P2P流量识别

王春枝　陈宏伟　叶志伟 ◎ 著

科学出版社

P2P 流量识别

王春枝　陈宏伟　叶志伟　著

科学出版社
北京

内 容 简 介

随着计算机网络的发展与对等网络的应用，P2P 流量已经成为网络流量中最主要的组成部分。P2P 应用带来网络发展繁荣的同时，也带来矛盾和挑战，有必要对 P2P 流量的进行控制和管理，首要问题便是实现对 P2P 流量的识别。

本书主要包括如下内容：P2P 流量识别方法概述，基于 DPI 抽样的 P2P 流量识别，基于 DPI 信任抽样的 P2P 流量识别，基于 DFI-SVM 模型的 P2P 流量识别，基于 DPI/DFI 结合的 P2P 流量识别，基于 CS-PSO 和 SVM 的 P2P 流量识别，基于菌群优化算法和小波 SVM 的 P2P 流量识别，基于人工蜂群算法和小波 SVM 的 P2P 流量识别。

本书可作为计算机科学与技术、网络工程与信息安全等相关专业研究生及高年级本科生的教材，也可作为科研人员的参考书，还可作为研究生、博士生及教师论文写作的参考书。

图书在版编目（CIP）数据

P2P 流量识别/王春枝，陈宏伟，叶志伟著. —北京：科学出版社，2016

ISBN 978-7-03-050702-0

Ⅰ. ①P… Ⅱ. ①王… ②陈… ③叶… Ⅲ. ①计算机网络-流量-识别-研究 Ⅳ. ①TP393

中国版本图书馆 CIP 数据核字（2016）第 278344 号

责任编辑：戴 薇 王 惠 / 责任校对：王万红
责任印制：吕春珉 / 封面设计：东方人华平面设计部

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京京华虎彩印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2016 年 12 月第 一 版 开本：B5（720×1000）
2016 年 12 月第一次印刷 印张：13
字数：247 000

定价：60.00 元

（如有印装质量问题，我社负责调换〈京华虎彩〉）

销售部电话 010-62136230 编辑部电话 010-62135397-2052

前　言

Internet 本身是世界上最大的非集中式互联网，但是 20 世纪 90 年代所建立的一些网络应用系统却是完全的集中式系统。传统的 Internet 应用采用客户-服务器模式，所有的内容与服务都在服务器上，客户向服务器请求内容或服务。对等网络（Peer-to-Peer Network，P2P）的出现打破了传统的客户-服务器模式。在对等网络中，每个结点的地位都是相同的，每个结点都有一些资源（如处理能力、存储空间、网络带宽、内容等）可以提供给其他结点，每个结点也都可以使用其他结点的资源，结点之间直接共享资源，不需要服务器的参与。也就是说，对等网络中的结点具备客户和服务器的双重特性，可以同时作为服务的使用者和提供者。它能充分利用互联网的边缘资源，即用户的计算能力、存储能力和带宽，甚至计算机硬盘的内容。如今 P2P 流量已经成为网络流量中最主要的组成部分。P2P 应用已经普及，影片下载、在线视频、文件下载等均采用这项技术。通过 P2P 技术，文件的下载速度、视频的观看效果均有极大的改善。然而，P2P 技术的飞速发展犹如一把双刃剑，一方面丰富了网络中的应用形式，另一方面也带来了许多负面影响。

首先，P2P 技术给网络的运营和管理带来一些不利影响。①带宽占用严重，扩容压力不断增加。据权威部门调查统计，P2P 已经取代 HTTP 等传统应用成为 Internet 上流量最大的一类应用，特别是我们国家，其占用网络带宽 50%～90%，并且其流量以每年 350%的速度增长，常常是运营商扩多少带宽，P2P 应用就占用多少；此外，P2P 还占用 HTTP 等端口的带宽，导致网页浏览等正常的互联网业务受到影响，一个 P2P 软件足可以拖垮任何一个单位的所有带宽，造成网络严重堵塞。P2P 流量消耗巨大的网络带宽，尤其是国际带宽，使得网络基础设施不堪重负，运营商苦不堪言。②语音、核心业务流失。P2P 技术使得支撑全球 7000 万用户的 Skype 只需少量 PC 服务器和成本即可实现业务运营，其语音通话质量堪比传统电话服务质量。③可增值业务的影响。目前，P2P 应用大多以免费形式提供，如 PPLive、PPStream 等软件，其服务质量直逼运营商重点推广的增收应用，如 IPTV 等。

