网络流量分类研究进展与展望

熊 刚[1,2] 孟 姣[1,2] 曹自刚[3] 王 勇[4] 郭 莉[1] 方滨兴[3,1]

[1](中国科学院计算技术研究所信息安全研究中心 北京 100190)

[2](中国科学院研究生院 北京 100049)

[3](北京邮电大学 北京 100876)

[4](国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 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。网络流量分类技术作为增强网络可控性的基础技术之一，不仅可以帮助网络运营商提供更好的服务质量，而且能够对网络进行有效的监督管理，确保网络安全。本文综述了网络流量分类领域的研究方法及研究成果，对这些传统方法进行比较，分别指出它们的优势和不足。并针对高速网络环境下的实时分类、加密流分类、精细化分类、协议动态变化时的分类等现实挑战，对相关研究进展进行阐述和分析。最后对未来的研究方向进行展望。

关键词 流量分类；高速网络；精细化；加密；协议混淆

Research Progress and Prospects of Network Traffic Classification

XIONG Gang[1,2] MENG Jiao[1,2] CAO Zi-gang[3] WANG Yong[4] GUO Li[1] FANG Bin-xing[3,1]

[1](*Research Center of Information Security, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2](*Graduate University of Chinese Academy of Sciences, Beijing* 100190)

[3](*Beijing University of Posts and Telecommunications, Beijing* 100876)

[4](*National Computer Network Emergency Response Technical Team/ Coordination Center of China, Beijing* 100029)

Abstract In recent years, the number of applications and the scalability of the Internet have experienced a rapid improvement. As one of the basic technologies for enhancing network controllability, traffic classification can not only provide better QoS for ISPs, but also supervise and manage network effectively, which can ensure the security of the Internet. In this paper, we first review the methodology and achievements in the field of traffic classification by comparing these traditional methods, and pointing out their advantages and disadvantages. Then we explain and analyze the related research progress aiming at challenges in reality such as real-time classification in backbone network, encrypted traffic classification, fine-grained classification, and constantly changing protocols classification etc. Finally, we look into the future of our research.

Keywords traffic classification; high-speed network; fine-grained; encryption; protocol obfuscation

1 引 言

近年来，Internet凭借其开放性、共享性等特点迅速普及并发展壮大，传统的互联网业务已经无法满足人们的需求，越来越多的新型网络应用应运而生。为了有效利用带宽，并提供更好的服务质量（quality of service，QoS），需要网络能够针对

基金项目：国家高技术研究发展计划（“863”计划）（2011AA010703）、国家自然科学基金项目(61070184)资助项目。**熊刚**，高级工程师，研究方向为信息安全。E-mail: xionggang@ict.ac.cn。**孟姣**，硕士研究生，研究方向为信息安全。E-mail: mengjiao@software.ict.ac.cn。**曹自刚**，博士研究生。**王勇**，高级工程师。**郭莉**，正研级高工。**方滨兴**，中国工程院院士，北京邮电大学教授，中国科学院计算技术研究所网络方向首席科学家。

不同应用进行分类。另一方面，Internet的开放性特点也意味着任何符合其技术标准的设备或软件都可以不受限制地接入互联网，导致了Internet的各类安全事件层出不穷，网络安全问题变得日益严重。

为了解决当前互联网发展过程中面临的这些问题，在保障网络安全的同时提供更好的服务质量，就需要对网络进行有效的监督和管理，增强网络的可控性。网络流量分类技术作为增强网络可控性的基础技术之一，可以帮助研究人员了解网络上的流量分布，允许网络运营商（internet service provider，ISP）优先一些重要的传输，并且阻止网络犯罪行为的发生。

传统的网络流量分类方法对于传统网络应用具有很好的分类效果，然而，随着越来越多新型应用的兴起，该技术也面临着巨大的挑战。例如：许多应用程序使用了私有的应用层协议，用加密的方式保护其数据内容，还有一些应用使用不规则的非标准端口号等。同时，为了更好地分析用户行为，对网络提供及时有效的监督和管理，精细化的流量分类和实时的流量分类思想也显得尤为重要。

本文主要介绍了流量分类研究的进展及面临的挑战，并提出了相关挑战的解决方案。全文内容安排如下：第2部分介绍了流量分类的基本概念及评价指标；第3部分介绍了流量分类研究的进展，并对其进行了评述；第4部分介绍了流量分类面临的挑战，对相关研究进展进行阐述和分析；最后，对流量分类技术进行了总结与展望。

2　流量分类基本概念与评价指标

很多网络应用具有自身的特性，对于网络环境的需求也不尽相同，因此只有对网络流量进行及时准确的识别和分类，才能准确地为不同应用提供合适的网络环境，有效利用网络资源，为用户提供更好的服务质量。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[1]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[2]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[3]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。

流量分类的一个关键度量标准是某个分类技术或分类模型对未知数据对象进行分类的准确率。通常用于衡量分类准确率的评估标准，主要包括以下四个方面：

真正（true positive，TP）：表示被分类模型正确预测的正样本数，即属于类别A并被预测为类别A的样本数。

假负（false negative，FN）：表示被分类模型错误预测为负类的正样本数，即属于类别A但被预测为不属于类别A的样本数。

假正（false positive，FP）：表示被分类模型错误预测为正类的负样本数，即不属于类别A但被预测为属于类别A的样本数。

真负（true negative，TN）：表示被分类模型正确预测的负样本数，即不属于类别A并被预测为不属于类别A的样本数。

此外，基于机器学习的分类方法通常采用另外两种度量标准对其分类结果进行评估，其定义如下：

召回率（recall）：recall=TP/(TP+FN)，表示类别A中被正确预测的样本所占比例。

精度（precision）：precision=TP/(TP+FP)，表示在所有被预测为类别A的样本中，真正属于类别A的样本所占比例。

目前，很多流量分类研究都使用流准确率或字节准确率作为其实验结果的度量标准，流准确率表示被正确分类的流所占的比例，而字节准确率则更关注被正确分类的流所携带的字节数。其中，准确率的定义如下：

