doi: 10.3969/j.issn.0490-6756.2019.01.008

一种基于 DPI 自关联数据包检测分类方法

贾 军[1], 杨 进[2], 李 涛[2]
(1. 四川大学计算机学院, 成都 610065; 2. 四川大学网络空间安全学院, 成都 610065)

摘 要: 随着互联网的不断发展,越来越多的非传统业务兴起,由于大量采用迂回机制、加密隐藏技术,使得这些业务变得难以控制管理,影响传统业务的正常性能. 现有识别方法普遍采用端口识别以及深度包检测技术 DPI,难以识别迂回流量以及加密流量. 因此本文提出一种基于 DPI 自关联检测分类方法,该方法首先通过与样本流之间七元组关联关系识别迂回流量,这部分称为强关联(SA),然后提取检测流特征值,通过本文提出的分类决策函数进行识别,这部分称为弱关联(WA),实验结果表明,该方法能克服 DPI 技术不能识别迂回流量以及加密流量的缺点,提高业务流识别准确率.

关键词: 自关联; DPI; 关联流量; 业务识别
中图分类号: TP393 **文献标识码**: A **文章编号**: 0490-6756(2019)01-0029-08

A DPI-based autocorrelationmethod for packet detection classification

JIA Jun[1], *YANG Jin*[2], *LI Tao*[2]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. College of Cyberspace Security, Sichuan University, Chengdu 610065, China)

Abstract: With the continuous development of the Internet, more and more non-traditional services are emerging and occupying a large amount of network bandwidth resources, which makes Internet services and security more and more difficult to be managed andaffects the normal performance of traditional services. The existing identification methods generally use port identification and DPI (Deep Packet Inspection) technology, which is difficult to identify roundabout traffic and encrypted traffic. This paper proposes a classification method based on DPI autocorrelation detection. This method firstly identifies the roundabout flow through the seven-tuple association relationship with the sample stream, called as strong autocorrelation (SA). Then, the detected stream features are extracted and identified by the classification decision function proposed in this paper. This part is called weak autocorrelation(WA). The experimental results show that the proposed method can overcome the DPI shortcomings in the roundabout and encrypted traffic identification and improve the traffic flow identification accuracy.

Keywords: Autocorrelation; Deep packet detection(DPI); Associated flow; Traffic identification

收稿日期: 2018-05-08
基金项目: 国家重点研发计划(2016yfb0800604, 2016yfb0800605); 国家自然科学基金(61572334,U1736212); 四川省重点研发项目(2018GZ0183)
作者简介: 贾军(1994—), 男, 四川绵阳人, 硕士生, 研究方向为网络安全. E-mail: 631642753@qq.com
通讯作者: 杨进. E-mail: 253960818@qq.com

1 引 言

随着互联网业务的不断渗透发展,当前网络环境充斥着各式各样的新的业务流量,常见的有 P2P(Peer-to-Peer)、VoIP(Voice over IP)、流媒体、音视频聊天等等. 由于这些流量大量采用了动态端口、自定义加密、流量伪装[1]以及启用迂回流等技术,使得当前的网络流量的管理难度日益增加,这也是导致国家网络安全问题日益复杂的原因之一. 因此对网络流量的分类具有重要的意义,一方面可以保证服务提供商能根据网络分类做服务质量 QOS(Quality of Service)保证[2],另一方面也能有效地发现、管控网络安全问题.

目前针对网络流量的分类主要的方法有基于端口的分类、基于有效载荷的分类以及基于机器学习的分类,以上方法都各有其优势和不足. 其中基于端口的分类实现简单、检测高效,但越来越多的业务启用动态端口技术,致使该方法目前的识别准确率只有 30%~70%左右[3,4];基于有效载荷的分类针对明文业务流具有很高的识别准确率,但该方法涉及到用户隐私数据且无法对迂回流及加密流进行识别[5];基于机器学习的检测方法能实现不依赖协议端口、数据包载荷的情况下利用网络流特征识别网络应用,因此成为当下研究热点[6],Velan 等人对现有的加密流量分类方法进行了总结[7],张泽鑫等人在文献[8]中研究了朴素贝叶斯分类方法的有效性.

