doi:10.3969/j.issn.1001 - 893x.2013.09.017

基于 AdaBoost 的组合网络流量分类方法*

赵小欢[1,**],夏靖波[1],连向磊[2],李巧丽[3]

(1.空军工程大学 信息与导航学院,西安 710071;2.解放军 71155 部队,山东 威海 264200;
3.解放军 94326 部队,济南 250023)

摘 要:针对单一分类方法在训练样本不足的情况下对于小样本网络流分类效果差的特点,通过自适应增强(Adaptive Boosting, AdaBoost)算法进行流量分类。算法首先使用 CFS(Correlation-based Feature Selection)特征选择方法从大量网络流特征中提取出少量高效的分类特征,在此基础上,通过 AdaBoost 算法组合决策树、关联规则和贝叶斯等 5 种单一分类方法实现流量分类。实际网络流量数据测试表明,基于 AdaBoost 的组合分类方法的准确率在所选的几种算法中是最高的,其能够达到 98.92%,且相对于单一的分类算法,组合流量分类方法对于小样本网络流的分类效果具有明显提升。

关键词:网络流;流量分类;相关特征选择;自适应增强算法;组合分类器
中图分类号:TP393 **文献标志码**:A **文章编号**:1001 - 893X(2013)09 - 1207 - 06

Ensemble Classification Overnetwork Traffic Based on AdaBoost

ZHAO Xiao-huan[1], XIA Jing-bo[1], LIAN Xiang-lei[2], LI Qiao-li[3]

(1. Institute of Information and Navigation, Air Force Engineering University, Xi'an 710077, China;
2. Unit 71155 of PLA, Weihai 264200, China; 3. Unit 94326 of PLA, Jinan 250023, China)

Abstract: To cope with the poor performance of single classification algorithms on minority flows when the train dataset is deficient, the AdaBoost (Adaptive Boosting) algorithm is introduced to classify network traffic. On the basis of selecting few but effective classification features with CFS (Correlation-based Feature Selection) method from a variety of flow's features, the AdaBoost algorithm is used to combine five single classification algorithms which belong to Decision Tree, Rules and Bayes respectively for the sake of traffic classification. The experiment over real network traffic shows that the AdaBoost algorithm has the highest precision up to 98.92% among the selected classification algorithms. Moreover, the AdaBoost algorithm achieves great improvement on the performance of minority flows' classification compared with single classification algorithms.

Key words: network traffic; traffic classification; correlation-based feature selection; adaptive boosting algorithm; ensemble classifier

1 引 言

随着近年来互联网的不断发展,社交网络、在线视频、电子商务、即时通信、微博、P2P 应用等多种新兴业务不断涌现并迅速占据互联网中主流应用位置,互联网流量在组成和性质上发生了较大的变化,网络的可控可管性变得越来越差。由于不同的网络应用对于带宽、时延等指标的需求不同,不同等级用户占用的网络资源不同,仅通过网络层和传输层流量实现网络流量管理是不够充分的,而需要将网络流量映射到特定的业务,根据网络业务实现网络流量的精细划分、分级管理和差异化服务。同时,精确

* **收稿日期**:2013 - 04 - 09;**修回日期**:2013 - 06 - 18 Received date:2013 - 04 - 09;Revised date:2013 - 06 - 18
基金项目:陕西省自然科学基础研究计划重点项目(2012JZ8005)
Foundation Item: The Natural Science Basic Research Project of Shaanxi Province(2012JZ8005)
** **通讯作者**:zxhzxh_2012@163.com **Corresponding author**: zxhzxh_2012@163.com

