



基于行为特征学习的互联网流量分类方法

刘珍¹, 王若愚²

(1. 广东药科大学医药信息工程学院, 广东 广州 510006;

2. 华南理工大学信息网络工程研究中心, 广东 广州 510006)

摘要: 基于连接图的互联网流量分类方法能反映主机间的通信行为, 具有较高的分类稳定性, 但是经验式总结的启发式规则有限, 难以获得高分类准确率。研究分析了主机间通信行为模式和 BOF 方法, 从具有相同{目的 IP 地址, 目的端口号, 传输层协议}网络流量中, 提取主机间连接相关的行为统计特征(HCBF), 采用 C4.5 决策树算法学习基于行为特征的分类规则, 其无需人工建立启发式规则。在传统互联网和移动互联网流量数据集上, 从基本分类性能和分类稳定性方面, 与现有的特征集进行比较分析, 实验结果表明, HCBF 特征集合的类间区分能力和稳定性较高。

关键词: 互联网流量分类; 行为特征; 机器学习; 通信行为; 网络测量

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016152

Internet traffic classification method based on behavior feature learning

LIU Zhen¹, WANG Ruoyu²

1. School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China

2. Information and Network Engineering and Research Center, South China University of Technology, Guangzhou 510006, China

Abstract: The connection graph based internet traffic classification method can reflect the connectivity behavior between hosts. Thus, it has high stability. But the heuristic rules summarized for traffic classification are generally incomplete, and they difficultly obtain high classification accuracy. Host communication behavior model and BOF method was researched, and a set of host connection related behavior features (HCBF) was extracted from the multiple flows with the same {destination IP, destination port and transport protocol}. To evaluate the performance of HCBF, it was compared with the existing feature set on the respect of basic classification performance and classification stability. The experiments were carried out on the traffic collected in the traditional and mobile networks. Results show that HCBF out performs existing feature sets.

Key words: internet traffic classification, behavior feature, machine learning, communication behavior, network measurement

收稿日期: 2016-04-11; 修回日期: 2016-05-09

通信作者: 王若愚, rywang@scut.edu.cn

基金项目: 国家自然科学基金资助项目(No.61501128)

Foundation Item: The National Natural Science Foundation of China(No.61501128)



1 引言

互联网流量分类是异常检测、网络计费、流量整形、网络规划、QoS 部署和网络协议研究等的重要基础^[1,2]。随着动态端口号、端口伪装和载荷加密技术的使用,传统的端口号映射和载荷特征检测方法逐渐失效。至今,学术界提出了多种互联网流量分类方法,特别是基于机器学习和通信行为的流量分类方法成为研究热点^[3,4]。

基于连接图(connectivity graph)的流量分类方法通过分析和描述网络应用的 IP 报文在主机间的传输行为模式,进而识别未知 IP 报文的应用类别。参考文献[5]基于主机传输层的行为特点提出 BLINC(blind classification)方法,从社会级别、功能级别和应用级别分析各类主机的行为,并建立通信行为模式,进而建立启发式规则用于流量分类。近期也发展出基于连接图的流量分类方法,此类方法利用主机^[6]、节点^[7]或网络流^[8]建立连接图,基于连接图的性质(例如顶点的出度数、入度数等)提取通信行为测度,进而基于这些测度建立分类规则。此类方法能描述网络应用在主机间的连接状况,体现通信行为,不易受网络环境的影响,较为稳定。但是难以提取完整的启发式规则进行流量分类,其分类精度和分类粒度有限。

基于机器学习的网络流量分类方法引入了机器学习算法,对网络 IP 报文根据五元组进行组流,在网络流的基础上进行统计特征提取(例如流持续时间、平均报文大小、报文数等),统计特征值描述网络流建立样本集合,作为分类算法的输入。在基于机器学习的网络流量分类方法领域,多种统计特征已被提出^[9,10],最有代表的是 Moore 等人提出的 248 个统计特征^[9],包括了报文大小、报文到达时间间隔等的统计特征,其得到了广泛的应用^[11-13]。此类方法与基于通信行为的流量分类方法相比,其分类粒度和分类精度方面更优,但是,这些特征不能反映主机间的通信行为,随着网络应用逃避检测策略的发展,它们的类间区分能力可能被模糊化技术弱化,例如,参考文献[14]总结出模糊化报文大小的一种方式随机清除发送缓冲区,这样可能削弱某种网络应用的报文大小的规律。

为提高基于机器学习流量分类方法的分类稳定性,参考文献[7]分析观察某类流量数据的通信特点,在多流数据集上根据节点信息提取行为统计特征,但是此特征集合包含的信息有限。已有的通信行为模式^[9]能包含多种网络应用类别的通信行为,信息量更丰富,但以连接图的形式存

在,缺乏研究文献在此基础上提取出行为统计特征。

针对上述问题,本文的主要贡献包括以下两方面。

(1)为结合基于连接图和基于机器学习的网络流量分类方法的优点,本文分析基于连接图的流量分类方法,根据其中的通信行为模式和启发式规则,并结合 BOF(bag of flow)方法的思想,在具有相同{目的 IP 地址,目的端口号,传输层协议}的网络流上提取行为统计特征集合,用于描述网络流,建立的特征向量作为机器学习算法的输入,用于分类器训练。

(2)在传统有线网的流量数据集和移动智能终端手机的流量数据集上,利用 C4.5 决策树分类算法,从基本分类性能、分类时间稳定性等多个方面,实验分析行为统计特征的性能,并总结得出:仅基于行为统计特征训练的分类器的分类性能欠佳,综合描述连接行为和通信过程的行为统计特征(host communication behavior feature, HCBF)能进一步提高流量分类性能;在移动互联网流量数据集上,基于节点间连接的行为特征不利于分类基于 HTTP 的多种移动服务,例如 Web 浏览、视频流等。

