在 80 世纪年代主要借鉴公众交换电话网络的泊松（Poison）模型。90 年代伴随数据网络和应用的出现，流量特性已经发生了显著变化，泊松过程已不能充分反映网络流量的高可变性[3]。随着研究深入，各种流量分类方法和模型相继出现。

网络流量分类发展到现在已有二十多年，技术也在逐步成熟，在过去的技术发展过程中，主要分为三个阶段[6]。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，但随着动态端口技术的出现，该方法分类的可信度大打折扣。第二阶段，基于有效载荷（payload-based）的分类方法，但目前由于涉及用户隐私的关系，大多数流量对其载荷加密。使得基于有效载荷的分类方法逐步退出了流量分类的舞台。第

三阶段，基于机器学习及其它技术的分类方法。由于机器学习的方法具有分类准确、快速、识别率高等特点[7]。因此受到广大研究人员的青睐，也是目前研究流量分类的主要方法。

图 1 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。

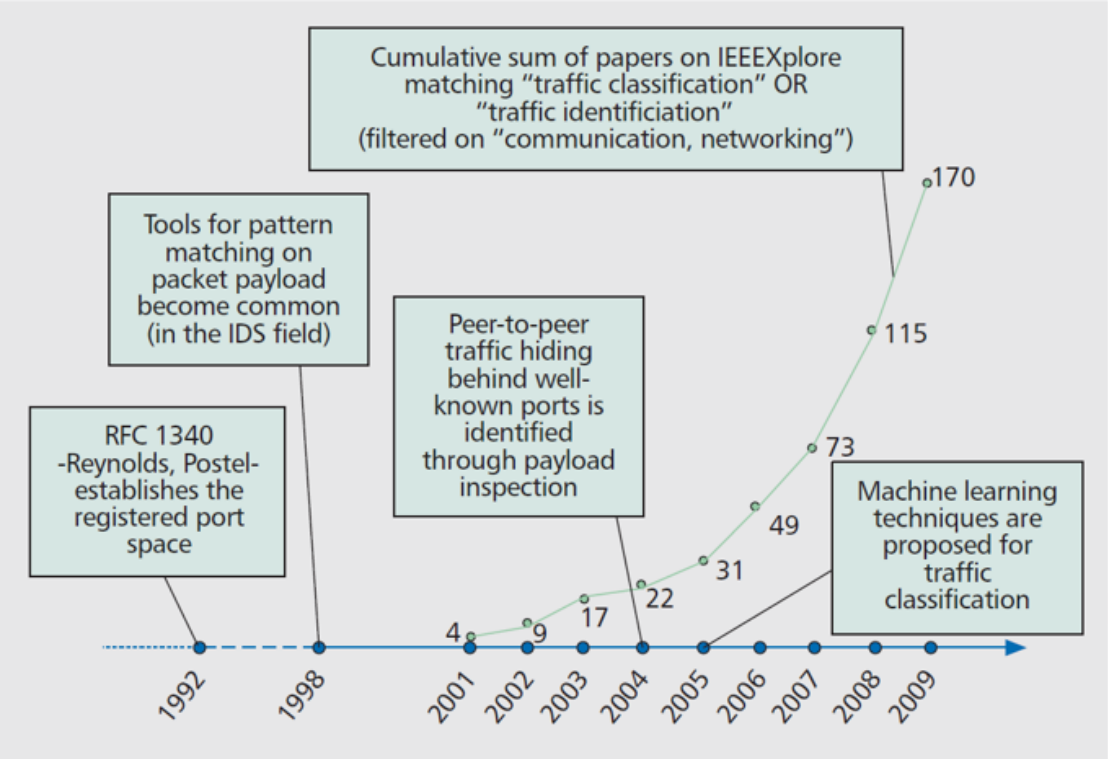

图 1 IEEEXplore 检索流量分类文章数

2.1.2 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[8]。以下对这四种方法分别进行简述。

2.1.2.1 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[9]。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[10]。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。

2.1.2.2 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[11]，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[12]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]。

2.1.2.3 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[13]。

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]。

2.1.2.4 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[14]。机器学习方法主要分为两大类，即有监督和无监督的算法。

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[8]。

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[14]。

但最近又涌现出一种半监督的机器学习方法，是有监督学习和无监督学习相结合的一种学习方法，半监督学习使用大量的未标记数据，以及同时使用标记数据，来进行模式识别工作。当使用半监督学习时，将会要求尽量少的人员来从事工作，同时，又能够带来比较高的准确性。[15]

