基于 DPI 和机器学习的网络流量分类方法*

李国平[1]，王 勇[1]，陶晓玲[2]

（1. 桂林电子科技大学 计算机科学与工程学院，广西 桂林 541004；
2. 桂林电子科技大学 信息与通信学院，广西 桂林 541004）

摘 要：网络流量分类是实现网络管理的重要技术之一，但是单一的基于 DPI 或是机器学习的分类方法分类精确度低。提出了一种基于 DPI 和机器学习相结合的网络流量分类方法。该方法采用 DPI 检测已知特征的网络流量，利用机器学习方法辅助分析未知特征以及加密的网络流。实验表明该方法能够提高网络流量分类的精确度。

关键词：流量分类；深度包检测；机器学习；朴素贝叶斯

中图分类号：TP393 **文献标识码**：A **文章编号**：1673-808X(2012)02-0140-05

A novel method for network traffic classification based on DPI and machine learning

Li Guoping[1], Wang Yong[1], Tao Xiaoling[2]

(1. School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China;
2. School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Network traffic classification is one of the important technology to implement network management, but the method for network traffic classification based on single DPI or machine learning is very poor. An algorithm based on DPI with machine learning for network traffic classification was proposed. Unencrypted network traffic was detected by DPI and others classified by maching learning. The experimental result shows that this method can get a more accuracy classification result.

Key words: network traffic classification; deep packet inspection; machine learning; Naïve Bayesian

随着网络技术以及网络应用的飞速发展，网络用户对网络连接速度和质量的要求越来越高。因此，通过有效的技术手段，管理和控制各种网络业务流量，区分不同服务，提供不同质量保障，满足用户的业务需求成为当前运营商面临的挑战之一。网络流量分类为区分不同应用业务流量提供了一种有效的技术手段。通过对网络流量的应用进行分类、识别和区分，从而对不同应用的流量进行细分，针对不同层次的用户提供有区分的网络服务，提高网络服务质量和用户满意度。

1 网络流量分类方法

网络流量分类是指基于 TCP/IP 协议的 Internet 中，按照网络的应用类型（WWW、FTP 和 P2P 等），将网络通信产生的双向 TCP 流或 UDP 流进行分类[1]。目前常用的 3 种方法：1）基于端口号分类；2）基于应用层特征字段分类；3）基于流统计特征的机器学习方法分类。

1.1 基于端口号的分类方法

基于端口号的分类方法是根据国际互联网代理

* 收稿日期：2012-03-05
基金项目：国家自然科学基金(61163058)；广西自然科学基金(2011GXNSFB018076)
通信作者：王勇(1964－)，男，四川阆中人，教授，博士，研究方向为网络安全。E-mail: wang@guet.edu.cn
引文格式：李国平，王勇，陶晓玲. 基于 DPI 和机器学习的网络流量分类方法[J]. 桂林电子科技大学学报，2012，32(2)：140-144.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

成员管理局建议的非强制端口号来分类不用的应用类型[2]。在互联网的早期可以通过 80、21 等端口号识别 HTTP、FTP 等应用程序，并且 P2P 应用兴起的早期，也可以通过 6346～6347、6881～6889 等端口识别 Gnutella、BitTorrent 等 P2P 应用。但是，随着网络新兴业务的不断出现以及 P2P 技术的不断发展，大量的新兴应用为了穿越防火墙或者躲避其他的封堵策略，开始使用端口伪装和动态端口技术。因此，基于端口的分类方法具有很大的局限性，分类结果很不准确[3]。

1.2　基于特征字段的分类方法

不同的应用程序都有各自不同的特征字段，这些特征字段可能是特定的字符串或者 bit 序列。深度包检测(deep packet inspection，简称 DPI)技术通过对网络数据包分解，依据模式匹配等方法检测网络交互过程或数据传输过程中 IP 数据包的载荷内容，根据不同的载荷内容确定应用程序的类型。DPI 检测技术能够快速地检测出网络流的应用类型且不受端口变更的影响，具有很高的准确性。Sen 等[4]根据数据包中有效载荷特征字段对 P2P 流的识别进行了研究，并验证了该方法具有较高的准确性、健壮性以及实时性。

1.3　基于流统计特征的机器学习分类方法

网络数据流由于其应用协议的不同，在数据流持续时间、数据包长度、数据包发送频率以及数据包速率等方面表现出不同的特点[5]。依据网络流表现出来的这些特征，利用数据挖掘中的分类技术，通过机器学习的方法能够实现很好的流量分类。贝叶斯分类[6]、支持向量机(support vector machine，简称 SVM)[7]、C4.5[8]等基于流统计特征的机器学习算法已被引入到网络流量分类的应用。