准确率（accuracy）：accuracy=(TP+TN)/(TP+TN+FP+FN)，表示被分类模型正确预测的样本数在总样本中所占比例。

3　流量分类研究进展与评述

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、基于有效负载的分类方法、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，如图1所示。

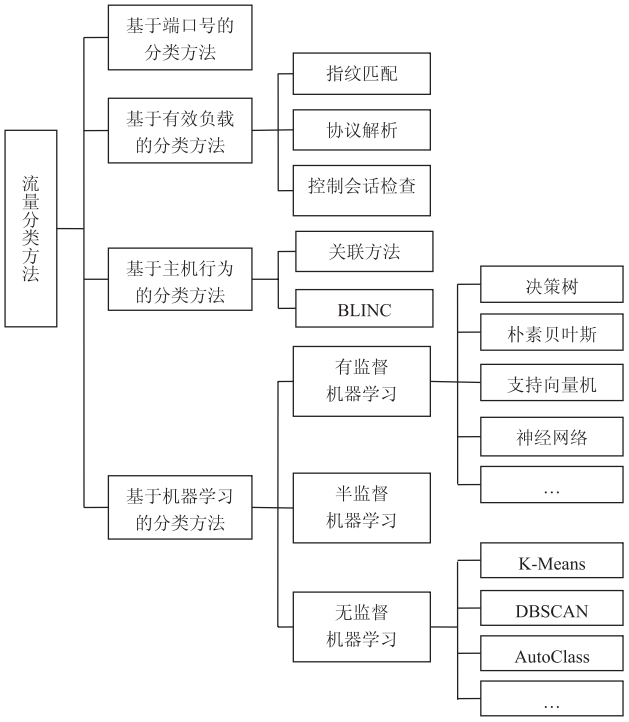

图1 流量分类的方法

3.1 基于端口号的流量分类方法

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，将熟知的端口号（IANA[4]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分类[5]。然而，它也具有一定限制。例如一些应用可能没有在IANA注册端口号，或者使用熟知端口号以外的端口。尤其是随着P2P应用的出现，它采用动态分配端口的技术，使通过端口号映射的方法检测应用类型受到了阻碍。此外，对于IP层载荷的加密也使得无法获取端口号，致使该方法无法实施。

2004年，Sen[6]等人对Kazaa P2P协议进行实验，测得默认的P2P端口号只占测试总流量的30%左右。2005年，Moore和Papagiannaki[7]通过实验测得：使用官方的IANA列表进行基于端口号的分类，其准确率不超过70%。Madhukar和Williamson[8]证实了使用端口号分类的方法，对于他们实验时30%-70%的流量都无法识别。

3.2 基于有效负载的流量分类方法

为了避免对端口号的过分依赖，提出了基于有效负载的分析方法。该方法通过分析包的有效负载是否包含已知应用的特殊签名进行流分类，具有较高的准确性。Sen[6]等人使用分析有效负载的方法对P2P流进行分类，有效降低了分类的假负率（FN）和假正率（FP）至实验总流量的5%。

虽然该方法具有很高的分类准确率，但分析代价太大。为了降低计算代价，可将其与一些分析代价较低的分类方法结合使用，先过滤出一些很容易分析出的流量，以减少计算开销。Moore和Papagiannaki[7]使用了一种端口号和有效负载相结合的技术来识别网络应用，实验测得69%的流可以通过端口号映射的方式被正确分类，在分析端口号的基础上分析流的前1kByte信息可将分类准确率提高到79%，对于上述两种方式都无法分类的流，分析其负载的全部内容可将其正确分类，分类的准确率接近100%。

基于有效负载的分类方法虽然避免了过分依赖端口号所带来的问题，但其自身也存在一定的限制：它只能识别那些已知的非加密流量，而无法分类其他未知流量；它的流量识别过程更加复杂，并需要对应用协议语义的大规模信息保持更新，需要较高的处理和存储能力；此外，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。

3.3 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。

2004年，Karagiannis[9]等人提出了基于P2P流量的连接模式来识别P2P应用类型的方法。实验采用了两个主要的启发式来检测不同流的行为特征。第一个启发式检测同时使用TCP和UDP两种传输层协议进行数据传输的{源IP，目的IP}对，第二个启发式用于监控{IP，port}对的连接模式。该方法不仅可以有效识别99%的P2P流和超过95%的P2P字节，而且还可以识别基

于有效负载方法未能识别出的P2P流。

2005年，在之前研究成果的基础上，Karagiannis[10]等人又提出了一个新的基于主机传输层行为模式进行流分类的方法，并称之为BLINC。该方法在社会层、功能层及应用层对主机行为进行分析。其中，社会层主要分析某一台主机在一定时间内与哪些其他主机进行通信，以及与某些特定主机进行通信的一批主机的集合；功能层主要分析一台特定主机在网络中是提供服务的一方、请求服务的一方，还是二者兼有；应用层将前两层分析后得到的信息和传输层端口信息结合起来推断原始应用的类型。使用该方法进行流量分类，可以识别出所有实验数据中80%-90%的流量，并且准确率超过95%。