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[9]。（2）不能很好地进行实时分类。

2.1.3 Apache Spark 技术发展历史

网络发展至今，越来越多的数据涌向网络，当今开源大数据越来越多，信息化越来越普遍，是一个大数据爆发的时代。为了解决大数据分布式计算问题，Spark 应运而生。Spark 源自加州大学伯克利分校的 AMPLab，现已捐献给 Apache 软件基金会。该基金会已经将其作为顶级开源项目[16]。由此对于数据科学家和未来想从事此类职业的人员来说，Spark 已经被他们所重点关注，同时也带来了更多的 Spark 开源项目。

Spark 是新一代开源大数据处理平台，是一种集流计算、数据查询、机器学习、和图挖掘于一体的通用计算框架[17]。Spark 技术是基于 Hadoop 的 MapReduce 技术演变而来，但是 Spark 却比 Hadoop 更加优秀，它的迭代计算能力要比 Hadoop 的快上 100 倍。因此在机器学习方面有着无与伦比的优势。（见图 1）

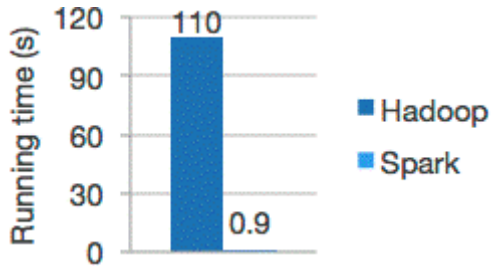

图 1 Spark 与 Hadoop 效率比较

2015 年是 Spark 技术快速发展的一年，在 Spark 东部峰会上，Databricks 公司发布了一系列相关标准，目的是为了指引内存数据处理框架 Spark 技术在今后的发展方向。Databricks 公司一直以来作为 Spark 技术核心的研发团队，采用不同领域的技术来推动 Spark 版本的改进。Spark2.0 版本在 2016 年春季发布，在 Hadoop 的生态下，Hadoop1.x 支持 Spark 早期版本，并且作为一种替代 Hadoop 计算框架而存在[16][17]。HDFS 文件系统是 Hadoop 支持 Spark 分布式计算的核心组件，预计未来 10 年，Spark 技术作为重要的大数据处理引擎技术，将衍生出各种大数据应用项目。

2.1.3.1 Apache Spark 技术简介

Spark由 UC Berkeley AMP Lab 开发。它拥有 Hadoop 所具有的优点，不同之处在于 Job 的中间结果可以保存在内存中，不再需要读写 HDFS，因此 Spark 能更好地适用于机器学习等需要迭代MapReduce 算法，与此同时Spark 拥有出色的容错和调度机制，能够确保系统的稳定运行。

Spark的一大特点是它的弹性分布式数据集（RDD）[18] 弹性分布式数据集（RDD）抽象使开发人员将流水处理线上的任何点物化在跨越集群节点的内存中。这样后续步骤如果需要相同数据集时就不必重新计算或从磁盘加载。这个特性使Spark可以应用于以前分布式处理引擎无法胜任的应用场景中。

Spark的另一大特色就是Spark的生态圈[19]，也称为BDAS（伯克利数据分析栈），是伯克利APMLab实验室打造的，力图在算法（Algorithms）、机器（Machines）、人（People）之间通过大规模集成来展现大数据应用的一个平台。Spark生态圈以Spark Core为核心，从HDFS、Amazon S3和HBase等持久层读取数据，以MESS、YARN和自身携带的Standalone为资源管理器调度Job完成Spark应用程序的计算。 这些应用程序可以来自于不同的组件，如Spark Shell/Spark Submit的批处理、Spark Streaming的实时处理应用、Spark SQL的即时查询、BlinkDB的权衡查询、MLlib/MLbase的机器学习、GraphX的图处理和SparkR的数学计算等等[19]。（图2为Spark生态圈模型）

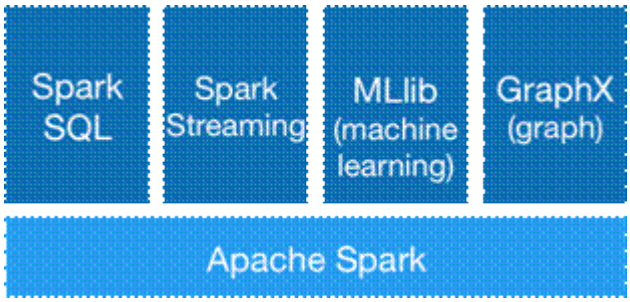

图2. Spark生态圈模型

2.1.4 目前存在的障碍和未来发展方向

2.1.4.1 数据的采集和标准

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。首当其冲的就是要平衡数据个人隐私、数据安全等问题，因为网络流量的分析涉及用户隐私的问题，如何权衡法律与科研之间的利弊，是一项亟待法律、政策制定者和研究人员商榷解决的问题[20]。目前而言，大多数运营商采取了对流量加密的措施来保护用户的隐私，但对于科学研究者来说，这无疑增大了数据分析的难度。