其次，P2P 技术带来一些新的安全和社会问题。①国家信息安全问题：开放式的结构使得色情、暴力等不健康内容无限制分发，目前 P2P 网络上的反政府言论和色情信息泛滥。②公司知识产权问题：合法用户可以利用 P2P 软件将公司内

部网络的数据传播出去，对企业造成严重威胁。③系统安全问题：P2P 文件共享没有文件存储中心，使得文件共享的集中可控制性、可管理性下降，给病毒、网络攻击、恶意软件在各个客户端之间传播提供温床；如果感染恶劣病毒，会导致单位的整个网络瘫痪，造成不可挽回的损失。④P2P 下载大量的网络资源，其信息和业务基本上没有版权，在非法地使用各种资源，这严重影响了内容厂商的合法权利和制作内容的积极性，若进一步发展将可能产生严重的法律问题。

P2P 应用带来网络发展繁荣的同时，也带来矛盾和挑战。为使 P2P 技术为人们生产生活提供更好的服务，就有必要对 P2P 流量进行控制和管理，首先要解决的问题是实现对 P2P 流量的识别。自 P2P 技术出现以来，无论是学术界还是网络运营商，都对 P2P 流量的识别和控制给予了极高的重视。在宽带用户和视频网站用户数量迅猛增长的今天，P2P 流量识别技术的研究和发展有利于对 P2P 网络流量进行有效的控制和管理；有利于用户关键业务的使用，为用户提供 QoS 保证；有利于制止非法内容在 P2P 网络中的传播；有利于合理利用互联网基础设施，解决因 P2P 流量而造成的网络带宽拥挤问题；有利于改善网络安全方面的问题，在保障 P2P 应用健康、规范和有序的发展方面具有非常重要的实用价值。

本书主要从如下几个方面展开论述：

第 1 章主要对当前的 P2P 流量识别方法进行概述。主要介绍基于端口号的识别方法、基于深层数据包（DPI）的识别方法、基于流量特征（DFI）的识别方法和基于机器学习的识别方法。

第 2 章和第 3 章主要研究基于 DPI 的 P2P 流量识别方法，给出字符串的几种匹配算法，以及 DPI 抽样模型和策略。在此基础上，提出基于信任抽样的 P2P 流量识别方法，提出基于信任度抽样和二阶贝叶斯信任抽样策略，并给出相应的信任抽样算法，可以较为准确地预测后续抽样周期 P2P 流量比例的波动程度，可以降低样本冗余。

第 4 章主要研究基于 DFI 的 P2P 流量识别方法。结合 SVM 算法的 P2P 流量特征选择，建立基于 SVM 的可行的 DFI 检测模型，将单个流行为特征和多个流之间特征结合来识别 P2P 流量。

第 5 章主要研究 DPI 和 DFI 相结合的 P2P 流量识别方法。将 DPI 与 DFI 通过协同策略机制进行结合，实现在系统中同时运行 DPI 和 DFI 程序，真正达到互补运行的目的，并取得更加准确的识别效果、更加广泛的适用范围和更加优秀的可扩展性。

第 6 章至第 8 章主要研究基于智能算法参数优化 SVM 的 P2P 流量识别方法。提出了融合杜鹃搜索和粒子群算法的 P2P 流量特征选择，基于人工蜂群算法的 P2P 流量特征选择方法，在众多特征集合中选择出具有最佳分类性能的特征子集；

提出采用融合杜鹃搜索的粒子群算法对支持向量机参数进行优化，基于菌群优化算法的 SVM 参数优化，以及基于人工蜂群算法对支持向量机参数进行优化，以避免 SVM 参数优化中计算费时、易陷入局部最优的问题。

参加本书相关专题研究和书稿撰写工作的有周昕、尤方萍、姜伟、宗欣露等老师，以及喻东阳、邓来、李沁沄、王泽琪、张会丽等研究生。徐慧、刘伟、严灵毓等老师参加了校对工作。

本书的撰写得到了国家自然科学基金、湖北省自然科学基金、武汉市晨光计划基金的资助。此外，在本书撰写过程中，参考了国内外相关研究成果，在此谨向相关作者表示衷心的感谢。