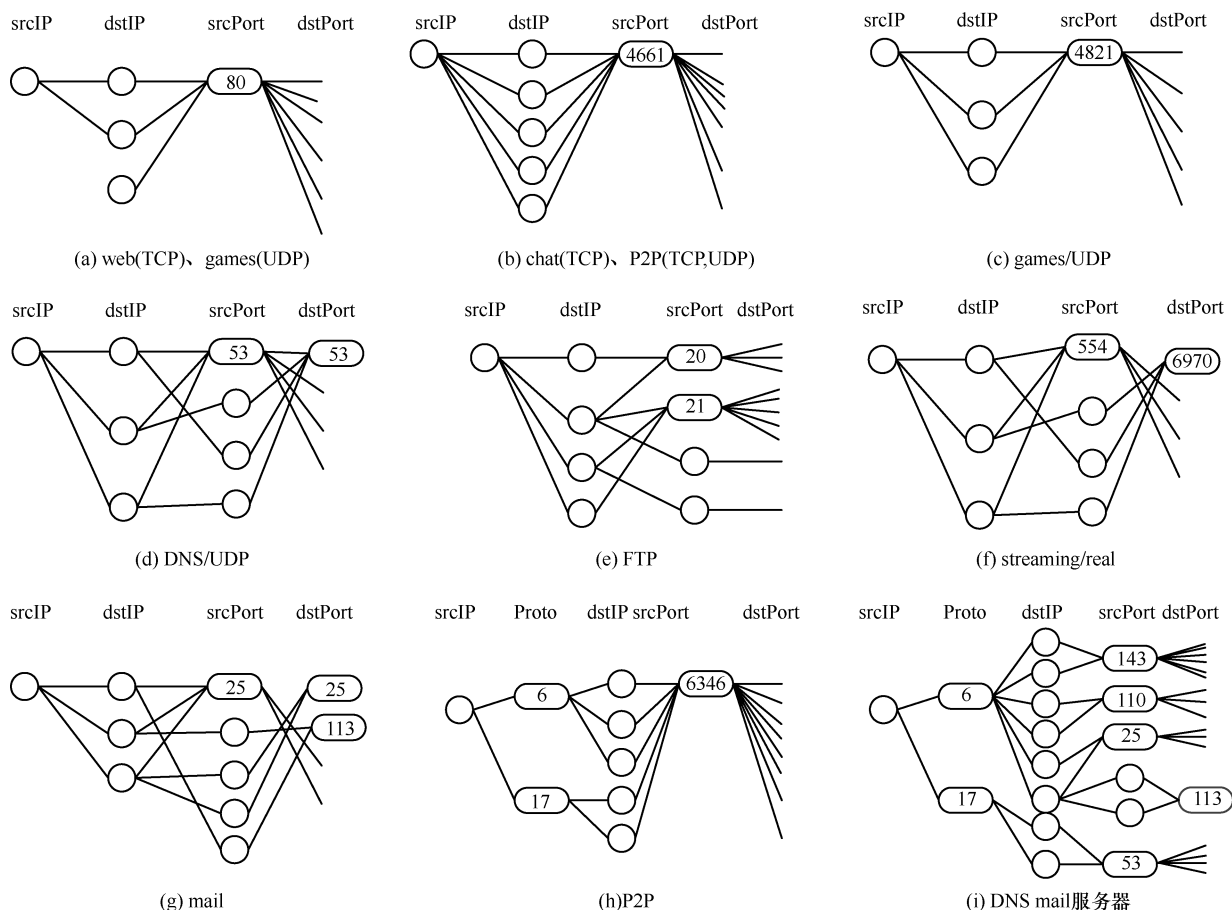
2 行为统计特征

本节主要介绍行为统计特征提取过程,并简要介绍现有的基于节点通信的行为特征集合,此特征集合将在实验部分与本文的特征集合进行比较分析。

2.1 通信行为模式

参考文献[5]通过观察和分析在传输层的主机间通信行为,提取其中的通信行为模式,并进一步提出启发式规则,实现网络流到网络应用之间的映射。此方法从 3 个级别分析主机的通信行为:社会级别、功能级别和应用级别^[5]。社会级别主要反映某个主机与其他主机通信的热门度(popularity)和主机间连接的集群性(community);利用目的 IP 地址的数量或 IP 地址范围进行度量。功能级别主要反映主机在通信中扮演的角色(服务器、客户端或者两者皆有);利用源 IP 地址和源端口进行度量,例如客户端使用多个端口与多台主机连接,而服务器端则通常使用一两个端口与其他主机通信。应用级别主要反映主机间在通信过程中表现出来的传输层连接模式;利用源 IP 地址、源端口、目的 IP 地址、目的端口进行度量。主流应用的通信行为模式如图 1 所示。

基于连接模式,多种启发式规则被建立用于分类未知网络流,相关的特征信息如下^[9]。

图 1 主机间通信行为模式示意^[9]

(1) 传输层协议

基于 TCP 的网络应用包括 P2P、Web、chat、FTP 和 mail 等, 基于 UDP 的网络应用包括网络管理流量和游戏等, 基于 TCP 和 UDP 的网络应用包括 P2P、streaming 等。

(2) 集合的基数

目的集合(目的端口和目的 IP 地址)的相对基数能够区分不同的应用行为, 例如区分 Web 与 P2P 和 chat, 或网络管理与游戏。此方法主要比较目的 IP 地址数和目的端口数, 例如 Web 服务器端的目的端口数多于目的 IP 地址数, 而 P2P 的目的端口数大约等于目的 IP 地址数。