其次高速网络不仅带来了快速大量存储的问题，也带来了大容量分布式处理问题[21]。在当今的高速网络环境下，网络信息的数据量也呈现爆炸式增长，并且呈现出4V特性（量大、多样性、速度快、价值密度低）[1]如何在高速网络中采集并转储大量网络流量数据，是网络流量研究者需要解决的问题。

2.1.4.2 网络的发展对流量分类的影响

随着技术发展，新的流量技术层出不穷。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[22][23]。如（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）。（4）动态端口技术，越来越多的程序应用采用了动态端口技术，使得有可能存在同一应用的网络流具有不同的端口特征。

2.1.4.3 网络流量的高速实时性

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[24]。现实网络具有网络规模大、链路速率高、流量并发性强等特点。目前，我国骨干网的传输速度达到了 OC192（10Gbps），部分经扩容以后，带宽容量已经升级到了 OC768（40Gbps）[25]并且流量内容日趋复杂，各种移动穿戴设备也开始接入网络，平板电脑、手机和台式电脑等各种上网设备也成为网络流量的主力军[21]。

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[26][27]。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[28]。在[29]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)。

2.1.4.4 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机器学习领域中。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[28]。

2.1.4.5 Spark 分布式的异构性

由于 Spark 采用分布式计算，需要使用多台计算机进行并行运算，但每台计算机的性能结构不同，导致分布式的计算负载效率不统一。论文[29]提出了一种基于负载的动态分配方法来解决分布式异构的问题，提出对性能低的机器分配较少的负载和计算任务，动态调度计算负载，从而使并行效率达到最大化。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[19][29]。

2.2 前人的研究成果

2.2.1 网络流量分类方法的性能评估策略

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[30][31]：

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：

$$\text{recall（i）}=\frac{\text{TP}_i}{TP_i+\text{FN}_i} \qquad (1)$$

$$\text{precision（i）}=\frac{\text{TP}_i}{TP_i+\text{FP}_i} \qquad (2)$$

$$\text{overall\_recall（i）}=\frac{\sum_{i=1}^{m}\text{TP}_i}{\sum_{i=1}^{m}TP_i+\text{FN}_i} \qquad (3)$$

$$\text{F-measure} = \frac{2\times precision\times recall}{precision+recall} \qquad (4)$$

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[7]。

2.2.2 网络流量分类的应用

在实际生产方面，网络流量分类对网络流量异常检测的应用很广泛，除了能帮助网络管理人员监控网络状况，预警网络异常之外，还针对不同的网络攻击做出了相对精确的识别，从而帮助网络人员发现并阻止异常流量的进入。主要的网络异常检测应用有以下方面：

（1）入侵检测的识别与分类：主要关注数据包（packet）的特征及其来源，检测并识别异常流量，分析数据的来源，将异常流量分为基于主机的入侵行为和基于网络的入侵行为。分析数据

特征可以分为基于特征的检测和基于异常的检测[32]。

（2）DDos（分布式拒绝服务）攻击检测：DDoS 攻击主要分为漏洞利用型攻击和资源耗尽型攻击[33]。该攻击属于当前网络的主要攻击手段，对现代网络有非常大的影响，其攻击具有攻击强度大、攻击影响范围广、攻击源分布性强、攻击隐蔽性强等特点。网络流量的异常检测能及时发现并防护 DDoS 攻击。

（3）用户行为识别：主要关注流量的信息熵来分析用户的网络行为，通过提取网络流量行为特征，研究信息熵的粗粒度来分析用户行为[22]。也有学者基于流量的关联相似性提出了识别监控僵尸网络的方法，来维护网络的安全[34]。

2.2.3 Spark 技术的发展现状

Spark 技术起源于 UC Berkeley AMP Lab。但于 2013 年 6 月才进入 Apache 孵化器项目。2015 年才进入了飞速发展的一年。目前，各大公司开始放弃 MapReduce，并开始转型 Spark。Spark 技术迎来的空间的发展机遇。

