2008年4月 北京邮电大学学报 Apr.2008
第31卷第2期 Journal of Beijing University of Posts and Telecommunications Vol.31 No.2

文章编号:1007-5321(2008)02-0015-05

基于流统计特性的网络流量分类算法

林 平, 余循宜, 刘 芳, 雷振明
(北京邮电大学 信息处理与智能技术重点实验室, 北京 100876)

摘要:针对传统基于单个流统计特性的网络流量分类算法识别率低、分类算法复杂的问题,在分析各类应用协议的基础上,发现了一组易于获取、可有效区分不同业务的网络流量特征.将这一组特征应用于网络流量分类,可以有效解决以往对等网络(P2P)业务识别率低下的问题;同时利用该组特征仅需采用多项逻辑斯谛回归算法即可实现网络流量的分类,较传统流量分类算法有较低的复杂度.实验结果表明,该组特征用于分类还具有较好的泛化特性,只需较少量训练样本即可在较长时间内保持较高的识别率.

关 键 词:网络流量分类;流;统计特征;多项逻辑斯谛回归
中图分类号:TP393.06 **文献标识码**:A

A Network Traffic Classification Algorithm Based on Flow Statistical Characteristics

LIN Ping, YU Xun-yi, LIU Fang, LEI Zhen-ming
(Key Laboratory of Information Processing and Intelligent Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Based on analysis of application protocols, a group of multi-flow characteristics with low complexity, high quality is proposed to mitigate the problem of low recognition rate and high implementation complexity associated with the traditional flow classification algorithms using single flow statistics. These characteristics can effectively identify peer-to-peer (P2P) traffic in network flow classification, and improve the recognition rate of the traditional algorithms. They also enable the use of multinomial logistic regression algorithm to classify the network flow, and reduce the complexity of the traditional algorithms. Experiment results show that the proposed characteristics can achieve good generalization, and only need a small number of training samples to get a model that can maintain good performance for a long time.

Key words: network traffic classification; flow; statistical characteristics; multinomial logistic regression

目前,通过网络监测手段对网路流量进行分类的最常用技术之一是利用标准端口,即通过分析报头,获得传输层端口信息,并把标准端口与特定的应用联系起来[1-2].但随着相当部分应用软件频繁采用非标准端口,甚至冒用其他应用软件的标准端口进行通信,传统基于端口的流量分类方法准确度越来越低.文献[3-5]提出采用匹配报文载荷中关键字的方法进行网络流量的分类,这种方法虽然准确率较高,但涉及用户隐私,且对载荷进行处理开销较大;随着加密协议和私有协议在网络上的迅速传播,这种方法应用的范围会逐渐缩小.

鉴于利用标准端口和关键字匹配的网络流量分

收稿日期:2007-06-19
作者简介:林 平(1982—),女,博士生,E-mail:linping.apple@gmail.com.

类算法的局限性,目前研究的热点主要集中在基于流的统计特征的流量分类算法.根据所利用特征的不同,目前常用的流分类算法可分为2类.第1类方法仅利用了待分类的单个流所包含的所有报文集合的统计特征,较有代表性的有文献[6-7];文献[8]总结了这类特征的特点——各种业务对应的特征参量是线性不可分的,因此只能采用复杂度较高的纯真贝叶斯和贝叶斯神经网络分类算法进行分类.第2类方法利用与待分类流有关的一组流集合的统计特征进行分类,充分利用了各种不同业务的社会特征,如文献[9]所提方法,其从社会、功能、应用3个不同的层次对一个流的主机行为作了分析,并按流的主机行为区分网络流量.本文利用文献[9]的思想,提出了一组易于提取、具有线性可分性的特征,可采用复杂度相对较低的多项逻辑斯谛回归分类算法进行网络流量分类,具有较高的识别率.