(3) 使用每流的平均报文大小

许多应用表现出不同的传输报文大小的模式。例如: 游戏、恶意软件和反垃圾邮件在通信过程中的报文大小几乎保持恒定。

(4) 社团 (community)

同一个社团的 IP 主机通常表现出一样的行为。

(5) 递归探测方式

主机提供的某些服务可以通过递归探测其交互得到, 例如反垃圾邮件服务器的识别可以通过递归式探测其与电子邮件服务器的交换, 因为通常情况下, 反垃圾邮件服务器只与电子邮件服务器通信。

(6) 基于无载荷流识别

无载荷的流量或者失败的流量可能来自于攻击行为或者 P2P 网络(客户端试图建立 P2P 连接, 连接可能失败)。

但是, 这些启发式规则难以覆盖完所有的通信情况, 而且适用范围有限, 互联网流量在不断变化, 而且目前移动互联网流行, 这些规则不一定适用于现有的网络流量数据。

2.2 HCBF

图 1 的通信模式表明服务端通常拥有大量的客户端与之通信。如图 2 所示, 腾讯和阿里巴巴表现为热点服务, 在某高校的移动终端网络的 5 min 内有大量的主机与这些服务器通信。

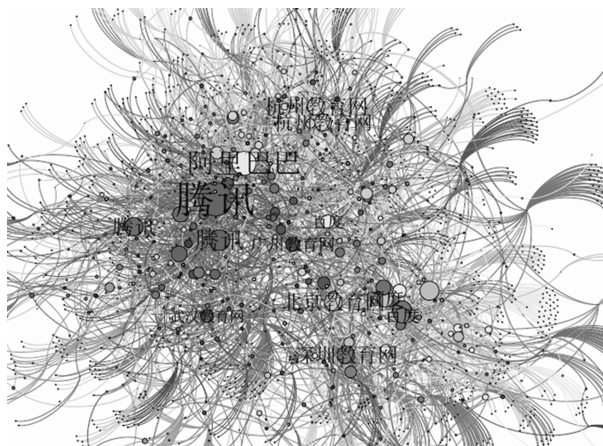


图2 校园网与外网的通信连接

腾讯服务与客户端通信的二部图如图 3 所示, 主要使用 80、443 和 8000 端口, 此行为模式与图 1(a) 的 Web 服务类似。但是, 客户端的端口数与 IP 地址数并没有出现明显的差距。因此, 仅依赖连接模式进行流量分类, 难以取得高分类性能。

为了建立适合于传统互联网和移动互联网环境的流量分类方法,本文提出行为统计特征。此类特征属于多流统计特征,即在多条流上提取统计特征,进而描述网络流,建立特征向量,将其作为分类算法的输入,利用机器学习算法学习分类规则,而非人工总结分类规则。

参考文献[17]提出 BOF 的思想,指出在一定时间间

隔内,相同目的 IP 地址(dstIP)、目的端口号(dstPort)和传输层协议(Proto)的网络流属于同一种应用。本文的多流特征基于相同的{dstIP, dstPort, Proto}的网络流量提出,包括通信行为特征和通信过程特征。在提取特征之前,先分析主流应用的通信情况,即与同一个{dstIP, dstPort, Proto}连接的来源主机信息。以 HTTP 和 BT 为例,分别代表 Web 和 P2P 类的应用。在某数据集上,HTTP 和 BT 应用的源 IP 地址数目和源端口数分别如图 4 和图 5 所示。横轴是源 IP 地址的序号,纵轴是源 IP 地址个数或源端口数。从图 4 和图 5 表明,HTTP 应用的端口数明显少于源 IP 地址数,表明客户端使用多个端口号与 Web 服务建立连接。BT 应用流量中,部分主机的源端口比源 IP 地