由于著者水平有限，书中的疏漏及不妥之处在所难免，敬请广大读者批评指正。

著　者

2016 年 11 月于武汉

目　录

第 1 章　P2P 流量识别方法概述

P2P 流量识别问题得到了越来越广泛的关注，围绕这一问题，很多专家学者从不同的角度提出了许多优化算法及识别方法。本章主要简单介绍 P2P 网络的发展历程，并分析 P2P 网络的特点，针对 P2P 流量特征分析 4 种 P2P 流量识别技术，并重点分析机器学习方法的模型，为后续的介绍打下一定的基础。

1.1　P2P 网络的发展及特点

1.1.1　P2P 网络的发展

P2P 技术是一种网络中的计算机间彼此无须中间媒介而互连通信、分享资源与服务的技术。早在 1979 年的新闻组（Usenet/NewsGroup）出现时，网络用户间就可以相互交流，阅读各类信息且可以参与讨论。而 1994 年 WWW 的出现使得互联网的发展进入全球领域，而 C/S 模式也成为主要的服务模式。虽然 C/S 服务模式能充分发挥客户端的作用，使用户从网站上获取自己所需要的网络信息和资源，但由于需要服务器的响应，从某一层面限制了信息的流动，阻碍了用户间的直接通信。

而 P2P 技术应运而生成为研究热点后，对 C/S 服务模式产生了重大影响，在对等网络计算、数据搜索、协同处理、即时通信、共享信息等许多方面发挥了明显的优势，真正方便了用户的使用。P2P 网络中的每个结点具有相同的地位，既能提供资源、分享服务，也能自由接收网络中其他用户提供的资源，充分享受服务。从技术的角度来看，P2P 技术使中央服务器作为网络核心应用的这种模式转向边缘化，将其扩散到网络边缘和终端设备上，网络构架也由集中式向分布式转变；从市场的角度来看，直接交互的 P2P 网络结点颠覆了传统电信的管理、控制和收费模式，使用户间的信息互通、资源共享更直接、更方便、更顺畅[1]。从某种角度上来说，P2P 技术是计算机网络真实本质的回归，是旧技术的新应用模式。

P2P 网络发展至今，按照它的体系结构大致可以分为 3 个发展阶段：集中式 P2P 网络、分布式 P2P 网络和混合式 P2P 网络[2]。

1. 集中式 P2P 网络

第一代 P2P 网络系统采用的是集中式结构，其网络架构如图 1.1 所示。它在结构上拥有中心服务器，负责保存共享资源的目录信息；要求所有结点都要连接

到中心服务器上才能实现信息共享，因而又称此种结构为集中式目录结构。早期的 P2P 应用是采用固定端口进行通信的，所以它的识别也是针对端口来检测的。

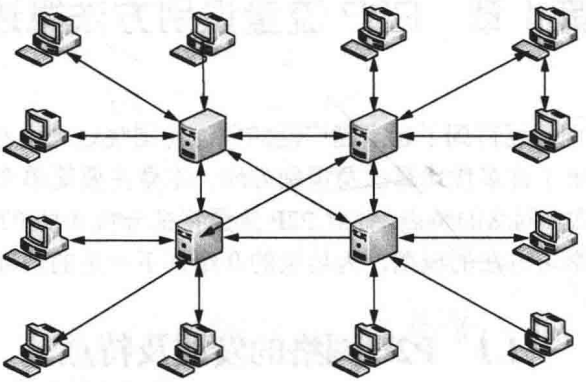

图 1.1　集中式网络架构图

这种网络最大的优点是响应快且便于管理，快速的搜索可以减少排队的响应时间。其缺点也很明显，由于采用集中式的中心服务器，且所有用户端均需与服务器相连，服务器一旦崩溃就会影响整个网络。而随着动态端口和加密技术的出现，固定端口的识别技术也逐渐淘汰，因而这种网络也在发展中被取代[3]。

2. 分布式 P2P 网络

第二代 P2P 网络系统无中心服务器，其网络架构如图 1.2 所示。网络中的结点均为对等端，在对等端之间实现相互通信与资源共享，实现了分布式的目录管理。随着 P2P 应用的发展，出现了随机的动态端口、伪装端口，因而对固定端口的识别检测已无法适用，采用的是基于应用层的数据检测技术[4]，通过检测数据包的载荷来识别 P2P 应用，此种识别技术将在后面详细介绍。