的流量分类对网络安全、网络计费、网络规划等也具有重要的意义。

为了应对互联网流量数据庞大、结构复杂、属性动态变化的特点,利用机器学习方法挖掘流量数据从而实现流量分类成为网络流量分类的研究热点,目前已有较多文献将多种机器学习算法引入到网络流量分类中。Thuy 在文献[1]中将网络流量分类方法分为无监督算法(聚类算法)、有监督算法和半监督算法 3 类,并详细综述了 2004 ~ 2007 年间网络流量分类领域的 18 项重要工作,最后从时间复杂度与持续分类能力、方向无关性、存储与计算复杂度、健壮性与鲁棒性等方面探讨了多种流量分类方法在实际应用时面临的挑战。文献[2]选取日本、韩国和美国的 7 组流量数据并通过实验全面对比了基于端口、基于主机行为和基于流特征的流量分类方法,文中指出各种流量分类方法均存在优势及不足,并且在所对比的几种机器学习分类方法中,SVM 算法具有最高的准确性和鲁棒性。文献[3]通过对比C4.5、Naive Bayes、L7 方法在相同时间段不同观测点及不同时间段同一观测点两种情况下的流量分类效果,指出 C4.5 方法具有良好的时间空间适应性。文献[4]指出现有的机器学习流量分类方法只是在特定的条件和假设下具有良好的分类效果,在多种网络环境及不同粒度的情况下,没有哪一种机器学习算法能够始终比其他算法分类效果好,文中通过最大似然组合、D - S 证据理论、增强型 D - S 证据理论等理论融合流量分类中常见的几种分类算法,改善了网络流量分类的效果和鲁棒性。文献[5]也指出多种分类方法的有效组合是网络流量分类发展的一个重要方向。

本文通过实际网络流量数据对比了流量分类中常见的 NBTree、PART、C4.5、B_ Net(Bayes Net)、B_ Kernel(Bayes Kernel)和 SVM 6 种分类方法,发现各种单一分类算法在训练样本不足的情况下对于小样本网络流的分类效果较差,基于此,本文采用基于 AdaBoost(Adaptive Boosting)的组合分类方法组合决策树、关联规则和贝叶斯等 5 种方法进行流量分类。实验结果表明,相对于单一分类算法,基于 AdaBoost 的组合分类方法具有较高的准确性,算法在一定程度上能够有效降低单一分类算法过于依赖特定假设分布的要求,算法具有更好的鲁棒性和适用性。

2 AdaBoost 组合网络流量分类方法

采用机器学习方法实现网络流量分类主要包括两方面的工作:首先,需要按照不同粒度(常见的粒度包括 TCP Connection、Flows、Biflows、Services、Hosts 等)将网络流量归并成网络流,从网络流中选择合适的流属性构建分类特征向量;其次,需要选择合适的机器学习分类算法实现网络流量分类。本文采用的基于 AdaBoost 的组合网络流量分类方法的框架如图 1 所示。

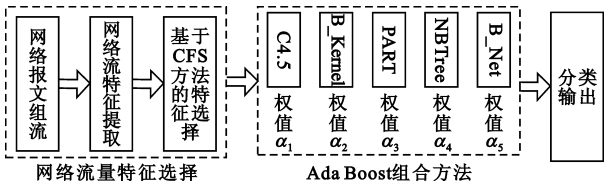

图 1 AdaBoost 组合分类框架

Fig. 1 Framework of ensemble classification based on AdaBoost

2.1 网络流量特征选择

网络流量分类特征的选择对于机器学习分类方法至关重要,过度相关或冗余的特征会对机器学习算法的性能造成负面影响,同时流量特征的增加使得机器学习算法需要的空间和时间复杂度急剧增加,因此选择足够少、但能够提供高效分类信息的特征子集十分必要[1]。

特征选择方法主要分为两种模式:过滤器方式和封装器方式。过滤器方式利用数据本身的特征作为特征子集的度量指标,而封装器方式利用机器学习算法的准确率作为特征子集的度量指标。考虑到机器学习算法的性能要求,按照 Moore 等在文献[6]中定义的网络流特征,本文首先采用 NetMate[7] 软件从网络流量中提取出 TCP 流特征,然后采用基于 CFS (Correlation-based Feature Selection)的过滤器方式从大量冗余流特征中选择合适的特征子集。