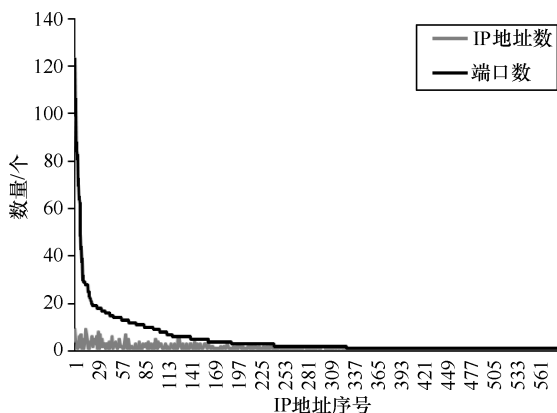


图 4 HTTP 应用的源 IP 地址数和源端口数

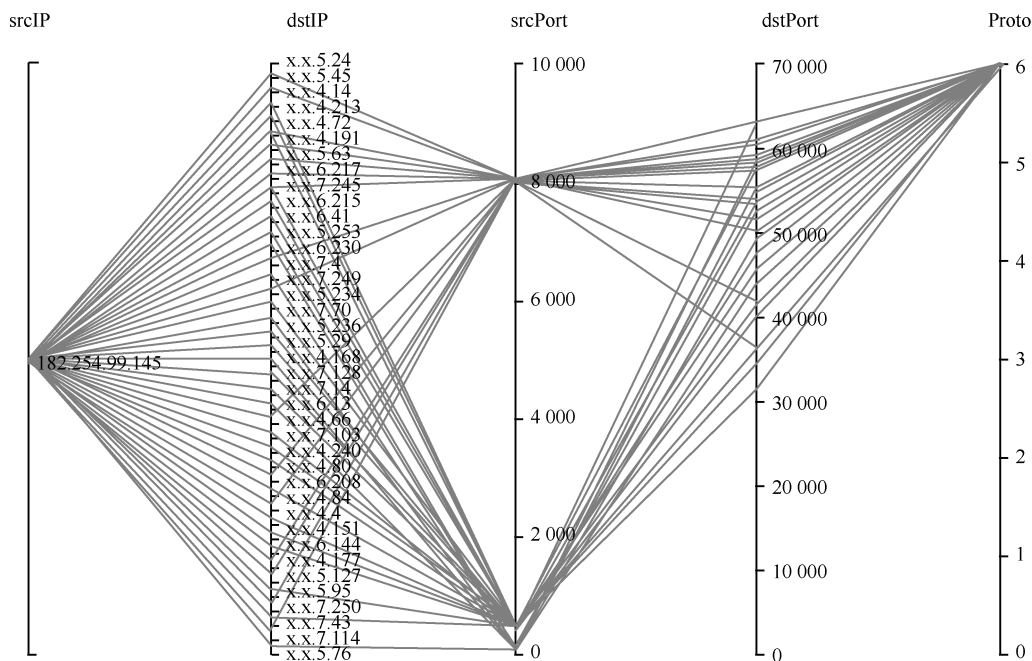


图3 腾讯服务的二部图

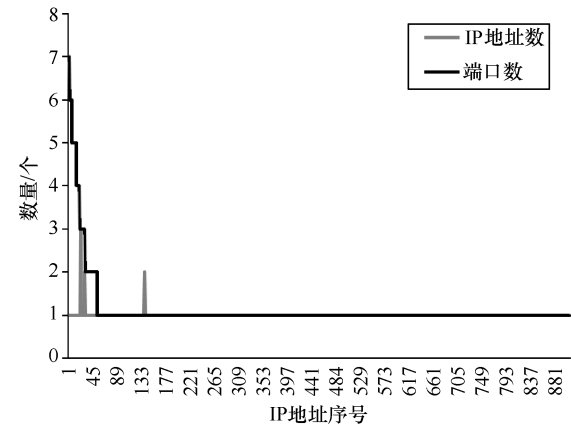


图 5 BT 应用的源 IP 地址数和源端口数

址多,而大部分主机的源端口和 IP 地址数一样,并且等于 1,即只与某个主机进行通信。

针对与相同{dstIP, dstPort, Proto}通信流量的平均报文大小,HTTP 和 BT 的情况分别如图 6 和图 7 所示。明显看出,BT 应用的报文大小的波动范围小于 HTTP 应用。

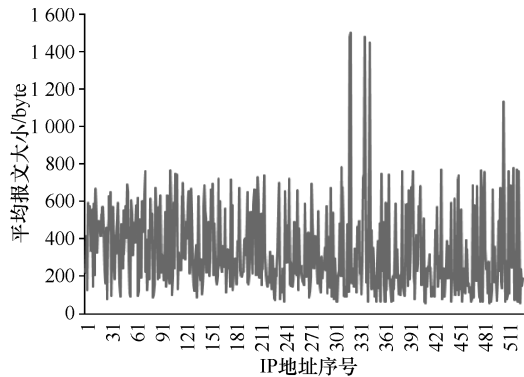


图 6 HTTP 应用的平均报文大小

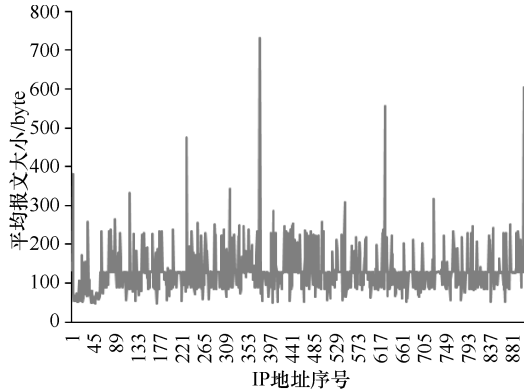


图 7 BT 应用的平均报文大小

基于上述分析,本文根据连接特征和传输过程的特征,建立通信行为统计特征集合,统称为 HCBF 集合,这些特征是在具有相同目的 IP 地址、目的端口号和传输层协

议的网络流量上计算得到,具体见表 1。

表 1 HCBF 特征集合

名称	含义
RaSrcIPPort	源端口数与源 IP 地址数的比值
destPort	目的端口号
AvgPkts	每流平均报文数
MinPkts	最小报文数
MaxPkts	最大报文数
StdPkts	报文数的均方差
AvgByts	每流平均字节数
MinByts	最小字节数
MaxByts	最大字节数
StdByts	字节数的均方差
Proto	传输层协议
AvgDuration	每流持续时间
FailFlowN	失败流数目

HCBF 特征解释如下。

(1)源端口数与源 IP 地址数的比值

从图 1 描述的各种网络应用的通信模式可看出,Web 服务器利用 80 端口与多个目的 IP 地址的多个端口进行通信,chat 服务则利用 4661 端口与多个目的 IP 地址进行通信。总体上,服务器利用少量的端口号与多个客户端通信,这意味着客户端利用多个端口号与服务端通信。图 4 和图 5 也表明在真实数据集上有这样的情况,因此,利用某主机的源端口数与 IP 地址数的比值,可以区分传统服务或 P2P 应用。

(2)目的端口号

图 1 表明,某些服务仍然使用固定的端口号进行通信,例如 DNS 采用 53,mail 使用 25 和 113 端口,端口号仍然具有一定的识别能力。

(3)报文数和字节数相关特征

多种应用表现出不同的传输报文大小的模式。例如:游戏、恶意软件和反垃圾邮件在通信过程中的报文大小(字节数)几乎保持恒定。图 6 和图 7 也表明 BT 的平均报文大小的波动范围小于 HTTP 应用。