基于以上分析,本文提出一种基于 DPI 自关联数据包检测分类方法 SACM(Self-Associated Classification Method,)来对网络数据包协议进行检测分类,该方法分为两个部分,分别命名为强关联 SA(Strong Association)和弱关联 WA(Weak Association),如图 1 所示. 对基于 DPI 检测方法,分类失败的网络数据包与已检测到数据流协议进行关联检测. 强关联利用相同应用层协议业务会在一定时间范围内采用不同的传输层协议或端口进行数据传输的特点来对业务进行关联,弱关联则利用已有标记数据集建立多特征分类决策函数,根据此决策函数将未知业务与已知业务应用层协议类别进行关联.

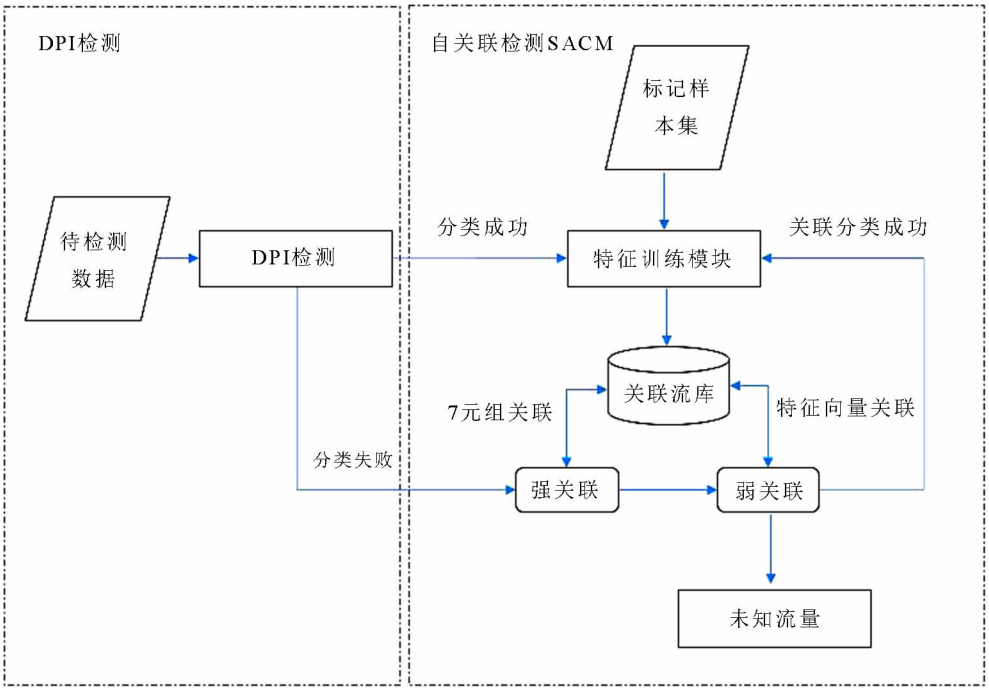

图 1 自关联检测方法流程图

Fig. 1 Flow chart of autocorrelation detection method

实验表明,本文提出的方法在相同数据集上与现有 DPI 检测方法相比,能克服 DPI 检测方法不能识别迂回流及加密流的缺点,提高了业务流分类准确率,同时减小误报率.

2 自关联检测方法

2.1 强关联

SA 部分主要分为两个步骤. 首先,它将所有

能通过 DPI 检测出来的业务流提取七元组(时间窗口、源 IP 地址、目的 IP 地址、源端口号、目的端口号、传输层协议、服务类型(TOS))信息，加上识别业务应用层协议信息一起保存在关联流表中，其中时间窗口是一变量值，表示关联信息失效时间，该值将在实验中具体给出.

然后当有新的未识别的业务流经过强关联部分时，提取未知流的七元组信息与关联流表进行关联检测. 如果关联成功则认为识别成功，反之进入弱关联检测. 强关联流程算法见算法 1. 由于一条网络会话流是由流五元组(源 IP、目的 IP、源端口、目的端口、传输协议)唯一决定的，本节使用流相关描述属性进行关联，具有较强的关联性，因此称之为强关联检测.