此外，Karagiannis等人还提出了依赖于以往经验的启发式方法，即除了以上端口和IP等基本信息之外，还可以根据实际情况，利用一些数据流中的其他信息来改进BLINC方法，并提高其分类准确率（例如，无负载的数据流或失败的链接往往代表这个流是一次攻击或者是一个P2P用户在尝试连接一个已经断开网络的IP地址）。

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。

3.4　基于机器学习的流量分类方法

随着流量分类需求的增大，一种基于流量的统计特征来识别应用的新方法被提出。这种方法潜在假设了对于某些类别的应用，网络层的统计特征（例如流持续时间的分布、包间隔时间和包长度等）是唯一的，可以用来将它与其他应用区分开[11]。由于对大规模数据集的分析需求，简单的统计特征方法已经无法满足需求，因此更加系统的机器学习技术被提出。

机器学习的过程通常由两部分组成[12]，即分类模型的建立和分类。首先采用训练数据建立分类模型，然后基于该模型产生一个分类器，并对未知数据集进行分类。1994年，机器学习首次被用于入侵检测中的网络流量分类领域[13]。目前，用于流量分类的机器学习方法主要包括无监督方法和有监督方法，此外还有将这两种方法相结合而产生的半监督方法。

3.4.1　无监督机器学习方法

无监督机器学习方法即聚类方法，它使用内在的启发式来发现数据中存在的簇[14]。同一个簇中的对象彼此相似，不同簇中的对象彼此相异。该方法通过发现和标记数据集中的簇来构造分类器，分类过程主要包括两部分，即评估一个对象与哪个簇具有更大的相似性，以及标记对象所属簇的类别[15]。其分类过程如图2所示。

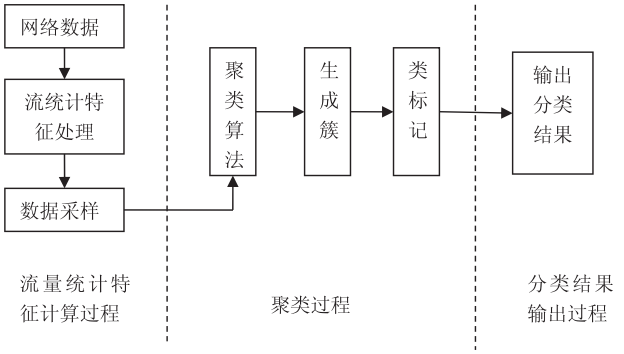

图2　基于无监督机器学习的流量分类过程

2004年，McGregor[16]等人以包长、包间隔时间、流持续时间等作为统计特征，利用EM（期望最大化）算法[17]，通过无监督学习的方式将流量按其类型（如批量传输、小规模传输等）进行了分类。

2005年，Zander[18]等人采用了一种AutoClass[19]方法，AutoClass是一个无监督的贝叶斯分类器，使用EM算法决定数据对象应该属于最优簇。该方法使用NetMate[20]工具计算流的特征值，并采用SFS（sequential forward selection）方法选择较优的属性集。实验将获取的流统计数据传给分类器进行学习，学习和分类的结果再作为下一次分类的评估标准。机器学习的时间越长，则分类准确性越高，一旦达到某个标准，就可以对后续的输入数据流自动分类[5]。

2007年，Erman[21]等人提出了针对网络核心流量分类的解决方案，在网络核心的入口点和出口点按照流量类别（如Web, P2P, FTP等）来识别和区分网络流量。实验采用了K-Means算法，以平均包长、流持续时间、平均包间隔时间等作为特征值，并取得了较好的实验效果，对服务器到客户端的单向流量识别的流准确率达到了95%。

目前比较常用的聚类算法有很多，它们在聚类

速度、聚类效果等方面具有各自的特点。2006年，Erman[22]等人比较了K-Means、DBSCAN[23]（density-based spatial clustering of applications with noise）和AutoClass这三种聚类算法。K-Means算法是一个划分算法，DBSCAN算法是一个基于密度的方法，而AutoClass则是一个基于概率模型的算法。实验结果显示，K-Means和DBSCAN比AutoClass的聚类速度更快，但AutoClass算法的准确率最高，DBSCAN算法在一个聚类的小的子集中包含了大多数的连接，具有最好的聚类效果，而K-Means虽然在准确率方面不及AutoClass，但其聚类速度却相当快。

3.4.2 有监督机器学习方法

有监督机器学习方法即分类方法，主要根据已标记样本的特点构造分类规则或分类器，将未知类别的样本映射到给定类别中的一个。机器学习过程的输入为一些已经分好类的样本实体的集合，输出为通过这些样本产生的一个分类模型。

用于训练和测试的数据需要具有相同的特征集，并且被标记好所属类别，数据的标准格式如图3所示[24]。

序号	特征 1	特征 2	…	特征 n	类别
1	x	xx	…	xxx	1
2	x	xx	…	xxx	1
…	…	…	…	…	…
m	x	xx	…	xxx	0

图3 数据集的标准格式

有监督机器学习的工作过程如图4所示，它主要包括两个过程：

（1）训练过程：根据提供的训练数据集构造一个分类模型。

（2）分类过程：利用训练过程中产生的分类模型对未知类别的样本进行分类。

目前，分类模型的构造方法主要包括决策树、朴素贝叶斯、支持向量机、关联规则学习、神经网络、遗传算法等。

1998年，Murthy[25]为机器学习领域的研究人员提供了关于决策树的一个概述。决策树是一个预测模型，它代表了对象属性与对象之间的一种映射关系。树中的每个节点表示对象的某个特征，每个分叉路径则代表某种可能的属性值，每个叶节点代表所属类别。每个实体从根节点开始，根据其特征值分类。决策树的代表方法有ID3[26]、C4.5[27]等。