(4)传输层协议

图 1 (a) 也表明 Web 和 games 的区别是 games 使用 UDP,另外 mail、FTP 等服务采用 TCP,P2P 可能使用 TCP 和 UDP,DNS 使用 UDP 等,传输层协议可区分这些应用。

(5)平均流持续时间

流持续时间反映一次通信连接的持续时间,P2P 网络存在节点的进入和退出的动态变化,大部分节点之间通信



的持续时间较小, chat 应用进行会话, 持续时间比较长, 网络流的平均持续时间特征的值较大。

(6) 失败流数目

无载荷的流量或者失败的流量可能来自 P2P 网络(例如客户端试图建立 P2P 连接, 连接可能失败)。

2.2 C4.5 决策树

在行为统计特征上, 本文采用 C4.5 决策树分类算法学习行为统计特征的分类规则, 用于流量分类。另外, 参考文献[18]表明, C4.5 决策树在互联网流量数据集上具有分类精度和分类效率方面的优点。C4.5 决策树^[19]是通过迭代式自顶向下选择测试属性作为树节点, 测试属性选择是基于信息增益率。叶子节点被标记为类别, 测试节点包括一个或多个输出, 每个输出对应一棵子树。假设一个有 m 个类别的数据集 S , 由特征向量 $A=\{A_1, A_2, \dots, A_n\}$ 描述, 假设特征 $A_i(i=1, 2, \dots, n)$ 有 v 个离散的取值。由 A_i 划分的信息增益率为:

$$GainRatio(S, A_i) = \frac{Gain(S, A_i)}{SplitInformation(S, A_i)} \quad (1)$$

其中:

$$Gain(S, A_i) = Entropy(S) - \sum_{i=1}^v \frac{|s_i|}{|S|} Entropy(s_i) \quad (2)$$

$$Entropy(S) = - \sum_{j=1}^m \frac{|s_j'|}{|S|} \lg \frac{|s_j'|}{|S|} \quad (3)$$

$$Entropy(s_i) = - \sum_{j=1}^m p_{ji} \lg p_{ji} \left(p_{ji} = \frac{|s_{ji}|}{|s_i|} \right) \quad (4)$$

$$SplitInformation(S, A_i) = - \sum_{i=1}^v \frac{|s_i|}{|S|} \lg \frac{|s_i|}{|S|} \quad (5)$$

s_j' 表示类别 C_j 的流量集合, s_{ji} 表示 C_j 的流量中 A_i 取第 i 个值的流样本, $s_{ji}(j=1, \dots, m)$ 的集合组成 s_i 。 s_i 表示特征

A_i 取第 i 个值的流量集合。当 C4.5 决策树分类未知流样本时, 从根节点开始, 从上往下比较测试属性的取值, 直到到达叶子节点。

3 实验与结果分析

3.1 实验数据

本文的实验数据包括传统互联网和移动互联网流量数据。有线网流量数据采集于某高校办公楼的出口路由器, 采集时间为 2011 年 9 月 17 日和 2011 年 9 月 25 日, 每次的采集持续时间为 30 min。这两天的流量数据包含了常用网络应用的流量, 例如 HTTP、BT、HTTPvideo、QQ 等, 其代表了用户日常使用网络的流量。对流量数据进行如下处理: 基于 L7-filter 和端口号对 IP 报文进行类别标记^[15]; 按照五元组对 IP 报文进行组流; 基于报文基本字段计算行为特征值, 建立特征向量; 每 10 min 的流量数据组成一个流样本集。两天的数据分别命名为 Day917 和 Day925, 它们的流和字节数见表 2 和表 3, 可见两者的类别分布有较大的区别。在 Day917 中, eDonkey 的网络流最多, 而在 Day925 中 HTTP 和 DNS 的网络流较多, 这与采集时间段内的用户上网行为有关。

移动网流量数据采集于志愿者的智能手机终端, 采集时间为 2016 年 2 月 29 日, 利用 GT 的方式^[16]进行类别标记, 并对智能终端 App 进行归类, 具体的类别分布情况见表 4。

3.2 评估指标

在互联网流量数据中, 分类性能可以通过流或字节进行评估。本文采用的性能评估指标包括: 单类流/字节分类准确率、总体流/字节分类准确率、流/字节 g-mean。这些评估指标都基于 4 个基本的测量指标, 即 TP(true positive)、FP(false positive)、TN(true negative) 和 FN(false negative)。对

表 2 Day917 数据集

类别	Day917_1		Day917_2		Day917_3	
	# 流/条	# 字节/MB	# 流/条	# 字节/MB	# 流/条	# 字节/MB
HTTP	6 421	64.080	9 072	199.051	8 678	133.504
DNS	6 714	1.253	11 722	2.070	9 912	1.779
eDonkey	14 474	31.439	12 945	83.960	18 064	43.705
SSL	130	2.026	88	0.943	140	1.604
BT	2 198	15.422	3 072	3.076	3 258	19.943
HTTPvideo	24	13.948	22	48.096	18	32.256
Xunlei	1 792	3.725	1 348	2.565	2 400	31.162
QQ	770	0.822	479	1.646	646	1.833
Kugou	220	58.150	400	2.437	247	4.894