由以上分析可知，SA 部分用于检测前后流量存在相关性的网络业务流. 常见的业务有被动 FTP、TFTP、BT 下载流量等，表 1 是 BT 下载流量启用迂回流量传输过程. 可以看出 BT 传输正常流量带有明显“BitTorrent protocol”的指纹特征，当正常流量受阻后，随即在本机开启新的端口采用相同的传输协议、目的 IP、目的端口来传输数据，这部分数据是无特征的. 例如当序号 1、2、3 数据受阻后序号 5 流量开启并正常传输，由于序号 5 流量与 1、2、3 流量存在七元组强关联，因此可以通过本 SA 部分关联检测出来.

算法 1　强关联算法

输入： 待检测网络数据流 X，关联流表 info_table

输出： 网络数据流关联分类结果

```
Begin
1)  提取网络数据流七元组信息 mete_info
2)  for i:=1 to info_table.len() do:
3)  if info_table[i].time<=current_time:
4)  表项关联信息失效，调用关联流表删除函数
5)  continue;
6)  if mete_info == info_table[i]. mete and
    mete_info. time<info_table[i]. time:
7)  调用关联流表添加函数，将新识别的流量加入关联流表
8)  更新当前表项 info_table[i]的有效时间窗口大小
9)  returninfo_table[i].protocol
10) else:
11) continue;
12) 调用弱关联入口函数，将待检测数据包传递给弱关联
    End
```

表 1　BT 下载迂回流量传输过程

Tab.1　The circuitous flow transmission process of BT

序号	到达时间	传输协议	源 IP	目的 IP	源端口	目的端口	特征
1	33:30	TCP	192.168.2.125	125.43.184.184	3528	16881	13 BitTorrent
2	34:00	TCP	192.168.2.125	125.43.184.184	3544	16881	13 BitTorrent
3	34:50	TCP	192.168.2.125	125.43.184.184	3565	16881	13 BitTorrent
4	35:30	TCP	192.168.2.125	37.187.127.62	3565	51413	13 BitTorrent
5	33:50	TCP	192.168.2.125	125.43.184.184	3554	16881	无特征
6	35:50	TCP	192.168.2.125	37.187.127.62	3572	51413	无特征

2.2　弱关联

针对加密网络业务流量，目前普遍的做法是使用机器学习的方法进行自动分类，本文采用朴素贝叶斯算法对网络业务流进行分类，为避免朴素贝叶斯假设各特征之间相互独立，重要性相同所带来的影响，本文使用层次分析法对业务流特征集赋予权重，并以此提高分类器的分类准确率，WA 分类决策流程如图 2 所示.

WA 部分也主要分为两个步骤. 首先提取标记训练数据集的特征向量，在该标记数据集上建立基于改进的朴素贝叶斯分类模型.

然后，当有未识别的业务流经过 WA 检测部分时，提取其相关特征向量，计算不同分类情况下的后验概率，取最大值为该业务流的分类结果，当该最大后验概率满足某一阈值则表示接受此分类结果，将该业务流加入到训练样本中.

由上可知，弱关联检测是使用流间特征作为关联因素，由于这些流特征并不属于流描述五元组，

因此其关联性较弱,因此称之为弱关联.

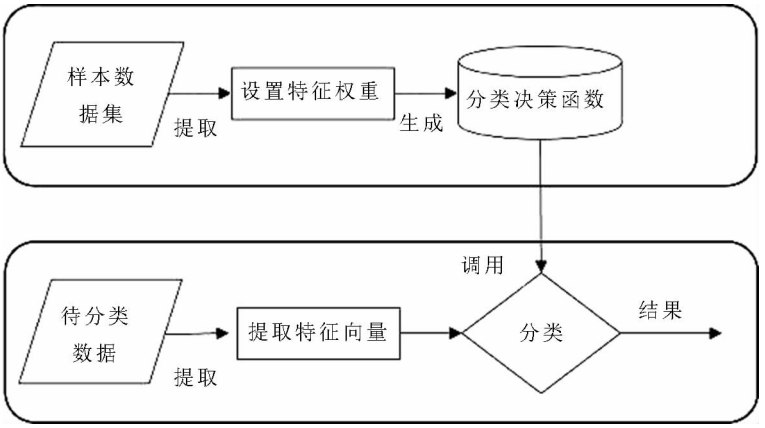

图 2 WA 分类决策流程

Fig. 2 Classification decision process of WA

3 分类决策函数建立

剑桥大学教授 Moore 等人通过傅里叶变换得到 249 项网络业务流特征[9],孙振在文献[10]中研究表明其中大部分特征之间存在冗余、对流量分类相关性不高,因此采用基于 NetFlow 技术统计特征作为本文的特征集,见表 2.