1.4　网络流量分类方法的比较

首先，基于端口的流量分类方法，实现原理简单，不需要复杂的计算分析，能够满足高速网络快速分类的要求。但由于新的网络应用的出现，尤其是 P2P 应用程序的发展，大多采用随机端口以及伪装端口等技术手段来保护自己的网络通信，这降低了基于端口的流量分类方法的准确性，该分类方法也在逐步退出历史舞台。

基于特征字段的分类方法针对载荷中的有效字段进行检测，不依赖应用程序的端口设置，能够很好地识别网络流，并且能够识别具体的网络应用，检测准确率高。该种方法可以通过只检测网络流的前几个特定数据包识别网络应用，检测速度快，能够实现快速的识别网络流量。但由于这种方法依赖于应用协议的特征字段，它只能对已知的应用进行识别，无法识别新型应用。另外，该种方法也不能识别载荷加密的网络流量。

基于流统计特征的机器学习分类方法利用数据挖掘中的分类技术，通过机器学习的方法实现流量分类，克服了前 2 种方法无法解决的难点，不受端口变动、协议特征变化的影响，能够识别新的应用。但是，这类基于机器学习的方法无论是贝叶斯分类还是基于支持向量机的分类方法，不能识别具体的应用、需要根据多个数据包形成流才能检测流量类型，检测表现相对滞后，而且容易受到流长度的影响，对小于特定时长的流误诊率高。另外，该分类方法的准确性容易受到网络动态变化以及流量属性集合的影响，并且这类方法的缺点是计算量大，不适合于高速网络的实时流量分类。

在分析和比较了上述几种流量分类方法后，依据基于特征字段的分类方法和基于流统计特征的机器学习方法原理，提出了一种基于 DPI 技术和机器学习的网络流量分类方法。

2　基于 DPI 和机器学习的网络流量分类方法

2.1　DPI 技术和机器学习分类方法

2.1.1　DPI 技术

DPI 技术是基于特征字段进行检测的一种技术，它通过深入读取 IP 包载荷内容对应用层信息进行重组，从而得到整个应用层的内容，然后根据已有的特征库对数据流内容进行扫描检测，从而识别具体的应用数据。深度包检测要求设备能够快速分析、检测以及重组应用数据，以避免给应用带来过大时延。

DPI 技术一般由 2 部分组成，一是扫描算法，二是特征库[9]。扫描算法是对 IP 包载荷的内容与特征库进行逐字节匹配，常用于 DPI 技术的字符串匹配算法有 AC 算法、WUMANBER 算法以及 SBOM 算法[10]。DPI 检测技术类似于杀毒软件中的特征匹配技术。杀毒软件把扫描的当前文件与自身的病毒库进行逐字节匹配，若发现相同的特征码，则判定病毒

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的类型以及名称。这种方式能够根据特征库来精确识别网络流，并且可以精确到网络流所属的具体应用程序，检测精确度高。但 DPI 无法识别那些特征码还未记录进特征库中的应用流量，滞后于新应用的发布，并且也不能识别加密的网络数据流。

2.1.2 机器学习分类方法

基于机器学习的网络流量分类方法的核心是计算机程序随着学习经验的积累能够不断地完善自我性能，从而完成常规方法不能完成的任务。在网络流量分类中，这种先验知识可以是网络流量呈现出的不同特征以及人的监督信息。在网络流量分类中，选择优秀的机器学习算法，能够很好地利用先验知识来完成流量的分类工作。