CFS 方法是一种经典的可以消除无关重复变量的基于过滤器方式的特征选择方法,使用如下变量子集评估方法给变量子集排序:

$$Merit_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}$$

其中,$Merit_s$ 表示一个包含 k 个特征的特征子集 S 的评价,$\bar{r}_{cf}$表示平均特征类别相关系数($f \in S$),$\bar{r}_{ff}$表示平均特征特征相关系数。通过该评价指标能够有效地给出特征对于分类的贡献度,从而清除贡献度低

的特征。表 1 描述了基于 CFS 方式选择的 TCP 流分类时使用的流特征。

表 1　TCP 流特征描述

Table 1 Characteristics of TCP flows

特征	描述
serv_port	server 端口号
clnt_port	client 端口号
push_pkts_serv	server 至 client 方向 PSH = 1 的 IP 包数量
push_pkts_clnt	client 至 server 方向 PSH = 1 的 IP 包数量
init_win_bytes_clnt	client 端初始窗口字节数
init_win_bytes_serv	server 端初始窗口字节数
avg_seg_size_serv	server 至 client 方向平均 IP 分片字节数
IP_bytes_med_clnt	client 至 server 方向 IP 包字节数中位数
act_data_pkt_clnt	client 至 server 方向 TCP 负载非空 IP 包数量
data_bytes_var_serv	server 至 client 方向 IP 包字节数方差
min_seg_size_clnt	client 至 server 方向最小分片大小
RTT_samples_clnt	client 至 server 方向 RTT 大小

2.2　AdaBoost 组合流量分类方法

根据“没有免费的午餐”法则可知,没有一种学习算法可以在任何领域总是产生最准确的学习器[8]。每一种机器学习算法都需要构建一个基于一组假设的某种模型,当假设在数据上不成立时,这种归纳偏倚将导致误差。学习是一个不适定问题,并且在有限的数据上,每种学习算法都收敛到不同的解,并在不同的情况下失效,可以通过性能调节使一个学习算法在确认集上达到尽可能最高的准确率,但即使对最好的学习算法也存在实例使其不能足够准确。同时由于不同地域、不同链路、不同应用产生的网络流量千差万别,文献[3]通过实验也指出即使是在特定时间、特定流量数据上性能最优的分类算法,当其适用于更长时间及不同地域的流量时,算法的分类准确性迅速降低,因此期望得到一种普遍适用的性能最优的网络流量分类算法是很难实现的。

通过多个单分类器组合能够克服单一分类器中对于某些实例分类效果差的问题,从而提升系统的分类精度。多分类器组合有两种常见类型:并联组合和串联组合。装袋法(bagging)和提升法(boosting)是并联与串联两种组合的典型代表。在 Boosting 算法中,首先需要根据已有的训练样本集选择一个准确率比平均性能要好的基分类器,分类器对样本正确分类后要降低该样本的权重,而错误分类时,则要增加错误分类样本的权重,而后加入的基分类器着重处理比较难的训练样本,最终得到一个分类准确率较高的组合分类器。

AdaBoost 算法是 Freund 和 Schapire 根据在线分配算法提出的一种利用大量分类能力一般的基分类器通过一定方法组合成分类能力强的组合分类器的方法,组合分类器为基分类器加权投票的线性组合。AdaBoost 算法的弱分类器组合方法和训练方法的有效性已经得到了证明并有大量应用验证,其中基于 AdaBoost 算法的人脸检测方法已经成为目前人脸检测最成功的方法之一[9]。本文选用 AdBoost M1 算法处理网络流量分类问题,它可以处理两种类别以上的分类问题。基于 AdaBoost 的组合流量分类算法的描述如下。