表 2 NetFlow 特征集

Tab. 2 NetFlow feature set

属性	描述
BYTES	业务流字节总数
PACKETS	业务流报文总数
DPORT	目的端口
DURATION	业务流持续时间
IAT	业务流平均报文间隔时间
PACKETSIZE	业务流平均报文大小
PPS	业务流每秒平均传输包数
BPS	业务流每秒平均传输字节数

3.1 改进朴素贝叶斯分类算法

朴素贝叶斯分类算法分类效率高、实现简单,因此被广泛应用于网络业务分类领域,文献[11,12]均采用了朴素贝叶斯的方法对网络业务流进行分类,但其假设条件每个流量特征对分类影响因子相同显然是不符合实际条件的,因此本节给出基于朴素贝叶斯分类器改进的形式化描述.

现假设有 M 条网络业业务流 $X=\{X_1,X_2,X_3,\cdots,X_m\}$,有 N 个分类类别选项 $C=\{C_1,C_2,C_3,\cdots,C_n\}$,已知每个网络业务流可用 8 个流特征集描述,分别为 $F=\{F_1,F_2,F_3,\cdots,F_8\}$. 对于任意网络业务流 X_i 其属于类别 C_j 的概率为

$$P(C_j|X_i)=\frac{P(F_1,F_2,F_3,\cdots,F_8|C_j)P(C_j)}{P(F_1,F_2,F_3,\cdots,F_8)} \quad (1)$$

式(1)是贝叶斯公式,对于在类别 C_j 下概率 $P(F_1,F_2,F_3,\cdots,F_8|C_j)$的计算,朴素贝叶斯给出了假设性等式.

$$P(F_1,F_2,F_3,\cdots,F_8|C_j)=\prod_{k=1}^{8}P(F_k|C_j) \quad (2)$$

式(2)等式的成立是在满足特征集 F 互相不影响且对分类结果影响相同的假设下,现假设特征集 F 的每个特征对分类结果的影响因子不同,且有权重向量 $W=(W_1,W_2,W_3,\cdots,W_8)^T$ 与之对应,则式(2)变为

$$P(F_1,F_2,F_3,\cdots,F_8|C_j)=\prod_{k=1}^{8}W_KP(F_k|C_j) \quad (3)$$

根据式(1)和式(3)可得改进后朴素贝叶斯分类器的后验概率计算公式为

$$P(C_j|X_i)=\frac{P(C_j)\prod_{k=1}^{8}W_KP(F_k|C_j)}{\prod_{k=1}^{8}W_KP(F_k)} \quad (4)$$

其中,先验概率 $P(C_j)$为类别 C_j 在整个网络业务流中所占比例,朴素贝叶斯分类器的目标是找出一个类别 C_j 使得式(4)取得最大值,由此可得分类决策函数.

$$f(X_i)=\max_{j=1\to n}(P(C_j|X_i))$$
$$=\max_{j=1\to n}\left(\frac{P(C_j)\prod_{k=1}^{8}W_KP(F_k|C_j)}{\prod_{k=1}^{8}W_KP(F_k)}\right) \quad (5)$$

3.2 特征集权重计算

上一节给出了分类决策函数(5),其中权重向

量 W 是未知的，在以往的研究中，往往是凭借经验依据不同的特征的重要性赋予初始值，这种方法主观性太强，可信度较差，而权重向量的设置对于整个分类模块意义重大，而层次分析法 AHP(Analytic Hierarchy Process)[13]适用于一些较为复杂、模糊的难于完全定量分析的问题，因此本文采用 AHP 来计算权重向量的值.