Spark 是一种分布式高计算的框架，在网络流量分类的应用也才刚刚起步，但许多前辈已经取得了巨大的成果。

2.2.4 基于 Spark 的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。高速复杂网络环境下的网络流量分类变得愈发困难。但随着 Spark 技术的兴起，在此过程中网络流量分类技术也逐步发展，并取得了长足的进步。

科研人员对此进行了大量的工作，如文献[19]利用 Spark 框架，使用 k-means 算法对实时的网络流量进行了分类，解决了网络流量分类的实时问题。文献[35]将 Python 与 Spark 技术相结合，利用 Python 和 IPython 进行数据预处理，并在 Spark 框架下使用 k-means 算法对流量进行异常检测。[36]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。[37] 提出了一种利用 Spark SQL 技术存储流量数据，并反馈训练。将 k-means 算法与随机森林算法结合，在分类流量的同时，进一步检测异常流量，达到了准实时的目的。

2.3 本课题的创新之处

通过研读各类文献，自选了本课题，本课题有以下创新之处。

2.3.1 基于 Spark 生态圈设计研发一套完整的网络流量分类系统

将 Spark 技术应用于网络流量的分类上，提高分类的效率。基于 Spark 的生态圈设计研发一套完整的网络流量分类系统。具体方案如下：

1. 利用 Spark Streaming 技术实现高速复杂环境下的网络流量实时分类，并区分检测其中的正常流量与异常流量。
2. 利用 Spark SQL 和 Redis 将分类的流量分别存储，并将 Redis 中存储的异常流量作为标记样本，重训练流量分类模型，提高其识别的准确率。
3. 利用 Spark R 将分类好的网络流量进行数据的可视化。通过 Web Service 将其呈现出来，以实现对网络流量状态的实时监控。

由于 Spark 生态圈功能强大，本课题的下一步工作就是在 Spark 框架之下，寻找更多的 Spark 技术来进行网络流量分类，也可将 Spark 与 Hadoop 相结合，取长补短，例如利用 Hadoop 的 HDPF 对数据进行预处理等，旨在寻找一种效率更高、效果更好的网络流量分类系统。

2.3.2 考虑网络流的相关性特征的半监督聚类算法

传统的网络流量分类，没有考虑流之间的相似性和相关性。往往造成聚类效果不佳。本课题在流聚类之前，考虑流之间的内在关联，提高其组内相似度，取得了更好的聚类效果。

此外，有监督机器学习方法往往需要大量的带标记的训练样本。对于无监督的聚类算法，其缺点是准确率不高，并且类簇数需要事先确定。若类簇数过大容易过拟合，若类簇数较小则准确度明显下降。本课题利用 Spark Streaming 技术与 Spark SQL 技术相结合，利用分类完的正异常网络流量数据，反向迭代指导标记训练模型，形成一种半监督的聚类算法，这种算法模型通过不断的自更新，反馈标记训练集。将会具有更高的识别率。

2.3.3 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。

2.4 已查阅的文献目录

[1]周显春，肖衡. Spark 框架下聚类模型在网络流量异常检测中的应用[J]. 网络安全技术与应用, 2017 (5) :62-63

[2]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).

[3]夏正敏. 基于分形的网络流量分析及异常检测技术研究[D]. 上海交通大学，2012。

[4]王宏. 基于行为分析的通信网络流量异常检测与关联分析[J]. 通讯世界 , 2015 (8) :90-90

[5]胡洋瑞，陈兴蜀，王俊峰，叶晓鸣. 基于流量行为特征的异常流量检测[J]. 信息网络安全 , 2016 (11) :45-51

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

[7]赵英，陈骏君. 基于流相关性的网络流量分类[J]. 计算机工程与应用 , 2015 , 51 (21) :25-29

[8]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.

[9]Can Bozdogan, Yasemin Gokcen, Ibrahim Zincir. A Preliminary Investigation on the Identification of Peer to Peer Network Applications[J].ACM,2015,07(11)

[10]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.

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

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

[13]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.

[14]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.