贝叶斯网络是一个带有概率注释的有向无环图，贝叶斯分类器是用于分类的贝叶斯网络，其分类原理

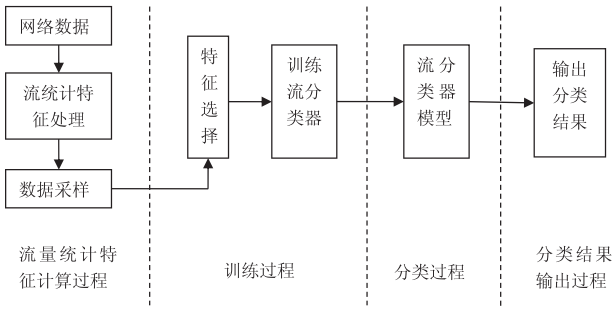

图4 基于有监督机器学习的流量分类过程

是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，选择具有最大后验概率的类作为该对象所属的类。目前研究较多的贝叶斯分类器主要有Naïve Bayes、TAN、BAN和GBN。Moore[28]等人使用简单的朴素贝叶斯技术进行流量分类的准确率约为65%，而使用基于核密度估计的朴素贝叶斯(NBKE)和基于快速相关过滤器的朴素贝叶斯(FCBF)降低了特征空间维度，大大提高了分类准确率，使准确率超过了95%。

支持向量机（support vector machine，SVM）是Corinna Cortes[29]等人于1995年首先提出的，该方法利用非线性变换和结构风险最小化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，其网络流属性不必满足条件独立假设，无需进行属性过滤，并能够在先验知识相对不足的情况下，仍保持较高的分类准确率，不依赖于样本空间的分布，具有较好的分类稳定性[30]。Zhu Li[31]等人在19种流量特征中选择了对分类影响最大的9种作为特征值，利用SVM技术将网络流量分成了Bulk traffic, Interactive, WWW, Service, P2P, Mail, Other七类，得到了95%以上的准确率。

神经网络（neural networks）是一种模仿动物神经网络行为特征，进行分布式并行信息处理算法的数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的，具有自学习和自适应的能力。Auld[32]等人通过246个统计特征，利用贝叶斯神经网络的技术进行流分类，获得了比朴素贝叶斯方法更高的准确率。用同一天的网络流量作为训练数据和测试数据时，分类准确率为99%。用相隔八个月的网络流量作为训练数据和测试数据时，分类准确率也达到了95%。

遗传算法是一种借鉴生物进化规律演化而来的随机化搜索方法，依据“适者生存，优胜劣汰”的遗传机制，自适应地调整搜索方向，而不需要确定的规则。2006年，Park[33]等人提出了一种基于遗传算法的特征选取技术，实验采用了决策树的方法，利用流的

前10个包的统计特征，将J48和REPTree这两种基于决策树的分类器和基于朴素贝叶斯的NBKE分类器进行了比较，结果显示两个基于决策树的分类器具有更高的准确率。

有监督的学习算法很多，每种方法都有其特点，需要根据实际应用环境的不同，选择合适的算法，才能达到更好的效果。Williams[34]等人从分类准确率和计算性能（分类模型建立时间、分类速度）这两方面，对朴素贝叶斯（包括NBD和NBK）、C4.5决策树、贝叶斯网络、贝叶斯树这五种机器学习算法进行了比较。实验结果显示，几种分类算法在分类准确率方面差别并不大，但在计算性能方面却相差很多。C4.5决策树算法的速度最快，而NBK的速度最慢。贝叶斯树构造分类模型所需时间远远超过了其他几种算法，而NBK建立分类模型的速度最快。文中还表明，对于大多数分类算法，减少特征值的个数可以有效提高算法的计算性能，提高其分类速度。

3.4.3　半监督机器学习方法

由于被标记的样本集并不容易获得，因此通常用于机器学习的训练样本数量不够广泛，无法较好的反映出各类别流量的特征。而且随着互联网的迅速发展，新的网络应用不断出现，这些新应用的特征无法提前预测，传统的有监督机器学习方法也无法对这些未知流量进行分类。为解决这些问题，出现了一种半监督机器学习方法，该方法是有监督和无监督两种方法的结合。

半监督机器学习方法的训练集由已标记样本和未标记样本两部分组成。首先，通过聚类算法将训练集分成不同的簇，然后通过被标记的流实现簇与类别之间的映射，那些不包含任何标记流的簇就被视为未知的新应用类型。Erman[35]等人较早的将半监督机器学习方法应用于流量分类领域，其训练集包含了少量标记流样本和大量未标记的流样本，实验证实该方法具有较好的准确率，并且可以识别未知应用流量和改变了行为模式的已知应用流量。2011年，Xiang Li[36]等人使用了一种半监督的SVM方法进行流量分类，该方法只需少量的标记流，并且分类准确率高，计算代价低。

半监督学习方法只需利用少量标注样本和大量未标注样本即可实现分类，可以有效减少标注代价，提高机器学习的性能，但它对于有噪声干扰样本数据的分类效果并不理想，而且由于它提出的时间比较短，其应用价值还需要更多的深入研究。