层次分析法主要分为 3 个步骤：首先，建立递阶层级结构；然后，按照因素重要性度量表 3 构造各层次中所有判断矩阵；最后，利用线性代数的方法计算出权重大小并进行各层次一致性检验. 以下是 AHP 中的一些重要定义.

表 3　Saaty 重要性度量表

Tab. 3　Saaty importance meter

标度值	描述
1	两个因素相比，两者相同重要
3	两个因素相比，前者稍比后者重要
5	两个因素相比，前者明显比后者重要
7	两个因素相比，前者强烈比后者重要
9	两个因素相比，前者极端比后者重要
2,4,6,8	上述两个相邻重要等级之间
倒数	因数与因数之间的判断值为互反数 $a_{ji}=\frac{1}{a_{ij}}$

定义 1　若判断矩阵 $A=(a_{ij})_{n*n}$ 满足：

$$(i)a_{ij}>0,(ii)a_{ji}=\frac{1}{a_{ij}}(i,j=1,2,\cdots,n)$$

则称矩阵 A 为正反矩阵.

定义 2　若矩阵 $A=(a_{ij})_{n*n}$ 为正反矩阵，且 A 满足：

$$a_{ij}a_{jk}=a_{ik},\forall i,j,k=1,2,\cdots,n$$

则称矩阵 A 为一致矩阵.

实际问题中，我们引构造的判断矩阵往往只能满足定义 1 而不能满足定义 2，为了对矩阵的一致性进行度量，我们引入一致性比率公式如下.

$$CR=\frac{CI}{RI} \tag{6}$$

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{7}$$

式(6)中，CI 为一致性指标公式，当判断矩阵完全一致时，$CI=0$，式(6)中 RI 为平均一致性指标，参考值见表 4. 规定当一致性比率 $CR<0.1$ 时，我们认为判断矩阵是满足一致性要求的，反之则说明该判断矩阵不满足要求，需重新设计修改判断矩阵并重新计算.

表 4　随机一致性指标 RI

Tab. 4　Random consistency index RI

N	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

因此，我们假设 w_1、w_2、w_3、w_4、w_5、w_6、w_7、w_8 分别表示表 2 中 8 种特征，并利用这八种特征构造判断矩阵为

$$A=\begin{bmatrix} a_{11}\cdots & a_{18} \\ \vdots\cdots & \vdots \\ a_{81}\cdots & a_{88} \end{bmatrix}$$

然后结合表 3 的 Saaty 重要性度量表，将此八种特征进行两两比较，可得出以下判断矩阵.

$$A=\begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{9} & \frac{1}{5} & \frac{1}{9} & \frac{1}{3} & \frac{1}{7} & \frac{1}{7} \\ 3 & 1 & \frac{1}{7} & \frac{1}{3} & \frac{1}{7} & 1 & \frac{1}{5} & \frac{1}{5} \\ 9 & 7 & 1 & 5 & 1 & 7 & 3 & 3 \\ 5 & 3 & \frac{1}{5} & 1 & \frac{1}{5} & 3 & \frac{1}{3} & \frac{1}{3} \\ 9 & 7 & 1 & 5 & 1 & 7 & 3 & 3 \\ 3 & 1 & \frac{1}{7} & \frac{1}{3} & \frac{1}{7} & 1 & \frac{1}{5} & \frac{1}{5} \\ 7 & 5 & \frac{1}{3} & 3 & \frac{1}{3} & 5 & 1 & 1 \\ 7 & 5 & \frac{1}{3} & 3 & \frac{1}{3} & 5 & 1 & 1 \end{bmatrix}$$

使用和法求解判断矩阵 A 的最大特征向量近似值，首先将矩阵 A 按列进行归一化得到矩阵 A'，见式(8).

$$A'_{ij}=\left(\frac{a_{ij}}{\sum_{i=1}^{n}a_{ij}}\right) \tag{8}$$

$$\widetilde{A}=\left(\sum_{j=1}^{n}\frac{a_{1j}}{\sum_{i=1}^{n}aij},\cdots\sum_{j=1}^{n}\frac{a_{nj}}{\sum_{i=1}^{n}aij}\right)^{T} \tag{9}$$