[15]周文刚，陈雷霆，Lubomir Bic，董仕. 基于半监督的网络流量分类识别算法[J]. 电子测量与仪器学报 , 2014 , 28 (4) :381-386

[16] Sandy Ryza, Uri Laserson,Sean Owen,Josh Wills,. Advance Analytics with Spark[M],人民邮电出版社，2015

[17] CSDN; Spark in action. http://www.cnblogs.com/shishanyuan/p/4699644.html

[18] 刘泽燊，潘志松. 基于 Spark 的大规模网络流量分类研究[J]. 计算机时代 , 2016 (4) :1-5

[19] M Čermák， T Jirsík， M Laštovička. Real-time analysis of NetFlow data for generating network traffic statistics using Apache Spark. IEEE,2016:1019-1020

[20] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M.McCauley,M. J. Franklin, S. Shenker, and I. Stoica,"Resilient distributed datasets:A fault-tolerant abstraction for in-memory cluster computing," in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation,2012

[21]王苏南. 高速复杂网络环境下异常流量检测技术研究[D]. 解放军信息工程大学,2012.

[22]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.

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

[24]韩德志， 陈旭光， 雷雨馨， 戴永涛， 张肖. 基于Spark Streaming的实时数据分析系统及其应用[J]. 计算机应用 , 2017 , 37 (5) :1263-1269

[25]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).

[26]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).

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

[28]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.

[29]M Kulariya ， P Saraf ， R Ranjan ， GP Gupta. Performance Analysis of Network Intrusion_Detection Schemes using Apache Spark[J]. International Conference on Communication & Signal Processing,2016 :1973-1977

[30]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.

[31]郭成林. 基于 Spark 平台的恶意流量监测分析系统[D]. 北京交通大学，2016

[32] 薛京花. K-means 聚类算法在网络入侵检测中的应用研究[D]. 中南林业大学，2012

[33] 王飞. 分布式拒绝服务攻击检测与响应技术研究[D]. 国防科技大学，2013

[34] 蒋鸿玲. 基于流量的僵尸网络检测方法研究[D]. 南开大学，2013

[35] 吴晓平，周舟，李洪成. Spark 框架下基于无指导学习环境的网络流量异常检测研究与实现[J]. 信息网络安全 , 2016 (6) :1-7

[36] 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.

[37] 周超. 基于 Spark 技术的实时网络流量异常检测研究[D]. 兰州交通大学，2016.

3 研究计划部分

3.1 论文选题的立论、目的和意义

3.1.1 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，规模不断扩大,应用领域不断扩展,由此产生的各种安全问题日益凸显,安全问题是互联网技术领域最受关注的问题之一。

虽然机器学习算法具有分类准确，快速等特点。但大部分基于机器学习的网络流量分类研究都是在单机环境下进行的，难以满足现实网络中大规模流量的实时分类任务。在大数据运维管理的应用背景下，银行业、证券业等是对安全性和稳定性要求较高的行业，需要实时保证上千台服务器的正常稳定工作，因此大数据运维管理应运而生。流处理的应用表现在实时性上，运维管理的预警功能在实时性方面的需求尤为突出，而Spark Streaming 将可以良好地应用到流模式分析当中。

在网络安全方面，网络流量的异常检测是保证网络安全性的一项重要技术，实时的网络流量异常检测也是大数据运维管理的方向之一。大数据的在线实时分类是当下网络流量分类急需研究

解决的问题之一。

3.1.2 本文的目的

本文的目的是设计一个基于 Spark 生态圈的（准）实时流量分类系统，首先该系统会在高速复杂的网络环境下，采集并分类大量网络流量，尽量做到实时性。然后根据获得分类的流量数据，进行反馈训练模型，形成一种半监督的自学习模型，提高分类效率与准确率。其次，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。本文拟采用双层模型就行流量分类，先经过改进的 k-means 算法进行流量分类，再利用随机森林算法作为第二层模型，对网络流量中的异常流量进行下一步判别，为异常检测提供依据。最后将实时分类的网络数据，通过 Spark R 进行数据降维可视化，最终通过 Web Service 这个媒介实时展现出网络流量状况，以供网络安全人员实时监控网络状态。

3.1.3 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）发现并挖掘新的网络流量，是新型网络应用与网络协议研究的需求。（5）分析网络行为，为用户提供更优质的网络服务。（6）网络流量分类是网络异常检测的关键一环。

利用 Spark 进行网络流量分类具有以下重要意义：（1）解决高速复杂环境下，网络流量实时分析难的问题。（2）解决时下利用单机分类网络流量，效率低的问题。（3）为流量分类及其可视化提供了另一种可行性方案。

3.2 本课题主要研究内容

3.2.1 流的分类技术与模型集成

（1）流的分类技术集成

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。本文在流量分类时，考虑流之间的相关性，寻找适当的算法发现并利用流之间的相关信息，增加分类的准确性。