表 3 Day925 数据集

类别	Day925_1		Day925_2		Day925_3	
	# 流/条	# 字节/MB	# 流/条	# 字节/MB	# 流/条	# 字节/MB
HTTP	10 786	87.707	8 687	123.760	9 109	82.783
DNS	11 480	2.645	7 371	1.190	7 097	1.201
eDonkey	589	16.937	360	16.912	329	7.526
SSL	133	9.474	348	108.942	146	13.335
BT	4 239	62.195	3 778	43.271	2 692	68.687
HTTPvideo	654	77.414	332	76.714	280	60.574
Xunlei	703	3.560	463	2.025	406	12.012
QQ	966	1.061	540	1.124	527	0.811
Kugou	1 557	1.707	139	5.657	129	8.749

表 4 移动网流量数据集

类别	# 流/条	# 字节/KB
social	230	261.877
Web	33	0.190
VoIP	25	0.037
mail	32	0.460
shop	204	1.094
video	31	91.036
navigation	1 112	3.442
service	321	1.037
others	272	3.838
unknown	676	6.325

某个类别 C_0 , TP 表示 C_0 的流量数据中被正确分类的流样本 (或字节) 数, FP 表示被错误分类为 C_0 类别的流样本 (或字节) 数, TN 表示非 C_0 的流量数据中被正确分类的流样本 (或字节) 数, FN 表示 C_0 的流量数据中被错误分类的流样本 (或字节) 数。

单类流/字节分类准确率表示每个类别的网络流中被正确分类的流/字节比率, 如式 (6) 所示。 R_i 表示类别 $C_i (i = 1, \dots, m)$ 的分类准确率。

$$R_i = \frac{TP_i}{TP_i + FN_i}$$

(6)

总体流/字节分类准确率表示总体的网络流中被正确分类的流/字节比率, 如式 (7) 所示。

$$OA = \frac{\sum_{i=1}^m TP_i}{\sum_{i=1}^m TP_i + FN_i}$$

(7)

互联网流量数据存在类不平衡问题, 即分类器可能偏向于分类大类 (拥有大量的流样本) 的流样本, 而忽略

小类 (拥有少量的流样本) 的流样本的分类性能。 $g-mean$ 是不平衡分类问题中常用的评估指标, 表示每类的分类准确率的几何平均^[11], 如式 (8) 所示。当所有类别的分类准确率为 100%, $g-mean$ 为 1; 当某个类别的分类准确率为 0%, $g-mean$ 为 0。好的分类模型在 $g-mean$ 上应当接近于 1。

$$g-mean = \left(\prod_{i=1}^q R_i \right)^{1/m}$$

(8)

3.3 实验结果分析

3.3.1 基本分类性能

本文采用 C4.5 决策树作为分类算法, 分别在 Day917 和 Day925 数据集上, 以前 10 min 数据作为训练集, 后两个 10 min 数据作为测试集, 例如 Day917_1 作为训练集, Day917_2 和 Day917_3 作为测试集。在以下的实验结果中, 分类性能最好的以粗体标出。

Abacus 和 HCBF 特征集合上的实验结果见表 5。在大部分数据集上, 基于 HCBF 特征集合的分类器的总体流分类准确率和总体字节分类准确率更高, 相比于基于 Abacus 的分类器, 分别提高了 5.16% 和 61.8%。Abacus 主要表征节点之间通信的过程, 即报文大小的分布情况, HCBF 从连接特征和通信过程两个方面描述网络流量, 例如源端口和源 IP 地址的比值能区分传统的服务端和客户端以及 P2P 应用。此外, 端口号也作为特征, 能区分使用固定端口号的传统应用, 例如 mail、DNS、FTP 等。

大类 (例如 HTTP、DNS 等) 包含大量的网络流样本, 对总体分类准确率的贡献大, 高的总体流分类准确率表明大类的流样本能较好地被分类。重型流 (具有高字节的网络流) 样本对总体字节分类准确率的贡献大, 高字节分类准确率反映对这些流的分类性能好。这些网络流会消耗更多



表 5 基本分类性能比较

数据集	特征集合	流 <i>g-mean</i>	字节 <i>g-mean</i>	总体流分类准确率	总体字节分类准确率
Day917_1 对比Day917_2	Abacus	0	0	87.8%	21%
	HCBF	41.2%	31.8%	90.4%	87%
Day917_1 对比Day917_3	Abacus	0	0	86.8%	48.7%
	HCBF	59.9%	34.7%	92%	70.5%
Day925_1 对比Day925_2	Abacus	23.3%	4.3%	80.4%	32.4%
	HCBF	46.8%	20.2%	88.3%	49.2%
Day925_1 对比Day925_3	Abacus	34.4%	16.9%	85%	57.8%
	HCBF	33.9%	26%	86.9%	52%

的网络带宽,因此正确识别这些网络流,有利于实施网络容量规划等活动。

网络流量存在类不平衡问题,除了大类和重型流,某些小类的分类性能对网络管理也非常重要,例如即时通信应用 QQ,正确识别并有效传输此应用的报文,提高即时传输性能,从而改善用户体验。但是总体的流或字节分类准确率不能反映小类的分类性能。*g-mean* 通常用于评估不平衡数据集上的分类器的分类性能。表 4 表明基于 HCBF 的分类器总表现出更高的流 *g-mean*、字节 *g-mean*,这表明分类器在类间的分类性能更均衡。Abacus 在 Day917 数据集上获得 0 的流 *g-mean* 和字节 *g-mean*。进一步分析单类的分类准确率(见表 6 和表 7),发现 HTTPvideo 应用获得 0 的流分类准确率和字节分类准确率。结合表 1 的网络流分布发现,此应用的网络和字节数较少,这可能由于类之间的网络和字节分布不平衡导致。Abacus 反映与某节点通信的报文大小和报文数目的分布情况,需要大量的网络流量,此特征集合在 HTTPvideo 应用和其他小类上的区分能力较弱。