然后根据式(9)按行求和得向量 $\widetilde{A}=(0.037,0.099,0.474,0.198,0.474,0.099,0.32,0.32)^T$，最后对向量 $\widetilde{A}$ 的每一个分量做归一化处理后，就可以得到权重向量 $W=(W_1,W_2,W_3,\cdots,W_8)^T$ 的近似解为 $W=(0.018,0.049,0.235,0.098,0.235,0.049,0.158,0.158)^T$.

$$\lambda_{\max}=\sum_{k=1}^{n}\frac{(AW)_k}{nW_K} \tag{10}$$

根据式(10)可计算出判断矩阵 A 的最大特征根 $\lambda_{\max}=8.765$，根据表 3 可得 RI 为 1.41，由式(7)可得 $CI=0.109$，由式(6)可得 $CR=0.077<0.1$，因此判断矩阵 A 满足一致性，则针对表 1 的八维

特征值的权重值为 $W=(0.018,0.049,0.235,\ 0.098,0.235,0.049,0.158,0.158)^T$.

$$A'=\begin{bmatrix}0.023 & 0.005 & 0.001 & 0.002 & 0.001 & 0.003 & 0.001 & 0.001\\ 0.068 & 0.014 & 0.002 & 0.004 & 0.001 & 0.008 & 0.001 & 0.001\\ 0.205 & 0.095 & 0.013 & 0.053 & 0.010 & 0.055 & 0.022 & 0.020\\ 0.114 & 0.041 & 0.003 & 0.011 & 0.002 & 0.024 & 0.002 & 0.002\\ 0.205 & 0.095 & 0.013 & 0.053 & 0.010 & 0.055 & 0.022 & 0.021\\ 0.068 & 0.014 & 0.002 & 0.004 & 0.001 & 0.008 & 0.001 & 0.001\\ 0.159 & 0.068 & 0.004 & 0.032 & 0.003 & 0.039 & 0.007 & 0.007\\ 0.159 & 0.068 & 0.004 & 0.032 & 0.003 & 0.039 & 0.007 & 0.007\end{bmatrix}$$

4 实 验

为了验证自关联 DPI 数据包检测分类方法的有效性,本实验针对同一数据集设置了两组对比实验,第一组只采用 DPI 检测方法对标记数据集进行分类;第二组基于 DPI 检测方法,加入自关联方法进行分类.

4.1 实验数据

表 5 数据集

Tab. 5 Data set

应用	SCUNET_1 流数	SCUNET_2 流数
腾讯 QQ	2223	2134
微信	2045	2141
BT 下载	2251	2132
迅雷下载	2062	2093
优酷视频	2281	2156
网易云音乐	2031	2251
FTP	2173	2016
TFTP	2034	2067

实验数据集来自四川大学校园网网关,总共采集了两个数据集,分别为 SCUNET_1, SCUNET2. 经分析发现,表 5 中的 8 种应用为校园网关的主要成分,因此本文选取这 8 种应用作为分类目标种类. 为了避免每种应用的训练数据样本不对称对实验结果造成影响,构造数据集时使每种数据流数基本保持一致,每种应用大约有 2000 条数据流作为样本集,这些数据流包含明文数据、迂回流量、加密流量等. 实验组 1 使用数据集 SCUNET_1,实验组 2 使用数据集 SCUNET_2 作为训练样本数据,数据集 SCUNET_1 作为测试数据集.

4.2 评估指标

为了评价自关联方法的有效性,我们假设总共有 n 个分类,TC_i 表示第 i 分类中正确分类的样本数目,FC_i 表示第 i 分类中被错误分类为其他类别的样本数目,FN_i 表示分类器错误分类为类别 i 的样本数目,因此可以得到如下样本评价指标.

(1) 准确率 $=\sum_{i=1}^{n}\frac{TC_i}{(TC_i+FC_i)}$,指分类器正确分类的样本数占所有样本的比例.

(2) 误报率 $=\sum_{i=1}^{n}\frac{FN_i}{(TC_i+FN_i)}$,指分类器错误分到某一类的样本占所有已分类样本的比例.

(3) 类召回率 $=\frac{TC_i}{TC_i+FC_i}$,指分类器正确分类到某一类别的样本数占该类别原所有样本数的比例.