基于机器学习的网络流量分类方法流程如图 1 所示。首先利用训练数据集训练分类模型，然后根据训练的模型建立分类器，待建立分类器后可以实现对流量的分类工作。

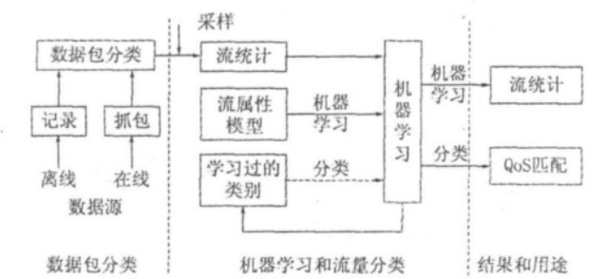

图 1 基于机器学习的流量分类流程图

Fig. 1 The flow chart of traffic classification based on machine learning

2.2 算法的设计

该网络流量分类方法采用 DPI 技术和机器学习相结合的方式实现网络流量分类，基本设计思想如图 2 所示。

1)DPI 检测阶段根据加载的协议特征库对网络数据流进行模式匹配检测，若能匹配相应协议字段，则对流量进行识别标识；否则，做未识别标识。

2)当网络流经过 DPI 识别后，流统计特征采集模块设置固定的采集时间，开始采集报文的特征信息。

3)当特征采集模块采集完成后，将未识别网络流统计特征信息交给已训练好的流量分类器进行识别。

4)针对 DPI 已经识别的网络流，将其流统计特征信息加入到训练样本库，作为训练样本集对分类器进行再学习。

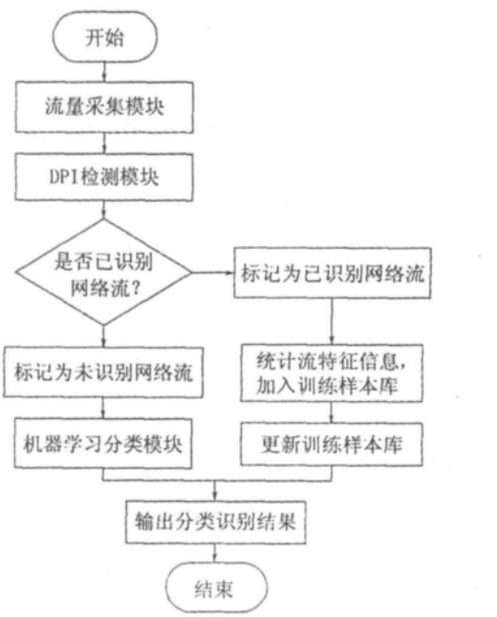

图 2 基于 DPI 和机器学习的网络流量分类流程

Fig. 2 The flow chart of traffic classification based on DPI and machine learning

2.3 算法的实现

基于 DPI 和流统计特征的网络流量分类方法主要由 DPI 检测模块和机器学习分类模块组成。

2.3.1 DPI 检测模块

DPI 检测模块主要根据特征库 RuleLib，通过模式匹配对数据流量进行深入分析，识别出具体的应用流量。其中，特征库以 XML 文件的形式存放各种协议的特征。检测流表根据流的五元组信息存储已检测出的数据流，当后续流的五元组信息和检测流表中的已有流信息相同时，则可以直接判定为相同的应用流量。DPI 分类引擎的工作流程如图 3 所示。

2.3.2 机器学习分类模块

朴素贝叶斯分类方法是通过计算后验概率来确定样本所属类别的，其基本思想是基于概率论中的贝叶斯公式和条件独立性假设，采用属性和类别的联合概率来估计新样本的类别。在机器学习分类模块中，采用朴素贝叶斯(Naïve Bayesian，简称 NB)分类方法对网络流量进行分类。朴素贝叶斯分类流程如图 4 所示。

1)准备工作阶段。特征选择是在保证应用数据原有价值的前提下去除与类别属性无关的冗余特征，获得对分类效果最有效的特征组合。特征属性选择

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

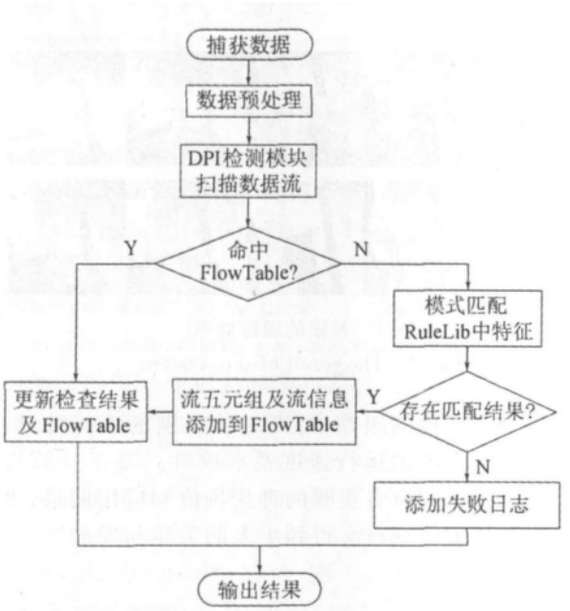

图 3　DPI 网络流量分类流程

Fig. 3　The flow chart of traffic classification based on DPI

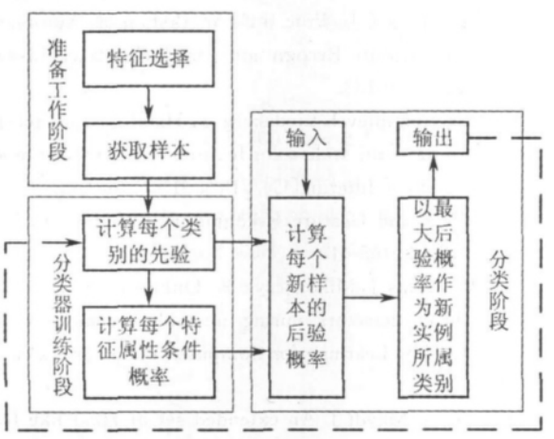

图 4　朴素贝叶斯网络流量分类流程

Fig. 4　The flow chart of traffic classification based on NB

采用 FCBF 算法从网络流的 248 个特征属性中选择 8 个特征属性作为特征集。

2)分类器训练阶段。朴素贝叶斯方法属于有监督分类方法,使用离线训练方式对训练样本进行预处理。实验中,采用 Matlab 对样本数据集进行了分析,计算了相应的网络流量应用类型的先验概率及条件概率。