表 6 Day917 的单类分类准确率

类别	流分类准确率		字节分类准确率	
	Abacus	HCBF	Abacus	HCBF
HTTP	86.6%	86.3%	33.2%	96.3%
DNS	97.5%	99.9%	96.3%	99.3%
eDonkey	95.3%	96.7%	3.3%	88.8%
SSL	4.5%	31.3%	1.3%	25.2%
BT	60.2%	76.4%	20.5%	83.3%
HTTPvideo	0	20.2%	0.0%	22.4%
Xunlei	44.5%	65.4%	15.4%	12.4%
QQ	58%	64.2%	18.0%	19.1%
Kugou	1.5%	18.5%	0.1%	38.6%

表 7 Day925 的单类分类准确率

类别	流分类准确率		字节分类准确率	
	Abacus	HCBF	Abacus	HCBF
HTTP	87.1%	92.2%	74.6%	76.8%
DNS	99.7%	99.9%	98.5%	98.7%
eDonkey	6.4%	13.0%	0.1%	2.5%
SSL	12.1%	73.4%	4.7%	55.1%
BT	60.2%	74.8%	3.2%	3.6%
HTTPvideo	68.1%	37.3%	29.7%	88.4%
Xunlei	3.0%	11.4%	0.7%	4.0%
QQ	37.6%	79.6%	19.3%	71.1%
Kugou	6.5%	21.7%	0.2%	33.2%

基于 HCBF 在多个小类上获得更高的流和字节分类准确率,例如 HTTPvideo、Kugou 和 SSL 等。在 Day925 数据集上,基于 HCBF 的分类器可为 SSL 获得 73.4% 的流分类准确率和 55.1% 的字节分类准确率。单类分类准确率再次证明 HCBF 特征集合在小类流样本上能获得更高的分类性能。

3.3.2 讨论

(1) 时间稳定性

Day917 和 Day925 数据集之间相差 8 天。为评估时间稳定性,此部分将 Day917 数据集作为训练集,Day925 数据集作为测试集,分类结果见表 8。结合表 4 和表 5,实验结果表明 Day917 上训练的分类器,在 Day925 上的分类性能比在 Day917 上的差。这说明网络流量的统计特征的取值分布发生了变化。结合表 1 表明,网络流在类间的分布也发生了变化。分类器的分类性能会随着时间的推移而弱化,即流量数据发生了概念漂移。表 8 表明 HCBF 仍然是性能最好的特征集合,因为此集合不仅反映主机间的连接行为,还提取了网络流量的通信过程,例如平均报文大小、流持续时间、报文大小的均方差等,具有更好的类间区分能力。

表 8 Day917 与 Day925 的分类结果

特征集	流 <i>g-mean</i>	字节 <i>g-mean</i>	总体流分类 准确率	总体字节分类 准确率
Abacus	17.2%	13.0%	83.0%	42.2%
HCBF	35.8%	13.2%	86.5%	52.5%

(2)移动互联网流量数据上的分类性能

随着移动智能终端的快速发展,有效分类移动互联网流量有利于实施移动互联网的网络管理活动,目前也成为互联网流量分类的热点研究领域。此部分研究 3 种特征集合在移动网流量数据集上的类间区分能力。以 C4.5 决策树作为分类算法,10 倍交叉验证的结果见表 9。实验结果表明依然是基于 HCBF 训练的分类器取得更高的流*g-mean*、字节*g-mean*和字节分类准确率。基于 Abacus 训练的分类器获得 0%的流/字节*g-mean*,进一步分析发现 Web, VoIP 和 video 都获得 0 的分类结果,这是由于智能手机 App 大多基于 HTTP,与传统的 Web 流量很类似,若仅基于报文大小分布的主机间连接特征,较难区分这些应用。HCBF 特征,不仅从 4 个网络流集合反映连接数、连接的 IP 地址数和端口数的比例、失败流数目,还有持续时间、报文数、字节数、传输层协议等信息,能从多方面表征每种网络应用的特性,从而具有更高的类间区分能力,在移动互联网流量数据上,取得 92.9%的流分类准确率和 85.3%的字节分类准确率。

表 9 移动网流量数据上的分类结果

特征集	流 <i>g-mean</i>	字节 <i>g-mean</i>	总体流分类 准确率	总体字节分类 准确率
Abacus	0	0	90.8%	71.1%
HCBF	46.2%	45.5%	92.9%	85.3%

(3)其他分类算法上的分类性能

上述实验都是基于 C4.5 决策树,为验证 HCBF 在其他分类算法上的分类性能,接下来的实验利用随机森林(random forest)和 1NN 在 Day917 与 Day925 上的实验结果见表 10。

表 10 random forest 和 1NN 分类结果

算法	特征集	流 <i>g-mean</i>	字节 <i>g-mean</i>	总体流分 类准确率	总体字节 分类准确率
随机 森林	Abacus	10.4%	13.5%	84.9%	56.5%
	HCBF	32.3%	19.2%	91.4%	59.2%
1NN	Abacus	17.1%	17.6%	84.2%	53.1%
	HCBF	24.1%	13.8%	80.4%	51.7%

实验结果表明,利用 1NN 算法,HCBF 的性能比 Abacus 差。但是,利用随机森林,HCBF 的分类性能仍然最优。本文的主要思想是利用 C4.5 决策树算法学习行为特征分类网络流量的分类规则,这说明基于分类规则的机器学习算法适合本文的行为特征。

(4)在其他数据集上的分类性能

在上部分只给出两天数据集上的实验结果,本部分给出在其他流量数据集上的分类结果,这些数据采集于 2012 年的 5 月 1 日和 6 月 3 日。前者数据作为训练集,后者作为测试集,C4.5 决策树的分类结果见表 11。实验结果表明 HCBF 的 *g-mean* 和总体分类准确率仍然较优。