(4) 检全率 $=\sum_{i=1}^{n}\frac{(TC_i+FN_i)}{(TC_i+FC_i)}$,指分类器识别出的所有样本数占总样本数的比例.

4.3 实验结果分析

第一组实验采用 DPI 技术分别对两组测试数据集进行实验,实验结果如表 6 所示,两组实验的平均分类准确率分别为 83.5%、84.3%,检全率分别为 95.3%、95.9%,从表 6 结果可以看出,DPI 分类技术的分类结果不是很理想,整体准确率较低,因为 DPI 技术不能有效的检测迂回流及加密流量.

表 6 DPI 技术分类结果

Tab. 6 Classification results of DPI Technology

数据集	流总数	TC_Total	PC_Total	PN_Total	准确率	误报率	检全率
SCUNET_1	17100	14278	2822	2025	83.5%	12.4%	95.3%
SCUNET_2	16990	14322	2668	1985	84.3%	12.2%	95.9%

第二组实验在 DPI 检测基础上，加入自关联检测分类方法，需要指出的是，在进行第二组实验之前，为确定强关联检测时间窗口大小，先进行了 5 组对比实验，分别取时间窗口为 0.5 s、1 s、2 s、4 s、8 s，实验结果表明，时间窗口取值为 4 s 时分类误报率最低、取值为 8 s 时拥有最高的检全率，因此本文取强关联检测时间窗口为 4 s 进行实验.

分别选取数据集中的一种作为样本集，另一组作为测试数据集，每次实验重复进行 10～15 次，最后取每次实验的分类结果平均值作为实验结果，由表 7 可以看出，加入自关联检测分类方法后，准确率分别达到 95.1%、94.9%，相比提高准确率 13.8%、12.6%. 系统检全率提升效果明显，分别达到 99.4%、99.3%.

表 7　基于 DPI 自关联检测方法分类结果

Tab. 7　Classification results based on DPI autocorrelation detection method

测试数据集	样本数据集	流总数	TC_Total	FC_Total	FN_Total	准确率	误报率	检全率
SCUNET_1	SCUNET_2	17100	16262	838	735	95.1%	4.3%	99.4%
SCUNET_2	SCUNET_1	16990	16123	867	754	94.9%	4.5%	99.3%

图 3 给出了每种应用在两组对比实验中的召回率，可以看出，所有应用的召回率都有所提高，其中 BT 下载、迅雷下载、FTP、TFTP 的召回率效果显著，平均分别提高了 19.65%、18.9%、38.13%、18.1%，因为 FTP 存在被动传输模式，传统的检测模式无法识别其数据流，因此 FTP 应用的召回率平均提高 38.13%. 优酷视频以及网易云音乐的召回率提升效果不明显，分别为 6.6%、3.95%，这是因为此类流媒体应用传输一般采用明文 UDP、或者 HTTP 模式传输，因此 DPI 技术可以达到很好的检测效果.