1 数据处理及特征提取

整个实验研究利用在某骨干网上采集到的真实数据,通过匹配报文关键字的方法对该数据从网络业务层面进行分类,以此作为实验的参考样本.实验中,通过程序对采集的数据进行特征提取,并用多项逻辑斯谛回归分类算法对获得的特征参量进行训练和检验.

1.1 采集数据的类别

本实验采用的报文包含了P2P(为peer与Tracker服务器的连接报文)、Email、RTSP(控制报文)和VoIP(为信令报文,没有语音通信的RTP报文),每类业务所包含的具体应用见表1.

表1 用于网络流量分类的网络应用类别

业务	具体应用	备注
P2P	BT, eDonKey, Soulseek, giFT	控制报文
Email	SMTP, POP 2/3, IMAP	控制、数据报文
RTSP	WMPlayer, RealMedia Player	控制报文
VoIP	H.323, SIP, MGCP	控制报文

1.2 流统计特征提取

定义1 流是指在超时约束下的2个主机对之间应用进程的双向通信的报文集合,即包含相同的主机IP地址对,使用相同的协议(如TCP、UDP、ICMP等),采用相同的进程端口对.

定义2 上行流指由流的发起方,即流的源地址,向宿地址发送消息的通信过程;下行流指由流的接收方,即流的宿地址,向源地址发送消息的通信过程.

定义3 存在流是指已经开始还未结束的流.源源集合是以待分类流的源地址作为源地址的所有存在流组成的集合;源宿集合指以待分类流的源地址作为宿地址的所有存在流组成的集合;宿源集合指以待分类流的宿地址作为源地址的所有存在流组成的集合;宿宿集合指以待分类流的宿地址作为宿地址的所有存在流组成的集合.

网络流量分类包括特征提取和分类器2个部分.本文所使用的流的统计特征包括单个流的特征和多个相关流的特征2大类.这些特征参数提取后将作为分类器的输入,具体特征表项见表2.

1) 单个流的特征,即对某个流分类时,仅利用组成该流的所有报文集合的统计特征,包括流中所有包的到达间隔的方差、报文大小方差,报文数等.该类特征利用了单个流所包含的信息,如不同的业务对应的流有不同的载荷长度、上下行流量等.

2) 多个相关流的特征.依据源源、源宿、宿源、宿宿集合的定义,可以提取与当前待分类流在时间/空间上有高关联的4组流集合特征.具体特征包括待分类流开始时刻各集合存在流的个数、不同端口数、不同IP数和不同4层协议类别数等.

以源源集合参数为例,在图1中表示出了4种业务(Email、RTSP控制报文、VoIP信令报文、P2P控制报文)对应的源源流集合特征关系.图中以流ID区分不同的流.一个流以连接线表示,其节点表明了该流的流ID、源IP、源端口、宿IP和宿端口,这些元素可以唯一标识一个流;虚线标出的为待识别流.所有连接线构成的集合即是以待分类流的源地址作为源地址的所有存在流组成的集合,对该集合,可以提取集合中存在流的个数、不同源端口数等特征参数作为分类依据.从图中可以看出,4种不同业务的各组源源参数差别显著,易于分类.

通过将2类特征有机结合,消除了仅采用第1类特征进行分类时每个流相互独立假设的局限性.引入第2类特征,可以充分利用与待分类流相关的一组流集合中隐含的信息,易于区分不同业务在宏观层面上的行为特征.实验结果表明,引入第2类特征还可以减少准确分类所需的第1类特征的数量,且所提取特征具有线性可分性,可采用复杂度相对较低的多项逻辑斯谛回归分类算法进行网络流量的分类.