表 11 其他实验数据集上分类结果

特征集	流 <i>g-mean</i>	字节 <i>g-mean</i>	总体流分类 准确率	总体字节分类 准确率
Abacus	0	0	87.1%	78.9%
HCBF	51.7%	9.1%	89.5%	79.1%

4 结束语

本文提取了基于多流的行为统计特征集合 HCBF,用于描述网络流,建立流样本作为 C4.5 决策树算法的输入,进而训练分类网络流量的分类规则。此特征集合建立在主机间通信行为模式之上,从{dstIP, dstPort, Proto}网络流中通过提取主机间连接特性和通信过程信息获得。此方法结合了基于机器学习和基于通信行为的流量分类方法的优点,将其与已有的行为统计特征进行比较。实验结果表明,在分类精度和分类稳定性方面,HCBF 表现最佳。本文在传统互联网和移动互联网中提取了包含常用网络应用的流量,可代表用户的日常流量,实验的分类性能反映了统计特征在常用网络应用流量数据上的分类性能。

但是,基于 HCBF 的流量分类器的字节分类准确率较低,特别是在不同天数的流量数据集上,这可能是错分重型流造成,未来工作将研究重型流在主机间的传输型行为,并提取新的统计特征;HCBF 在全流的网络流量数据上获取到,难以用于在线流量分类,未来将研究在子流上提取行为统计特征,提高分类速度。

参考文献:

[1] WANG Y, XIANG Y, ZHANG J, et al. Internet traffic clustering with side information [J]. Journal of Computer and System Sciences, 2014, 80(5): 1021-1036.

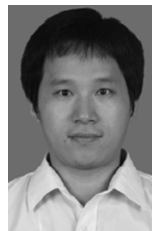


- [2] ZHANG J, CHEN X, XIANG Y, et al. Robust Network Traffic Classification [J]. IEEE/ACM Transactions on Networking, 2015, 23(4): 1257-1270.
- [3] DAINOTTI A, PESCAP A. Issues and future directions in traffic classification[J]. IEEE Network, 2012, 26(1): 35-40.
- [4] CALLADO A, KAMIENSKI C. A survey on Internet traffic identification [J]. IEEE Communications Surveys & Tutorials, 2009, 11(3): 37-52.
- [5] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark [J]. ACM SIGCOMM, 2005, 35(4): 229-240.
- [6] ILIOFOTOU M, KIM H, FALOUTSOS M, et al. Graption: a graph-based P2P traffic classification framework for the internet backbone [J]. Computer Networks, 2011, 55(8): 1909-1920.
- [7] BERMOLLEN P, MELLIA M, MEOB M, et al. Abacus: accurate behavioral classification of P2P-TV traffic[J]. Computer Networks, 2011(55): 1394-1411.
- [8] ASAI H, FUKUDA K, ESAKI H. Traffic causality graphs: profiling network applications through temporal and spatial causality of flows [C]//The 23rd International Teletraffic Congress, Sept 6-9, 2011, San Francisco, CA, USA. New Jersey: IEEE Press, 2011: 95-102.
- [9] MOORE A, ZUEV D, CROGAN M. Discriminators for use in flow-based classification [C]//In Passive & Measurement Workshop 2003 (PAM2005), August1, 2005, London, England. [S.l.: s.n.], 2005.
- [10] HAJJAR A, KHALIFE J, DAZ-VERDEJO J. Network traffic application identification based on message size analysis [J]. Journal of Network and Computer Applications, 2015 (58): 130-143.
- [11] LIU Z, WANG R Y, TAO M, et al. A class-oriented feature selection approach for multi-class imbalanced traffic datasets based on local and global metrics fusion [J]. Neurocomputing, 2015(168): 365-381.
- [12] FAHAD A, TARI Z, KHALIL I, et al. Toward an efficient and scalable feature selection approach for internet traffic classification [J]. Computer Networks, 2013, 57(9): 2040-2057.
- [13] HONG Y, HUANG C C, NANDY B, et al. Iterative-tuning support vector machine for network traffic classification[C]//The 2015 IFIP/IEEE International Symposium on Integrated Network Management, May 11-15, 2015, Ottawa, ON, Canada. New Jersey: IEEE Press, 2015: 458-466.
- [14] HJELMVIK E, JOHN W. Breaking and improving protocol obfuscation: No. 2010-05, ISSN 1652-926X [R]. [S.l.: s.n.], 2010: 1-34.
- [15] LEE S, KIM H, BARMAN D, et al. NeTraMark: a network traffic classification benchmark [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(1): 23-30.
- [16] GRINGOLI F, SALGARELLI L, DUSI M, et al. GT: picking up the truth from the ground for internet traffic [J]. ACM SIGCOMM Computer Communication Review, 2009, 39 (5): 13-18.
- [17] ZHANG J, XIANG Y, WANG Y, et al. Network traffic classification using correlation information[J]. IEEE Transactions on Parallel & Distributed Systems, 2013, 24(1):104-117.
- [18] WILLIAMS N, ZANDER S, ARMITAGE G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. SIGCOMM Computer Communication Review, 2006, 30(5): 5-16.
- [19] 徐鹏, 林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009, 20(10):2692-2704.
- XU P, LIN S. Internet traffic classification using C4.5 decision tree[J]. Journal of Software, 2009, 20(10): 2692-2704.

[作者简介]



刘珍(1986-),女,博士,广东药科大学讲师,主要研究方向为互联网流量分类、机器学习和移动互联网。



王若愚(1977-),男,博士,华南理工大学工程师,主要研究方向为计算机网络和模式分类。