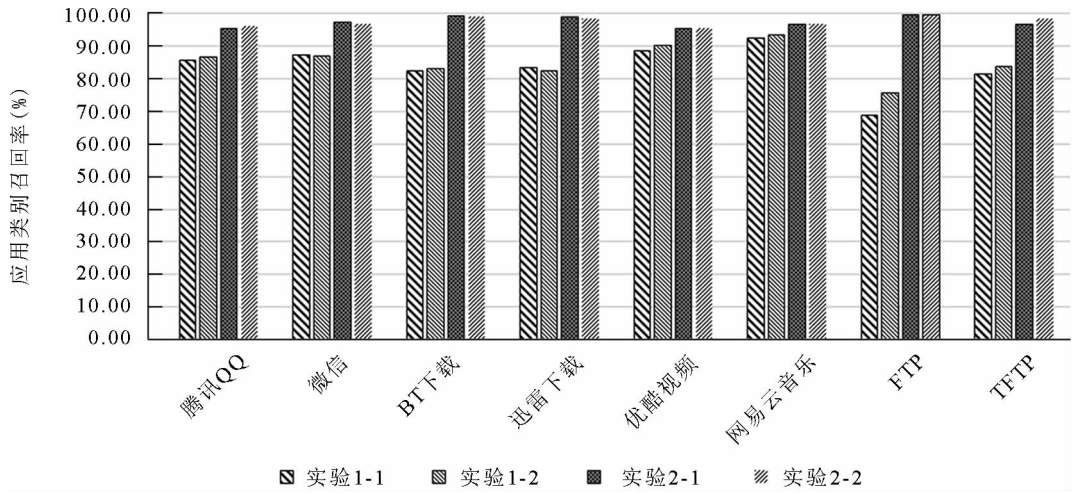

图 3　网络应用类别召回率

Fig. 3　Recall rate of network application category

实验结果显示，采用基于 DPI 的自关联的数据包检测分类方法可以在 DPI 检测技术方法上，有效检测网络迂回流及加密流，对整体数据包分类准确率有所提升，并提高整体检全率，每一类的应用准确率也相应有所提升.

5　结　论

实际网络环境下充斥着大量的迂回流及加密流量，目前常规 DPI 检测技术只能处理带有明显特征指纹的明文数据，并不能有效识别迂回流及加密流量，为此本文提出了一种基于 DPI 的自关联数据包检测分类方法. 该方法有如下优势：(1)强关联部分针对应用程序迂回流量，和数据传输流信息再握手阶段进行协商的流量分类效果明显. (2)弱关联部分能克服 DPI 技术不能识别加密流量的缺点，并且通过层次分析法计算各流特征的重要程度，以此提升朴素贝叶斯分类器的准确率.

实验结果表明，SACM 方法对迂回流量的检测效果明显，并且通过引入改进朴素贝叶斯分类算法，克服了 DPI 技术不能检测加密流量的缺点.

参考文献:

[1] Wright C V, Coull S E, Monrose F. Traffic morphing: an efficient defense against statistical traffic analysis [C]// proceedings of the Network & Distributed System Security Symposium. [s. l.]: DBLP, 2009.

[2] 耿道渠, 郭春, 李小龙, 等. 基于 ARM9 的嵌入式 Linux 系统移植研究与 QoS 功能实现 [J]. 四川大学学报: 自然科学版, 2014, 51: 719.

[3] Sen S, Spatscheck O, Wang D. Accurate, scalable in-network identification of p2p traffic using application signatures [C]//Proceedings of the International Conference on World Wide Web. NewYork: ACM, 2004.

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

[5] Dainotti A, Pescape A, Claffy K C. Issues and future directions in traffic classification [J]. Network IEEE, 2012, 26: 35.

[6] Nguyen T T T, Armitage G. A survey of techniques for internet traffic classification using machine learning [J]. IEEE Commun Surv Tut, 2009, 10: 56.

[7] Velan P, Čermák, Milan, Čeleda, Pavel, *et al*. A survey of methods for encrypted traffic classification and analysis [J]. Int J Netw Manag, 2015, 25: 355.

[8] 张泽鑫, 李俊, 常向青. 基于特征加权的朴素贝叶斯流量分类方法研究 [J]. 高技术通讯, 2016, 26: 119.

[9] Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques [J]. ACM SIGMETRICS Perform Eval Rev, 2005, 33: 50.

[10] 孙振. 基于机器学习的网络流量特征选择 [J]. 电子测量技术, 2017, 40: 131.

[11] Raveendran R, Menon R. An efficient method for internet traffic classification and identification using statistical features [J]. Int J Eng Tech Res, 2015, 4: 1.

[12] 邱密, 阳爱民, 刘永定, 等. 使用贝叶斯学习算法分类网络流量 [J]. 计算机工程与应用, 2010, 46: 78.

[13] 阮永芬, 高春钦, 李志伟, 等. 基于改进 AHP 与熵权法的膨胀土胀缩等级云模型评价 [J]. 江苏大学学报: 自然科学版, 2017, 38: 218.

引用本文格式:

中 文: 贾军, 杨进, 李涛. 一种基于 DPI 自关联数据包检测分类方法 [J]. 四川大学学报: 自然科学版, 2019, 56: 29.

英 文: Jia J, Yang J, Li T. A DPI-based autocorrelationmethod for packet detection classification [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 29.