给定训练集

$$L = \{(x_1, y_1), \cdots, (x_N, y_N)\}$$

其中,x_i 表示网络流对应的特征向量,y_i 表示网络流对应的应用类型,$i = 1, 2, \cdots, N$,$y_i \in \{1, 2, \cdots, J\}$,令 T 表示基分类器的数目,令 $h_t(x)$表示第 t 个基分类器,其中 $t = 1, \cdots, T$,令 m_t 表示训练第 t 个基分类器$h_t(x)$时使用的分类方法,其中分类方法包括 C4.5、B_Kernel、PART、NBTree、B_Net 5 种类型。

(1)对于 $t = 1$,初始化样本权值 $D_t(i) = 1/N$,$i = 1, 2, \cdots, N$,以概率分布 $D_t(i)$从训练样本集 L 中可放回重复抽样得到样本数为 N 的新的训练集 L_1,使用分类方法 m_1 对训练集 L_1 进行训练得到基分类器 $h_1(x)$,应用基分类器 $h_1(x)$对原始样本集 L 上所有样本进行分类,计算错误率

$$\varepsilon_1 = \sum_{i=1}^{N} D_1(i)[y_i \neq h_1(x_i)]$$

若 $\varepsilon_1 \geqslant 0.5$,$T = 0$,算法结束;否则 $\alpha_1 = \frac{1}{2}\lg\left(\frac{1-\varepsilon_1}{\varepsilon_1}\right)$;

(2)for $t = 2, 3, \cdots, T$

$$D_t(i) = \frac{D_{t-1}(i)}{Z} \cdot \begin{cases} e^{-\alpha_{t-1}} \text{ if } h_{t-1}(x_i) = y_i \\ e^{\alpha_{t-1}} \text{ if } h_{t-1}(x_i) \neq y_i \end{cases}$$

其中,$Z = \sum_{i=1}^{N} D_t(i)$ 为确保 D_t 为概率分布的归一化参数。

以概率分布 $D_t(i)$从原始样本集 L 中可放回重复抽样得到样本数为 N 的新的训练集 L_t,使用分类方法 m_t 对训练集 L_t 进行训练得到基分类器 $h_t(x)$,应用基分类器 $h_t(x)$对原始样本集 L 上所有样本进行分类,计算错误率 $\varepsilon_t = \sum_{i=1}^{N} D_t(i)[y_i \neq h_t(x_i)]$,若 $\varepsilon_t \geqslant 0.5$,$T = t - 1$,结束;否则 $\alpha_t = \frac{1}{2}\lg\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$。

(3)组合 J 个基分类器 $h_t(x)$得到组合分类器

$H(x)$,对于 x_i,使得

$$H(x_i) = \arg \max_{y_i \in \{1,\cdots,J\}} \left\{ \sum_{j:h_j(x_i)=y_i} \alpha_j \right\}。$$

3 实验与评价

3.1 实验数据基本信息

早期的网络流量分类方法中大部分采用的均是2003 年的 Moore 数据集[6],主要原因是出于隐私的考虑,我们能够得到的网络流量基本上都去除了有效载荷,并且 IP 包头信息也采用了匿名化技术,导致研究者无法有效地获得各网络流对应的应用类型来评估自己的分类方法。一种可能的解决方案是网络流量发布者在匿名处理流量之前先采用特定的流量标记工具(如 L7-Filter、GVTS(Ground Truth Verification System))标记好网络流对应的应用类型,然后将匿名化的流量及对应的应用类型共同发布,如 UNIBS[10]数据库就提供了部分标记好的网络流量数据。由于 Moore 数据集采集的时间较早,许多新出现的网络应用在 Moore 数据集中并没有体现,基于此,本文选用文献[3]提供的一条吉比特链路 2007 年的流量数据集,该数据集为剑桥大学计算机实验室所提供,且该数据集提供了详细的应用类型标记信息以供研究者评估自己的算法。从数据集中提取30 min的 TCP 流评估流量分类方法,其中 TCP 流对应的分类结果、流数量和具体的网络应用情况如表 2 所示。