3.5　流量分类方法的比较

前面的几节主要介绍了四种流量分类技术，并结合一些已有的实验及成果对其进行了评述。下面将从分类准确性、适应场景、优缺点等方面对这几种流量

表1 流量分类方法的比较

分类方法	基于端口号的流量分类方法	基于有效负载的流量分类方法	基于主机行为的流量分类方法	基于机器学习的流量分类方法
准确性	低	极高	较高	较高
使用场景	较简单的网络流量结构，使用IANA中注册端口号的传统网络应用。	数据包内容为明文的流量分类与识别	大流量的骨干网络中（汇聚的行为主机越多，流量识别准确率越高）	传统及新兴的网络应用，其他几种方法无法应对的场景。
优点	技术简单、易操作、计算开销较低，分类速度快。	分类准确率很高，并且可以对P2P等新型网络应用进行准确的分析。	能够对加密数据流量进行分析，开销较低、容易实施。	准确率较高、可扩展性强，能够对加密数据流量进行分析。
缺点	若网络应用没有使用IANA中注册的端口号，或者使用动态端口，则无法使用该方法。	计算开销太大，涉及到隐私侵犯问题，并且无法解决私有协议及加密流量的分析问题。	无法识别一些特定应用的子类型，当传输层数据被加密时无法使用。	耗费资源较多，需要事先了解数据样本集，应用于在线环境具有一定困难。
单独使用时对高速网络流量分类的适用性	不适合单独使用	适用范围有限	比较适用，但需其他方法配合使用	不太实用，处于实验阶段
单独使用时是否适用于加密流量分类	否	否	是	是
单独使用时是否适用于精细化流量分类	否	是	是	是
单独使用时是否适用于动态变化协议的流量分类	否	否	处于试验阶段	处于试验阶段

分类技术做一个比较，如表1所示。

4 流量分类面临的挑战与技术手段

随着网络的迅速发展，一些新的网络应用不断涌现，网络规模不断扩大，应用类型纷繁复杂，不同应用的流量也呈现了不同的特征，并处于一种动态变化的过程，这给流量分类领域带来了巨大的挑战。

4.1 高速网络环境中的流量实时分类

目前针对校园网等环境下的流量分类技术有较多研究，但对骨干网环境下的流量分类研究较少。

骨干网环境中的流量特征与接入网环境存在较大差异，且其吞吐量较高（单光纤可接近10Gbps），这就对流量分类技术提出了更高的要求，即不仅要保证较高的识别准确率，而且要减少分类需要的代价，及早进行分类判定（如数据流只流经几个包就能得出分类结果），尽可能提高分类速度和性能。

从当前流分类领域的研究成果来看，大多数的研究工作都围绕提高流量分类的准确率展开，虽然这些分类技术都宣称能达到较好的分类效果，但大多都是在小规模实验环境下的测试结果，若将其应用在高速网络环境中，分类效果并不尽如人意。

另外，在高速网络环境流量分类中，对流量分类效果评价指标的侧重点也可能与小规模网络环境有所不同，网络管理者可能更关注于某一类流量是否都被识别出来，即首要关注特定流量的实时检全率或实时漏检率，其次才关注误检率，这种需求也使得很多传统的分类技术在此情况下难以适用。

传统的基于端口号的流量分类方法所需代价小且速度快，很适合高速网络环境对分类性能的需求，但它只能识别出一些熟知端口的应用，无法应对很多其他应用，因此无法满足骨干网环境下的分类要求。基于有效负载主机行为的分类方法虽然弥补了基于端口号分类方法的不足，但其分类代价大、消耗资源多，也无法满足实时高速网络环境的分类要求。基于机器学习的分类方法由于需要进行大量的统计分析，目前难以适应高速网络环境中的流量实时分类，相关研究成果还处于实验阶段。基于主机行为的分类方法虽然在分类效果方面存在一定局限，但它消耗资源较少，并且识别的准确率也要优于基于端口号的方法，因此很适合应用于高速网络环境下。

目前在这一领域的研究较少，但其应用价值却很大，因此在这方面还有很大的研究空间，值得进一步的研究和探索。

4.2 加密流量分类

随着网络资源和带宽的逐渐完善，用户更注意保护隐私，加密应用被广泛使用，加密流分类成为流量分类中的挑战性问题之一。

传统的基于熟知端口号的流分类方法在大多应用端口随机可变和端口共用情况下已经失效，而基于有效负载的流分类方法对加密流量也束手无策，并且存在隐私侵犯的问题，耗费资源较大。

2006年以来，在流量分类领域，对加密流的分类问题成为一个重要的研究方面，大多数研究都采用了基于流特征的统计识别方法。目前研究的加密流量主要包括四类典型流量：SSH隧道、IPSEC隧道、SSL，以及P2P的加密流量。其中，对于SSH的研究较多，主要侧重于对SSH隧道中的应用进行有效的分类识别，使用的方法基本都是基于流特征的识别，主要使用机器学习的方法，例如朴素贝叶斯、C4.5决策树、SVM、k-means、k-nearest neighbor等算法，还有基于高斯混合模型（Gaussian mixture models）、隐马尔可夫模型（hidden Markov model, HMM）和最大似然分类等，另外还有使用基于主机行为的分类方法[37]，主要基于协议连接模型，通过源IP、目的IP和连接特征等特征值来识别，但这种方法自身具有一定局限性，且不容易达到高精确度，因此研究者一般较少使用。

对于SSL和SSH流量的分类通常分为两步[38][39]，第一步根据协议自身特性（如协议首部特定字节的特征或密钥交换特征等）识别出SSL或SSH流量，第二步对加密隧道中的不同应用数据流进行分类。Dusi[40]等人使用高斯混合模型（GMM）和支持向量机（SVM）技术，对运行在SSH隧道中的应用（假设同一时刻SSH隧道中只有一种应用）进行分类（包括HTTP、POP3、POP3S、EMULE、unknown）。实验选取包长度和数据包方向（服务器到客户端或客户端到服务器）作为特征值，结果显示两种分类方法对POP3和POP3S类的应用识别效果最好，真正率均超过了98%，GMM对HTTP应用的识别准确率高于SVM，对EMULE的识别准确率则低于SVM。