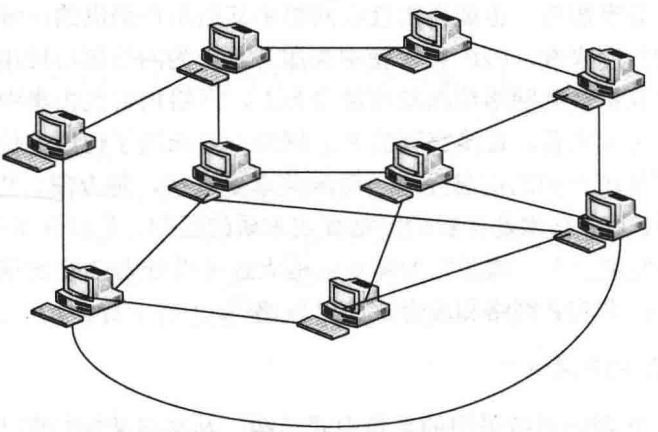

图 1.2　分布式网络架构图

这种网络是真正的分布式网络，可扩展性良好。由于没有中央服务器，不会因中心故障而导致整个网络的瘫痪。但与之相应的缺点也很明显，因其搜索遍布全网，相应时间较长，造成网络性能下降，影响了用户的使用。

3. 混合式 P2P 网络

第三代 P2P 网络系统采用了混合式架构，其网络架构如图 1.3 所示。它结合了集中式网络与分布式网络的优点：分布式的中心服务器是由 P2P 应用随机选择的超级起点（Super-Peer）来与其他结点互连，实现快速搜索与资源共享的[5]。网络中的结点本身还是对等端，只是根据具体情况来随机选择一组结点作为超级结点，为另外的普通结点提供目录服务。第三代 P2P 系统中会采用 SSL 协议来进行加密技术的处理，因而又出现了基于流量特征的识别方法来解决这个问题。

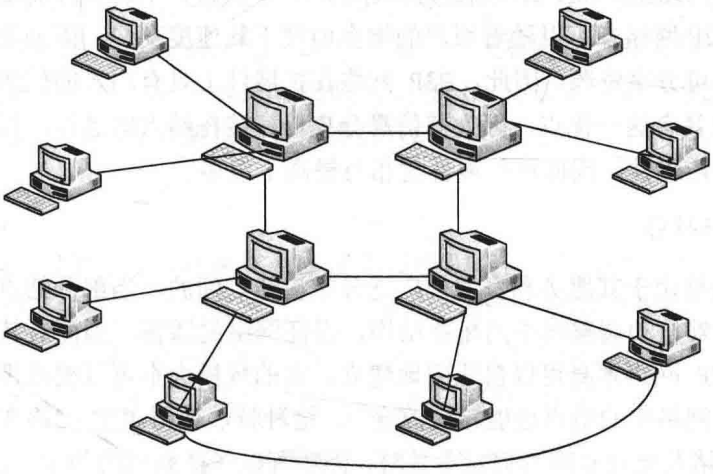

图 1.3　混合式网络架构图

这种网络模式运用了分布式的超级结点来取代中心服务器，又能利用分层的快速检索节省检索时间，因而其既能缩短排队的响应时间，又能提高检索的性能[6]，不会因为中心服务器的崩溃致使整个网络的瘫痪，所以成为当前普遍使用的 P2P 网络。

1.1.2　P2P 网络的特点

P2P 网络的特点是针对网络本身以及各种 P2P 应用而言的，P2P 应用包含即时通信、文件分享、IP 电话、分布式搜索、视音频播放、移动 P2P 应用、电子商务等[7]。P2P 网络主要具有以下几个特点。

1. 非中心化

非中心化是 P2P 网络最基本的特点。P2P 网络中的信息传输和服务传送都无

须经过媒介，直接发生在对等的结点之间，提高了网络的性能。即便对于混合的 P2P 网络结构，虽然在检索和定位方面需要建立超级结点来节省响应时间，但结点本身仍为对等的，超级结点也是随机选取并据情况与一组普通结点相连，信息的分享与交换仍然无中间结点。良好的可扩展性和健壮性是建立在非中心的网络特点基础之上的。

2. 可扩展性良好

P2P 网络由于非中心化，相对于传统的 C/S 服务模式，能够在用户增长与服务的需求量增加时，同步地增强资源供给与服务能力[8]。而传统的系统因为在用户增多时要增强中心服务器的性能，必然要增大系统的开销来组建更大的网络，从而限制了网络的扩展。比如采用传统的 FTP 下载时，不能始终满足用户的下载需求；而 P2P 网络却可以随着用户的增多而使下载速度变快，因为用户的增长提供了更多的可分享资源。因此，P2P 网络在扩展性上具有明显的优势。混合式的 P2P 网络也具有这一优点，因为其信息分享仍直接在结点间进行，极大地降低了对服务器的依赖性，因而可扩展性就相对提高了很多。