（2）使用多层模型

以往的网络流量分类技术往往采用的是单层模型的无监督聚类算法。本文拟采用双层模型的半监督聚类算法训练模型。先利用改进的 k-means 算法做第一层模型进行流量分类。再利用随机森林算法作为第二层模型对分类流量进行具体识别。然后将识别后的流量作为标注数据集回馈训练模型，形成一种半监督的自学习模型。

（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。

3.2.2 Spark 生态圈技术

（1） 数据的采集以及预处理

如何在高并发的复杂网络环境下采集流量，是网络流量分类研究人员所面临的一大难题。本课题拟利用 Spark Streaming 技术，在高速复杂的网络环境下进行数据采集，并利用中间件 kafka 和 FCBF 对采集的流量数据进行预处理。

（2） 数据的存储

怎样存储大容量的网络流量数据也是本课题需要研究解决的内容。本课题拟在使用 Spark SQL 技术，与 Redis 相结合，考虑到真实网络环境中，正常流量远远多于异常流量。将正在流量存入 SQL 之中以便后期流量分析使用，将异常流量存入 Redis 之中，然后将其作为标注数据集反馈训练模型。

（3） 数据可视化

将处理好的流量分类可视化呈现给网络管理人员也是至关重要的一步。本课题拟在使用 Spark R 技术将分类好的网络流量进行降维可视化，并通过 Web Service 这个媒介将数据展现出来。

3.2.3 测试数据

（1）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，拟在使用上一次的分类信息作为标注回馈训练下一次的数据分类模型。有效的减少了时间和人力。

（2）性能分析

对于使用 Spark 技术分类的流量在性能上相较于单机分类效果怎样？算法分类的准确度怎样？以及对 CPU 的使用率是怎样的？都是需要经过具体的实验分析比较，才能得出结果。

3.3 研究方案

3.3.1 技术方案

3.3.1.1 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，研究主要集中在对网络流量分类的基础上使用 Spark 技术做分布式计算，搭建网络流量分类模型。此阶段的数据可以通过 CAIDA（Center for Applied Internet Data Analysis）获得，也可以采用 KDA 99 发布的网络流量数据。

第二阶段，研究主要集中在对高速复杂环境下的网络流量实时分类，通过第一阶段建立的网络流量分类模型，来验证真实环境下的网络流量分类。因为网络流量涉及用户隐私问题，此阶段的数据来源个人路由流量出口，采集该端口的流量来进行实时分类，验证系统的准确率与可行性。如能获得允许，将把该系统用于某校网络端口进行测试。

3.3.1.2 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 Scala 语言编写所有接口的代码，调用部分 Spark API 进行编程
- 使用 Spark Streaming 采集网络流量，并使用 Spark SQL+Redis 存储流量数据
- 使用 Spark MLib 的改进 K-means+随机森林算法进行网络流量分类。
- 使用 R 语言进行数据可视化，使用 PHP+Nginx+Mysq 开发 Web Service 进行数据展示
- 对模型进行测试校验，验证系统性能以及准确率。

3.3.1.3 算法设计测试方案

- 设计网络流量分类的双层模型
- 设计网络流量的采集方法以及预处理方案
- 将系统结果与单机结果，单层模型算法，各类机器学习算法进行性能比较
- 使用 Scala 实现算法，综合各类 Spark 生态圈技术

3.3.2 实施方案所需条件

- 流量数据集，以及真实环境的网络流量出口数据
- Spark 开源框架，MySQL，Redis 等开源软件
- 多台高性能服务器（至少两台）
- 能互联的网络出口

3.4 本课题难点

- 高速复杂网络环境下的流量采集及预处理
  --由于现在真实网络中的接入设备纷繁复杂，并且网络会因为各种情况产生坏的，无意义的网络流量。所以如何进行流量采集及预处理，是本课题的难点之一。
- 不同算法结果的综合方法的选择
- 流量的特征提取

- 系统的计算开销
- Spark 技术的应用

3.5 预期的研究成果及创新点

- 利用双层模型的分类检测效果高于单一的分类模型
- 解决高速复杂网络环境下的实时分类问题。解决集群异构问题，使系统具有可行性和普适性。
- 对分类好的网络流量进行数据展示，给网络监管人员提供更直观的网络状态展示。
- Spark 技术的应用能加快流量分类的效率，并且利用集群特点，能实行系统实时备份。

3.6 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分类算法及 Spark 生态圈技术。
ii. 2018 年 1 月至 2018 年 3 月，设计算法、设计软件框架，设计系统流程。
iii. 2018 年 4 月至 2018 年 7 月，编写系统，实现系统功能。
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：