一些研究者对多种分类方法进行比较，并通过实验提取出用于识别加密流量的一般性流特征。2007年，Alshammari[41]等人使用了两种有监督的机器学习算法AdaBoost和PIPPER识别网络中的SSH流量，并进一步对SSH应用类型（如ssh，scp，sftp，tunnel等）进

行分类，通过两种方法的对比，PIPPER的分类效果要优于AdaBoost，其准确率达到99%，假正率为0.7%。实验同时说明了，通过基于流统计特征的方法，而不使用IP地址、端口、有效载荷等信息对SSH流量进行精确识别是可行的。之后，Alshammari[42]等人又对以SSH和Skype为代表的加密流量进行了进一步的研究，并使用五种机器学习算法（AdaBoost、SVM、朴素贝叶斯、RIPPER、C4.5决策树）进行对比，对比结果显示基于C4.5决策树的分类器的分类效果最好。

单独的一种分类方法通常在某方面有优势，但不可避免的在另一方面存在劣势，为了获得更好的分类效果，一些研究者使用了一种或多种分类方法结合的技术进行研究。Bar-Yanai[43]等人提出了一种实时的加密流量分类方法，将k-means和k-nearest neighbor两种分类器相结合，使用17个流特征，分析每个流的前100个包的统计信息，并进行分类。传统的k-nearest neighbor分类算法分类准确率高，但分类速度慢，而k-means算法则相反。实验将两种分类器相结合，分类速度快、分类准确率高，比使用单独一种分类算法的效果要好很多。

SSH隧道下可承载很多类型的应用，虽然识别隧道应用的存在并不困难，但要识别出隧道中的具体应用类型就有很大挑战。Tan[44]等人提出了一种利用最大似然分类器对SSH隧道中的应用进行分类的方法，该方法首先需要识别出SSH Tunnel流的建立边界，然后丢弃该边界之前的数据包，只需要分析该边界之后的连续L个数据包的统计信息，对SSH隧道中的HTTP、POP3、FTP、SMTP四种类型的应用进行分类，分类的真正率为90%左右。

为了提供更好的网络服务质量，需要及时识别出网络中的延迟敏感应用（如VoIP、音视频等），并为其提供高优先级。这些应用通常使用非标准端口号，并且其数据被加密，或者被放入加密隧道中执行。Yildirim[45]等人以包长作为特征值，采用了机器学习的方法对IPSec隧道中的VoIP和非VoIP流量进行了分类，并通过对比实验证实了及时识别出VoIP应用并给予其高优先级对于提高QoS的重要性。

对P2P应用（如BT、Emule、Skype等）进行分类的挑战主要集中于数据加密和混淆协议，目前使用较多的方法是基于流特征的统计识别方法，也有一些研究使用了有效负载和流特征相结合的方法，他们使用的有效负载往往是一个连接的前几个包的靠前数据，也是用于统计特征，而不是直接把某个特定字节串作为分类依据。Hjelmvik[46]等人全面展示了利用统计分析和统计协议识别（statistical protocol identification, SPID）算法对混淆协议进行分类的有效性，并通过实验总结出了对一些特定模糊协议（包括BT的MSE、eDonkey的模糊协议、Skype、Spotify）进行分类的最有效统计属性。

总的说来，加密流量分类是流量分类中最具挑战性的问题之一，目前的主要研究思路主要集中在各种行为特征提取及统计分析的方法上。我们认为，基于主机行为的关联分类方法，辅以主动验证等其它的技术手段，也是较好的研究思路之一。

4.3　精细化流量分类

目前，在流分类领域中的大部分研究都关注对网络协议和应用的分类。很多研究者都使用已有分类方法的变种，或者将多种分类方法相结合的技术，改善之前研究的分类效率及准确率。然而，随着网络的迅速发展，网络流量的规模逐渐扩大，流量特征也越来越复杂，因此很难找到一种准确率达到100%的流分类方法。相比于改善已有分类算法，使其准确率提高1%-2%来说，在一个新的分类层次研究流分类问题显得更加有意义[47]。

精细化流分类是一种新的分类思想，指在细粒度的层次上对网络流量进行分类，比如对某个特定协议上承载的应用类型进行分类，或者对某个特定应用中的不同功能模块进行分类。这种精细化的流分类思想可以帮助研究者更好地分析网络流量的组成，了解用户行为，以便提供更好的网络服务质量。

2008年，Li[48]等人使用了基于有效负载的方法，对HTTP流量按其应用目的（如网页浏览、web应用、爬虫、文件下载、广告、Webmail、多媒体等）进行了细粒度的分类。实验发现，HTTP中的非web浏览应用在HTTP流量中占有相当比例，利用web收发邮件、下载文件等应用被广泛使用，约占HTTP流量的一半。