3. 健壮性强

P2P 网络由于其服务和资源是广泛分布在结点间的，当部分结点或者网络出现故障时能够自动调整网络的拓扑结构，保证网络的通畅，因而对其他结点的影响较小。P2P 网络本身可以自适应地建立，它的规模大小可以根据带宽和负载自由地调整，网络中的结点也能相应改变[9]。这种特性使得 P2P 网络在面对网络中断、网络拥堵及结点故障等突发事件时，仍然可以保持系统的稳定与服务的顺畅，从而具有良好的健壮性。

4. 性价比高

P2P 网络能合理利用结点的资源与服务，将分布在边缘的结点也包含在内，使资源存储与计算工作广泛分布于各个结点，因而充分利用了网络闲置的结点与资源，可以在计算性能与存储能力不断增长的网络中保证在耗费成本低的状况下提供更好的服务[10]。

1.2 P2P 流量识别方法

随着 P2P 网络的广泛应用与 P2P 技术的蓬勃发展，P2P 流量产生了不可忽视的影响。对于用户而言，P2P 网络中的流量具有上下行流量对称的特点[11, 12]，因而易占用大量的网络带宽资源，会在传输过程中造成网络拥堵的现象；对于企业

而言，P2P 网络的非集中的特点给管理方面带来了诸多不便，且动态端口或伪装的随机端口也对企业进行有效的监控管理提出了挑战，而网络的安全隐患是企业担忧的另一大问题；对于网络运营商而言，各种 P2P 应用使网络管理增加了难度，无法对网络流量进行有效的控制和管理，也会导致网络的总体性能和服务质量相对下降，因而需要识别网络流量，并对其进行规范与管理，以维护良好的网络环境[13, 14]。另一方面，网络运营商的管理与维护也需要考虑耗费成本，因而一种有效的 P2P 流量识别技术成为各种研究的关键所在。

针对 P2P 流量所具有的上下行流量对称和非均衡分布的特性、动态端口和加密技术，以及流量本身具有很强的隐蔽性等，对 P2P 流量进行有效的识别需要不断加强，因而也成为 P2P 技术研究与应用的重点[15-17]。

从技术方面看，现有 P2P 流量识别的主要技术大致可分为基于深层数据包识别方法（Deep Packet Inspection，DPI）、基于流量特征的识别方法（Deep Flow Inspection/Transport Layer Identification，DFI/TLI）及基于机器学习的流量识别方法[18-20]。除此之外，还有最初的基于端口号的识别技术。

1.2.1　基于端口号的识别方法

基于端口号的识别方法是在 P2P 应用发展初期使用固定端口或默认端口时所采用的识别方法，这种方法简单、直接、解析速度快，通常只需分析 IP 包头找出端口号，然后匹配端口表中的端口号[21]，若一致则表明是 P2P 流量，否则不是。表 1.1 所示为 11 个常用 P2P 应用的端口号及其使用的协议。

表 1.1　常用 P2P 应用的端口号和协议

P2P 应用	端口号	协议
eMule	4662/4672	TCP/UDP
Gnutella	6346～6347	UDP
eDonkey	4662	TCP
Morpheus	6346/6347	TCP/UDP
KuGoo	7000	UDP
BitTorrent	6881～6889	TCP/UDP
Skype	4048	TCP
Yahoo Messenger	5000	UDP

这种识别方法的实现虽然简单，只需了解端口号就可以匹配检测是否为 P2P 流量。但随着 P2P 技术的发展，P2P 应用不断增多，预设端口表中的协议不断增加，端口号信息也要不断更新，而且随着动态端口的出现，许多 P2P 应用采用非固定端口或者伪装成服务端口来躲避流量限制，还有加密技术的使用，都使得采用固定端口的识别能力受限，无法有效地识别 P2P 流量[22]。

1.2.2　基于深层数据包的识别方法

基于深层数据包的识别方法（DPI）是一种基于应用层的数据检测方法，由于动态的随机端口和加密技术使得传统的基于端口的识别方法不再有效，DPI 针对动态端口的特性，通过提取并分析应用层上数据的 P2P 载荷所包含的协议特征值，建立一个特征库[23, 24]。由于不同数据包的特征字符串不同，对待检测的数据包进行协议层分析，匹配特征库的特征字符串，判断是否为 P2P 数据。表 1.2 所示为 6 种常见 P2P 应用的协议特征字符串。