表 2　用于区分网络业务的流的特征参数

流特征属性	流特征		每一类流特征的具体表现形式
单个流的特征	流 4 层协议		TCP、UDP 层协议类别
	流持续时间/s		会话时长
	流报文数		会话交互总报文数*，上行流报文数，下行流报文数
	报文流量/B		上行流的字节数，下行流的字节数
	报文速率/(B·s^{-1})		流速率，上行流速率*，下行流速率*
	载荷信息		一个会话中载荷的总字节数*(B)，载荷平均字节数(B)，载荷长度的方差(B^2)
	包到达间隔		一个会话中报文平均到达间隔(s)，到达间隔的方差(s^2)
多个相关流的特征	流开始时刻(共 20 个参数)	源源集合参数	集合中存在流个数、不同源端口数、不同目的 IP 数、不同目的端口数、采用的不同 4 层协议类别数
		源宿集合参数	集合中存在流个数、不同目的端口数、不同源 IP 数、不同源端口数、采用的不同 4 层协议类别数*
		宿源集合参数	集合中存在流个数*、不同源端口数、不同目的 IP 数、不同目的端口数*、采用的不同 4 层协议类别数*
		宿宿集合参数	集合中存在流个数、不同目的端口数、不同源 IP 数、不同源端口数、采用的不同 4 层协议类别数*

* 参数不在分类的最终模型中，为冗余参数或参数显著性水平不在 95% 置信区间内，为无用参数.

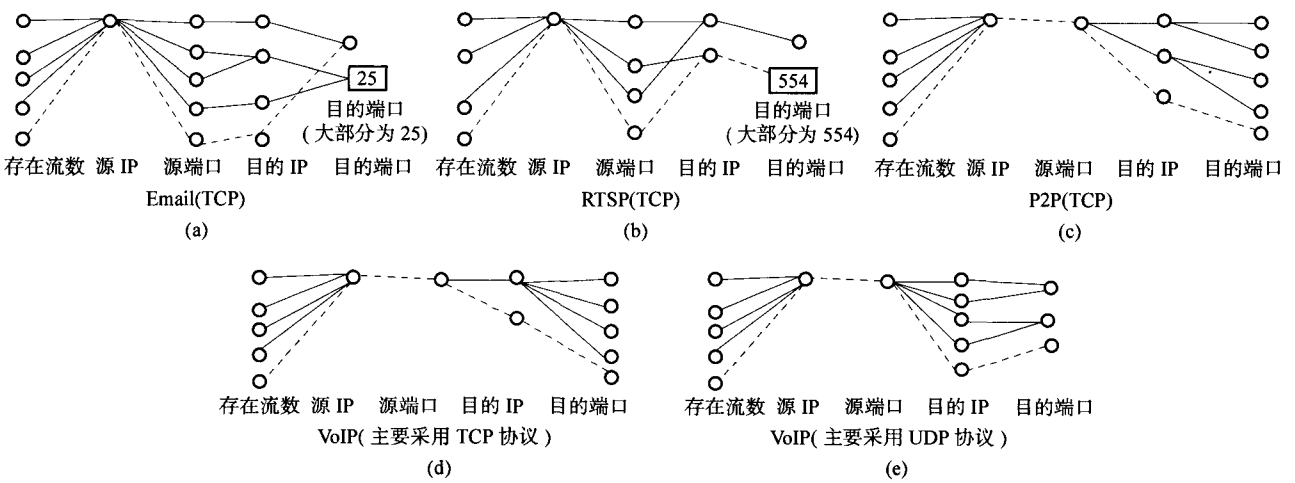

图 1　Email、RTSP、P2P、VoIP 流开始时源源关系图

2　多项逻辑斯谛回归算法

多项逻辑斯谛回归是多类模式识别的经典算法. 多项逻辑斯谛回归模型假设因变量的 logit 变换与各自变量有线性关系. 训练过程通过极大化一组已知分类的训练样本的后验概率确定模型参数. 多项逻辑斯谛回归较纯真贝叶斯和贝叶斯神经网络分类算法有模型简单、训练和分类速度快等优点.