还有一些研究者对同一种类的应用类型进行了细粒度的分类。2009年，Valenti[49]等人使用了SVM的方法，以特定的一段较短时间内双方交换的包数作为特征，对P2P-TV应用进行细粒度的分类，将其分为PPlive、SopCast、TVAnts和Joost四类。该方法在最坏情况下的识别率也达到81%以上。2011年，Archibald[50]等人通过SVM的方法，消除了http的模糊性，实现了对http协议下应用的细粒度分类。他们使用了统计和谱分析的特征提取方法，选择包长度和包间隔作为分类的特征值，对基于http的三种典型应用

facebook、gmail、youtube进行了分类，分类准确率达到93%。

精细化的流分类思想对分析用户行为也有很大的积极作用。2009年，Bonfiglio[51]等人采用了主动和被动的技术对Skype流量进行深入的分析，通过分析比特率、包间隔时间、包大小等特征值来区分语音和视频通话产生的流量特征。他们通过不同地区在通话期间的单位时间内产生的流量规模对用户行为进行了分析，并惊奇地发现通话时长与所在地的关税政策有所关联。2011年，Park47等人通过使用基于文本检索技术和多签名的流量分类器，将应用按其不同功能模块（如登陆、下载、浏览等）进行分类。实验中主要针对两种P2P应用Fileguri和BitTorrent进行细粒度分类。该分类思想能够帮助相关人员更好地分析网络组成和用户行为。

目前针对精细化分类的研究较少，但随着相关需求的增加，我们认为，精细化分类是未来网络流量分类的重要发展方向之一。

4.4 协议动态变化时的流量分类

当前针对各种网络流量分类的研究，基本都集中已有流量的分类问题上。但从流量分类在信息安全及网络管理的总体技术架构中所处的位置看，流量分类的结果必然伴随着相关网络管理手段的实施。例如，某些电信运营商可能为维护话务量，采用技术手段对Skype流量进行检测、分类，并对这些流量进行管理，为防止网络流量被识别，Skype可能会频繁发布新版本，且不断调整其网络流量的外显行为特征，即Skype的网络协议出现了动态变化，在此情况下，原来的流量分类方法可能失效。

在上述例子中，Skype的流量出现了新的变化。而在其他情况下，某类网络应用的流量不光出现新的变化，还可能进行流量混淆，如2012年2月，TOR（the onion router，最典型的匿名通信网络）推出obfsproxy（obfuscated proxy）模糊代理软件，据称该软件将让SSL或TLS加密流量看起来像未加密的HTTP或即时通讯流量。

目前，针对此类协议动态变化时的流量分类技术的研究还未有效展开，我们认为这是未来网络流量分类的可能发展方向之一。

5 流量分类总结与展望

网络流量分类是信息安全领域的经典问题。各类新型网络应用的不断涌现，使得该问题一直为研究者持续关注。

本文综述了网络流量分类领域的研究方法及研究成果，对基于端口号的分类方法、基于有效负载的分类方法、基于主机行为的分类方法，以及基于机器学习的分类方法进行比较，分别指出它们的优势和不足，及适用的场景。

本文还分析了流量分类面临的四大挑战与技术手段。

高速网络环境中的流量实时分类成果较少，应用价值很大，在这方面还有很大的研究空间。基于机器学习的分类方法由于需要进行大量的统计分析，目前难以适应高速网络环境中的流量实时分类，相关研究成果还处于实验阶段。基于主机行为的分类方法消耗资源较少，较适合应用于高速网络环境下，值得进一步探索。

加密流量分类是流量分类中最具挑战性的问题之一，目前的主要研究思路主要集中在各种行为特征提取及统计分析的方法上。基于主机行为的关联分类方法，辅以主动验证等其它的技术手段，可能是较好的研究思路之一。

精细化流分类能在细粒度的层次上对网络流量进行分类，有助于对网络的精细化管理。目前对它的研究较少，但随着相关需求的增加，精细化分类是未来网络流量分类的重要发展方向之一。

针对网络协议动态变化时的流量分类技术的研究还未有效展开，这可能是未来网络流量分类的方向之一。

参 考 文 献

[1] Fraleigh C, Moon S, Lyles B, et al. Packet-level traffic measurements from the sprint IP backbone [J]. IEEE Trans on Networks, 2003, 17(6): 6-16.

[2] Barakat C, Thiran P, Iannaccone G, et al. Modeling internet backbone traffic at the flow level [J]. IEEE Transactions on Signal Processing (Special Issue on Networking), 2003, 51(8): 2111-2124.

[3] He T, Zhang H, Li X, et al. A methodology for analyzing backbone network traffic at stream-level [C] //International Conference on Communication Technology Proceedings. 2003.

[4] IANA. Internet assigned numbers authority [EB/OL] . http://www.iana.org/assignments/port-numbers.

[5] 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.

[6] Sen S, Spatscheck O, Wang D. Accurate, scalable in network identification of P2P traffic using application signatures [C] //In WWW2004. New York(USA), 2004.

[7] Moore A, Papagiannaki K. Toward the accurate identification of network applications [C] //Proceedings of Passive and Active Measurement Workshop (PAM2005). Boston(USA), 2005.

[8] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification [C] //In MASCOTS. 2006

[9] Karagiannis T, Broido A, Faloutsos M, et al. Transport layer identification of P2P traffic [C] //In ACM/ SIGCOMM IMC. 2004.

[10] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark [C] //Proceedings of the Special Interest Group on Data Communication Conference (SIGCOMM). Philadelphia(USA), 2005.

[11] Nguyen T T T, Armitage G. A survey of techniques for internet traffic classification using machine learning [J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56-76.

[12] 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, 2008, 25(5): 1492-1495.

[13] Frank J. Artificial intelligence and intrusion detection: current and future directions [C] //Proceedings of National 17th Computer Security Conference. Washington D.C.. 1994.

[14] Fisher H D, Pazzani J M, Langley P. Concept formation: knowledge and experience in unsupervised learning [M]. Morgan Kaufmann, 1991.

[15] Erman J, Mahanti A, Arlitt M. Internet traffic identification using machine learning [C] //Proceedings Of 49th IEEE Global Telecommunications Conference (GLOBECOM 2006). San Francisco(USA), 2006.