表 1.2　常见 P2P 应用的协议特征字符串

P2P 应用	协议	特征字符串
eDonkey	TCP/UDP	Oxe319010000 Ox53f010000
BitTorrent	TCP	"Oxl3Bit" "GET TrackPak"
Fasttrack	TCP	"Get/.hash" "GIVE"
Gnutella	TCP	"GNUT""GIV" "GND"
Ares	TCP	"GET hash:" "Get shal:"
Direct Connect	UDP	"$Pin" "$SR"

这种识别方法能针对已知协议的 P2P 流量进行精确的识别判断，识别精度高，误判率较低，维护也较为容易。但其缺点也很明显，由于每次匹配判断时都需解开数据包并分析协议的特征字符串，计算速度较慢；若数据包的特征字符串复杂，则会加重计算工作量。另外，还有针对新的 P2P 协议来不断更新特征库的协议字符串，目前很多 P2P 应用对数据包的内容进行了加密处理或使用传输分块机制，使得获取数据包的特征字符串变得异常困难；而 DPI 技术也无法识别未知的 P2P 应用，因此 DPI 的可扩展性很差，无法对使用加密技术的 P2P 应用进行有效的识别[25]。

1.2.3　基于流量特征的识别方法

基于流量特征的识别方法（DFI）能解决 DPI 无法识别加密技术、扩展性差的不足，因为 DFI 不涉及高层的协议，计算量相对较小，速度比 DPI 快，且能识别加密的 P2P 应用。不同的 P2P 应用会表现出不同的流量特征，因此可以通过分析 P2P 的流量特征来进行识别。流量特征可以是报文中的特征字符串，也可以是应用行为特征，具体包括应用的连接数、IP 的连接模式、上行流量和下行流量的

比例关系、数据包发送的频率等。统计数据的流量特征，然后进行分析，以判断该流量是否为 P2P 流量[26]。

虽然 DFI 较 DPI 具有可扩展性好、可识别加密数据的优点，但也有很多不足之处：首先，由于所分析的流量特征不具备辨别应用层协议类型的能力，DFI 的分类性能不强；其次，由于不同的应用可能会有相同的流量特征，因而 DFI 的准确性差，误判率高，需要结合其他识别方法或检测技术来进行；最后，DFI 在数据包传输过程中，易发生不对称路由以及数据包的丢失和重传现象，因而也会导致流量特征不能被准确地区分，健壮性较差。

1.2.4　基于机器学习的识别方法

机器学习是计算机模仿人类学习的一种方法，计算机能不断学习新的经验来达到完善自身学习能力和处理能力的目的，目前在生物、医学、商务、科技和军事等领域都有着广泛的应用。而基于机器学习的识别方法，是将流量的样本依据一些流量特征进行训练，然后根据训练的结果产生一个分类模型，再对实时输入的流量依据训练模型进行测试，判断检验出样本的分类情况。训练所依据的流量特征是要从一个或多个数据包中计算并统计出的，包括流持续时间、传输时上下行包长均值、总字节数、包到达时间等流特征[27]。

基于机器学习的识别方法可不依赖端口信息，不用检查 IP 包的负载，也不涉及用户的隐私，相对于前面 3 种方法而言具有明显的优势：

① 识别准确性较好，且对于动态端口的检测也适用。

② 能够识别加密的 P2P 流量，能改善 DPI 无法识别加密数据包和未知 P2P 流的缺陷，且对于应用层和网络层的加密都能检测识别，这是 DFI 也无法做到的。

③ 不用检测 IP 包的载荷，因而计算复杂度相对 DPI 而言大大降低。

表 1.3 所示为 4 种 P2P 流量识别技术的比较，从中可以看出，基于机器学习的 P2P 流量识别方法能综合其他 3 种方法的优点，是最有潜力的，所以基于机器学习的识别方法是研究的热点。

表 1.3　4 种 P2P 流量识别技术的比较

比较项目	基于端口的识别方法	基于深层数据包的识别方法	基于流量特征的识别方法	基于机器学习的识别方法
准确性	差	好	一般	一般
可扩展性	差	差	好	一般
动态端口检测	不支持	支持	支持	支持
应用层加密检测	支持	不支持	支持	支持
网络层加密检测	不支持	不支持	不支持	支持

基于机器学习的识别方法仍然存在着下述问题：准确性一般，P2P 流量识别的分类正确率有待提高。因而，如何提高基于机器学习的识别方法的分类正确率