假设共有 K 个分类，输入特征参数样本，即自变量为 N 维向量 $\boldsymbol{x}_i$、样本所属类别为 $C(\boldsymbol{x}_i)$，则模型因变量为 K 维向量 $\boldsymbol{\omega}$，且满足

$$\omega_k = \begin{cases} 1 & k = C(\boldsymbol{x}_i) \\ 0 & k \neq C(\boldsymbol{x}_i) \end{cases} \tag{1}$$

多项逻辑斯谛回归算法通过设定与自变量满足线性关系的中间变量 z_{ik}，计算因变量各分量为 1 的概率. 建立因变量和自变量的关系为

$$\pi_{ik} = \frac{e^{z_{ik}}}{e^{z_{i1}} + e^{z_{i2}} + \cdots + e^{z_{iK}}} \tag{2}$$

$$z_{ik} = b_{k0} + b_{k1}x_{i1} + \cdots + b_{kj}x_{ij} + \cdots b_{kN}x_{iN} \tag{3}$$

式中，π_{ik} 表示第 i 个观测样本属于第 k 类的概率，即 $P(\omega_{ik}=1|x)$；x_{ij} 表示第 i 个观测样本的第 j 个自变量；b_{kj} 为特征参数第 j 个自变量属于第 k 个类别的

回归系数.

训练过程利用一组已知分类的训练样本,采用最大似然估计法、对数似然法和非线性迭代法等方法估计逻辑斯谛回归模型的参数[10]. 训练后的模型可以预测某样本属于某分类的后验概率,即 π_{ik}. 依据最大后验概率准则即可对样本进行分类.

$$C(\boldsymbol{x}_i) = \arg\max_k \pi_{ik} \tag{4}$$

通过设定判决门限,分类算法还可以分离出无法准确分类的样本,以便及时发现、分析新增特殊业务类型,提高分类器性能.

3 实验结果

基于上述基本原理,对提供的网络流量原始数据,按表2进行特征提取的数据处理,并采用多项逻辑斯谛回归分析对网络流量的特征集进行业务层面上的分类. 随机选取2006-09-01的各类2 500个数据,即总共1万个数据作为训练样本;再以当天和2006-09-06的各类2 500个数据作为检验样本.

首先利用单个流的特征参数对Email、RTSP、P2P、VoIP这4种业务进行多项逻辑斯谛回归判别,训练样本的分类结果见表3. 从表中可以看出,在仅采用单个流的特征参数分类的情况下,训练样本中各类业务不完全线性可分,因此仅采用简单的多项逻辑斯谛回归分类算法无法获得满意性能.

表3 按单个流的特征训练样本分类结果

观察值	预测值				
	P2P	Email	RTSP	VoIP	百分比校正/%
P2P	1 620	880	0	0	64.8
Email	333	2 152	14	1	86.1
RTSP	701	19	1 766	14	70.6
VoIP	275	18	66	2 141	85.6
总百分比/%	29.3	30.7	18.5	21.6	76.8

然后利用本文提出的两类流的特征,即单个流特征和多个相关流特征,按这些特征对Email、RTSP、P2P、VoIP 4种业务进行多项逻辑斯谛回归分类,训练样本的分类结果见表4. 在实验过程中,发现有些参数是冗余的,或者是一些参数的显著性水平不在95%的置信区间内,因此在产生最终模型时把这些参数去掉了. 这些无用的参数为表2中标有*号的参数,因此最终模型从35个特征参数降维到26个,进一步降低了复杂度.

表4 结合单个流和多个相关流特征训练样本的分类结果

观察值	预测值				
	P2P	Email	RTSP	VoIP	百分比校正/%
P2P	2 449	8	37	6	98.0
Email	43	2 442	12	3	97.7
RTSP	86	2	2 381	31	95.2
VoIP	31	0	74	2 395	95.8
总百分比/%	26.1	24.5	25.0	24.4	96.7

以2006-09-01和2006-09-06的检验样本,输入到由训练样本得到的多项逻辑斯谛回归的模型中,检验样本的分类结果分别如表5和6所示. 从这2个表可以看出,随着间隔时间的增大,总体分类的准确性会有少许下降,但整体分类准确性较高,都还在90%以上. 因此在实际应用中,可以在分类准确率低于可以忍受的分类下限时,再进行一次训练,得到新的回归参数. 因此本文得到的回归模型对于具有时变特性的网络业务具有较强的适应性.