[16] McGregor A, Hall M, Lorier P, et al. Flow clustering using machine learning techniques [C] //In PAM 2004. Antibes Juan-les-Pins(France), 2004.

[17] Dempster A, Paird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[18] Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning [C] //In IEEE 30th Conference on Local Computer Networks(LCN 2005). Sydney(Australia), 2005.

[19] Cheeseman P, Strutz J. Bayesian classification (autoclass): theory and results [J]. In Advances in Knowledge Discovery and Data Mining. USA: AAAI/MIT Press, 1996.

[20] Netmate [EB/OL]. http://sourceforge.net/projects/netmate-meter/, as of August 14, 2007.

[21] Erman J, Mahanti A, Arlitt M, et al. Identifying and discriminating between web and peer-to-peer traffic in the network core [C] //In WWW' 07: Proc. 16th international conference on World Wide Web. Banff(Canada): ACM Press, 2007: 883-892.

[22] Erman J, Arlitt M. Traffic classification using clustering algorithms [C] //In MineNet' 06: Proc. 2006 SIGCOMM workshop on Mining network data. New York(USA): ACM Press, 2006: 281-286.

[23] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] //In 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD 96). Portland(USA), 1996.

[24] Kotsiantis S B. Supervised machine learning: a review of classification techniques [J]. Informatica, 2007, 31: 249-268.

[25] Murthy S K. Automatic construction of decision trees from data: a multi-disciplinary survey [J]. Data Mining and Knowledge Discovery, 1998, 2(4): 345-389.

[26] Quinlan J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[27] Quinlan J R. C4.5: Programs for machine learning [M]. California: Morgan Kaufmann, 1993.

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

[29] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20: 273-297.

[30] 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, 46(3): 407-414.

[31] Li Z, Yuan R X, Guan X H. Accurate classification of the internet traffic based on the SVM method [C] //Proceedings of the 42th IEEE International Conference on Communications (ICC 2007). 2007.

[32] Auld T, Moore A W, Gull S F. Bayesian neural networks for internet traffic classification [J]. IEEE Trans. Neural Networks, 2007, 18(1): 223-239.

[33] Park J, Tyan H R, Kuo C J. GA-based internet traffic classification technique for QoS provisioning [C] //Proceedings of 2006 International Conference on Intelligent Information Hiding and Multimedia Signal. Pasadena(California), 2006.

[34] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review, 2006, 36(5): 5-16.

[35] Erman J, Mahanti A, Arlitt M, et al. Semi-supervised network traffic classification [J]. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) Performance Evaluation Review, 2007, 35(1): 369-370.

[36] Li X, Qi F, Xu D, et al. An internet traffic classification method based on semi-supervised support vector machine [C] //2011

IEEE International Conference on Communications (ICC). 2011.

[37] Gu C J, Zhang S Y, Xue X Z. Encrypted internet traffic classification method based on host behavior [J]. International Journal of Digital Content Technology and its Applications, 2011, 5(3): 167-174.

[38] Bernaille L, Teixeira R. Early recognition of encrypted applications [C] //Proceedings of the Eighth Passive and Active Measurement Conference (PAM' 07). 2007.

[39] Hirvonen M, Sailio M. Two-phased method for identifying SSH encrypted application flows [C] //In 7th International Wireless Communications and Mobile Computing Conference (IWCMC). 2011.

[40] Dusi M, Este A, Gringoli F, et al. Using GMM and SVM-based techniques for the classification of SSH-encrypted traffic [C] //Proceedings of the 44th IEEE International Conference on Communication(ICC' 09). 2009.

[41] Alshammari R, Zincir-Heywood A N. A flow based approach for SSH traffic detection [J]. Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on,2007: 296-301.

[42] Alshammari R, Zincir-Heywood A N. Machine learning based encrypted traffic classification: identifying SSH and skype [C] //Proceedings of the 2009 IEEE Symposium on Computation Intelligence in Security and Defense Applications (CISDA 2009). 2009.

[43] Bar-Yanai R, Langberg M, Peleg D, et al. Realtime classification for encrypted traffic [C] //In SEA, 2010: 373-385.

[44] Tan X B, Su X Q, Qian Q M. The classification of SSH tunneled traffic using maximum likelihood classifier [C] //In 2011 International Conference on Electronics, Communications and Control (ICECC). 2011.

[45] Yildirim T, Radcliffe P. VoIP traffic classification in IPSec tunnels [C] //In 2010 International Conference on Electronics and Information Engineering (ICEIE). 2010.

[46] Hjelmvik E, John W. Breaking and improving protocol obfuscation [R]. Technical report, Chalmers University of Technology, 2010.

[47] Park B, Hong J W, Won Y J. Toward fine-grained traffic classification [J]. Communications Magazine, 2011, 49(7): 104-111.

[48] Li W, Moore A W, Canini M. Classifying HTTP traffic in the new age [C] //In ACM SIGCOMM. Poster Session, 2008.

[49] Valenti S, Rossi D, Meo M, et al. Accurate, Fine-grained classification of P2P-TV applications by simply counting packets [C] //In Traffic Measurement and Analysis (TMA). Springer-Verlag LNCS 5537, 2009.

[50] Archibald R, Liu Y L, Corbett C, et al. Disambiguating HTTP: classifying web applications [C] //In 2011 7th International Wireless Communications and Mobile Computing Conference (IWCMC). 2011.

[51] Bonfiglio D, Mellia M, Meo M, et al. Detailed analysis of skype traffic [J]. IEEE Transactions on Multimedia, 2009, 11(1): 117-127.