表 2 TCP 流组成

Table 2 Composition of TCP flows

分类结果	全部流数	训练集流数	网络应用
WWW	212 492	2097	Web browsers/applications
MAIL	10 871	116	IMAP/POP/SMTP
BULK	555	3	FTP/wget
ATTACK	4 008	39	Port scans/worms/viruses/ sql injections
CHAT	506	4	MSN Messenger/Yahoo IM/Jabber
P2P	17 851	190	Napster/Kazaa/Gnutella/ eDonkey/BitTorrent
MULTIMEDIA	11	1	Windows Media Player/ Real/iTunes
VOIP	1 043	17	Skype
SERVICES	465	4	X11/DNS/IDENT/LDAP/NTP
INTERACTIVE	317	6	SSH/TELNET/VNC/GotoMyPC
GAMES	150	3	Microsoft Direct Play
GRID	93	3	Grid computing

3.2 分类结果及对比

本文选用 C4.5、NBTree、PART、B_Net、B_Kernel、SVM 和 AdaBoost 组合分类方法共 7 种算法测试分类效果。其中 C4.5 与 NBTree 分类方法为基于决策树的算法,PART 分类方法为基于关联规则的算法,B_Net 与 B_Kernel 分类方法为基于贝叶斯原理的算法,SVM 支持向量机方法通过非线性映射和结构风险最小化原则实现分类,且 SVM 类型选为 C-SVM,核函数选为径向基 RBF 函数,AdaBoost 方法通过 AdBoost M1 算法组合除 SVM 外其余 5 种分类算法实现分类(通过选择不同原理、互有差异的分类方法构造基分类器,有利于实现各种方法的优势互补,提高组合分类模型的分类效果)。本文采用数据挖掘软件 Weka[11]实现 6 种单一的分类算法,而 AdaBoost 组合分类方法通过 Matlab 和 Weka 软件共同实现。

由于在实际的网络环境中获取网络流对应的应用类型是比较困难的,为了能够更加贴近实际网络环境,本文选用小训练样本集测试流量分类效果。首先从整个数据集共 248 362 条 TCP 流中无放回等概率抽取 1%共 2 483 条 TCP 流构成训练集,以整个 TCP 数据集作为测试集,训练集中各种应用类型对应的流数如表 2 所示。同时采用 AdaBoost 算法计算出的组合分类方法中各基分类器的权值,如表 3 所示。

表 3 AdaBoost 方法基分类器权值

Table 3 Weights of base classifiers for AdaBoost

基分类器	权值
C4.5	0.987 7
B_Kernel	0.307 4
PART	1.117 8
NBTree	1.121 0
B_Net	0.936 4

表 4 显示了 7 种分类算法在测试集上的分类效果,其中各种应用类型对应的单元格格式为:准确率(Precision)/召回率(Recall),例如采用 NBTree 算法分类 WWW 类型得到的结果为 0.996/0.998,该分类结果表示:分类结果为 WWW 的流中有 99.6%的流是正确的 WWW 流,同时测试集全部 WWW 流中有 99.8%的流被正确分类为 WWW 流。