表5 结合单个流和多个相关流特征,2006-09-01检验样本分类结果

观察值	预测值				
	P2P	Email	RTSP	VoIP	百分比校正/%
P2P	2 499	8	38	5	98.0
Email	34	2 488	16	2	97.9
RTSP	100	12	2 363	25	94.5
VoIP	26	1	93	2 380	95.2
总百分比/%	26.6	25.1	25.1	24.1	96.4

表6 结合单个流和多个相关流特征,2006-09-06检验样本分类结果

观察值	预测值				
	P2P	Email	RTSP	VoIP	百分比校正/%
P2P	2 362	16	55	67	94.5
Email	48	2 431	19	2	97.2
RTSP	138	9	2 311	42	92.4
VoIP	30	8	72	2 390	95.6
总百分比/%	25.8	24.6	24.8	25.0	94.9

4 结束语

本文将业务分类常用的两类特征有机结合,提出一组易于提取、具有线性可分性的相关流特征用

于网络流量分类. 本文方法有 5 个显著的特点.

1) 所提特征易于提取,向量维度少,复杂度低,处理速度快.

2) 所提特征向量具有线性可分性,可以利用多项逻辑斯谛回归模型进行分类,分类模型简单,复杂度低.

3) 算法对各类业务都有较高的识别率,且特征具有泛化特性,不需频繁训练.

4) 不依赖熟知端口号和报文关键字匹配方式进行分类,适用于多种私有和加密协议.

5) 特征提取方法适用于 UDP 流、完整 TCP 流和不完整 TCP 流,算法具有较高鲁棒性.

该方法较好解决了目前网络流量分类算法存在的问题,具有较高的可用性和可靠性. 今后将进一步优化分类算法,降低算法复杂度,并研究其在实时网络流量分类中的应用.

参考文献:

[1] Logg C. Characterization of the traffic between SLAC and the Internet [EB/OL]. [2007-06-13]. http://www.slac.stanford.edu/comp/net/slac-netflow/html/SLAC-netflow.html.

[2] Moore D, Keys K, Koga R, et al. Claffy CoralReef software suite as a tool for system and network administrators [C]// Burgess M. Proceedings of the LISA 2001 15th Systems Administration Conference. San Diego: USENIX, 2001: 133-144.

[3] San S, Spatscheck O, Wang D. Accurate, scalable in-network identification of P2P traffic using application signatures[C]// Proceeding of the 13th Internation World Wide Web Conference. NY: ACM Press, 2004: 512-521.

[4] Haffner P, Sen S, Spatscheck O, et al. ACAS: automated construction of application signatures[C]// ACM SIGCOMM Workshop on MineNet 2005. Philadelphia: ACM Press, 2005: 197-202.

[5] Cisco IOS Documentation. Network-based application recognition and distributed network-based application recognition[EB/OL]. [2007-06-13]. http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122newft/122t/122t8/dtnbarad.htm.

[6] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]// ACM SIGMETRICS. New York: ACM Press, 2005: 50-60.

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

[8] Okabe T, Kitamura T, Shizuno T. Statistical traffic identification method based on flow-level behavior for fair VoIP service[C]// IEEE Workshop on VoIP Management and Security. Vancouver: IEEE Press, 2006: 35-40.

[9] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark[C]// ACM SIGCOMM 2005. Philadelphia: ACM Press, 2005: 229-240.

[10] Webb A R. Statistical pattern recognition[M]. 2nd ed. England: Wiley, 2004: 161-162.