3)分类阶段。分类阶段根据训练阶段得到的决策模型训练生成分类器,实现网络流量的在线分类。

3　实验结果与分析

流量分类算法在 CentOS 系统中实现,使用 Wireshark 在校园局域网络中捕获数据,然后对其进行处理,只保留 BitTorrent、PPStream 这类 P2P 流量以及属于 WWW 类型的 HTTP 流量,最后分别采用基于 DPI 的流量分类方法和基于本模型的流量分类方法对流量数据进行分析。

从表 1 可以看出,基于 DPI 的分类方法对 PPStream 产生的流量的识别率要明显低于对 BitTorrent 流量的识别。这是因为 BitTorrent P2P 文件共享软件是开源的,通过对其程序和应用协议的分析比较容易发现其协议特征,使用 DPI 技术能够很好的识别对应的网络流量。而对于 PPStream 私有商业化应用软件,只能通过分析网络数据包以及反编译的手段获取其协议特征,其准确率在一定程度上受到了限制,而导致对这些流量的分类识别率有所降低。对于 DPI 不能识别的网络流通过机器学习的方法进行识别,把 BitTorrent 和 PPStream 流量判定为 P2P 流量,弥补了 DPI 识别的不足。如表 2 所示,将 BitTorrent、PPStream 产生的流量判定为 P2P 流量,本研

表 1　基于 DPI 算法的流量识别结果

Tab. 1　The results of flow classification based on DPI

协议名	实际流量/Byte	识别流量/Byte	流量识别率	实际连接数	识别的连接数	连接识别率
BitTorrent	1 090 976 041	1 002 761 045	91.9%	251	239	95.2%
WWW	81 975	79 875	97.4%	10	10	100%
PPStream	38 409 170	27 654 605	72%	245	184	75.1%

表 2　基于本算法的流量识别结果

Tab. 2　The results of flow classification based on DPI and ML

协议名	实际流量/Byte	识别流量/Byte	流量识别率	实际连接数	识别的连接数	连接识别率
P2P	1 129 385 211	1 036 801 859	94.0%	496	472	95.2%
WWW	81 975	79 875	97.4%	10	10	100%

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

究所采用的分类方法由于结合了 DPI 技术和机器学习算法来检测网络流量，对 BitTorrent、PPStream 这类 P2P 流量识别有了显著提高，提高了对网络流量的整体识别率。

4 结束语

通过分析基于特征字段和基于流统计特征的机器学习 2 种网络流量分类方法，提出了一种基于 DPI 和机器学习的网络流量分类方法。该方法以 DPI 技术为主识别大多数网络流量，减少了需要通过机器学习方法识别的工作量，同时 DPI 技术能够识别具体的应用流量，提高了识别的精确度。以基于流统计特征的机器学习方法辅助识别加密和未知特征的网络流，弥补了 DPI 技术不能识别新应用及加密流量的缺点，提高了网络流量的识别率。

参考文献：

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

[2] 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，2010，30(3)：216-219.

[3] Moore A W，Papagiannakik. Toward the accurate identification of network applications[C]//PAM 2005 Proceedings of the 6thIntenational Workshop on Passive and Active Network Measurement. Berlin：Springer-Verlag，2005：41-54.

[4] Sen S，Spatscheck O，Wang D. Accurate，scalable internet-work idetification of P2P traffic using application signatures[C]//Proceedings of the 13th International Conference on World Wide Web. New York：[s. n.]，2004.

[5] Zhao Honglai，Galis A，Rio M，et al. Towards automatic traffic classification[C]//Third International Conference on Networking and Services. Athens：[s. n.]，2007：19-29.

[6] Zander S，Nguyen T，Armitage G. Automated traffic classification and application identification using machine learning[C]//Proceedings of the IEEE Conference on Local Computer Network 30th Anniversary. Sydney：IEEE Press，2005：250-257.

[7] Moore A W，Zuev D. Internet traffic classification using Bayesian analysis techniques[C]//Preceedings of the ACM SIGMETRICS Internet Conference on Measurement and Modeling of Computer Systems. New York：ACM Press，2005：50-60.

[8] 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(20)：2692-2704.

[9] Karagiannis T，Broido A，Faloutsos M，et al. Transport layer identification of P2P traffic[C]//Proc of the 4th ACM SIGCOMM Conference on Internet Measurement. New York：ACM Press，2004：121-134.

[10] Kumar S，Turner J，Williams J. Advanced algorithms for fast and scalable deep packet inspection[C]//Proc of IEEE/ACMANC'06. New York：ACM Press，2006：81-92.

编辑：曹寿平

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net