表 4　TCP 流分类效果对比
Table 4 Comparison of classification accuracy of TCP flows

网络流	WWW	MAIL	BULK	ATTACK	CHAT	P2P	MULTI MEDIA	VOIP	SERV ICES	INTERA CTIVE	GAMES	GRID	总体流准确率
NBTree	0.996/ 0.998	0.987/ 0.95	0.861/ 0.391	1/ 0.987	0.928/ 0.636	0.977/ 0.945	0/0	0.317/ 0.711	0.053/ 0.009	0.735/ 0.779	0.926/1	0.308/ 1	98.603%
PART	0.994/ 0.995	0.99/ 0.963	1/ 0.515	0.94/ 0.979	0.695/ 0.589	0.958/ 0.922	0/0	0.459/ 0.623	0.167/ 0.161	0.314/ 0.795	0.5/ 0.44	0.437/ 1	98.26%
C4.5	0.996/ 0.998	0.993/ 0.995	0.658/ 0.526	0.97/ 0.984	0.913/ 0.52	0.958/ 0.958	0/0	0.432/ 0.402	0/0	0.674/ 0.801	0.408/ 1	0.877/ 1	98.809%
B_Net	0.994/ 0.987	0.958/ 0.914	0.42/ 0.658	0.986/ 0.986	0.743/ 0.607	0.983/ 0.832	0.019/ 0.455	0.299/ 0.712	0.058/ 0.312	0.579/ 0.801	0.207/ 0.78	0.207/ 1	96.793%
B_Kernel	0.993/ 0.991	0.927/ 0.931	1/ 0.524	0.998/ 0.986	0.932/ 0.516	0.912/ 0.838	1/ 0.091	0.167/ 0.509	0/0	0.552/ 0.782	1/1	1/1	97.105%
SVM	0.857/ 1	1/ 0.011	1/ 0.005	1/ 0.014	1/ 0.008	1/ 0.011	1/ 0.091	1/ 0.016	0.8/ 0.009	1/ 0.019	1/ 0.04	1/ 0.032	85.723%
Meta_AB	0.996/ 0.997	0.986/ 0.996	0.921/ 0.314	0.983/ 0.986	0.893/ 0.629	0.964/ 0.976	1/ 0.364	0.653/ 0.594	0.142/ 0.183	0.81/ 0.792	0.975/ 0.793	1/1	98.92%

从表 4 中可以看出,即使是在 1% 的抽样率的情况下,7 种算法的总体流分类准确率仍然较高,效果最差的 SVM 算法其总体流准确率也达到了 85.723%,同时表中几种单一的分类算法对于不同的应用类型分类效果高低不一,很难准确地评判出哪种算法优于其他算法。各种单一分类算法对于训练集中小样本网络流的分类效果较差,对于训练集中流数为 1 的 MULTIMEDIA 流,NBTree、PART、C4.5 3 种算法完全无法识别出 TCP 流测试集中对应的应用类型。

值得注意的是,与文献[2]得出 SVM 算法具有最高准确性和鲁棒性的结论不同,本文实验中 SVM 算法的流准确率是几种单一分类算法中最低的。对于 BULK、CHAT、MULTIMEDIA、SERVICES、INTERACTIVE、GRID 6 种小样本网络流,SVM 算法在测试集中检测到的流数与训练集中检测到的流数完全相等,这意味着 SVM 算法完全没有识别出测试集中新出现的 BULK、CHAT 等 6 种应用流。可以推断出 SVM 算法在训练样本分布不均衡的情况下为了寻求最优分类平面导致出现了过拟合(overfitting)的现象。同时,由于 SVM 算法的时间复杂度明显高于其他 5 种单一分类算法,因此组合分类方法中并未采用 SVM 算法构造基分类器。

通过对比各种单一分类算法以及组合分类方法可以看出,相比于单一的分类算法,基于 AdaBoost 的组合分类方法的分类准确率在各种算法中是最高的,算法的总体流准确率达到 98.92%,算法的分类效果要优于文献[2-3]采用的 SVM 和 C4.5 分类方法;同时对于各种小样本网络流,如训练流数为 1 的 MULTIMEDIA 流及训练流数为 4 的 SERVICES 流等,AdaBoost 组合分类方法的分类效果相对于单一分类方法具有明显的提升。

由于数据分布不均衡是网络流量的一个重要特征,在各种情况下获取的流量数据中必然存在部分网络流的规模远大于其他网络流的现象,而各种单一分类方法对于小样本网络流的分类效果波动较大,基于 AdaBoost 的组合流量分类方法通过赋予前一次分类错误的样本更高的权重实现多个基分类器的加权组合,算法能够取得更加稳定的分类效果。因此 AdaBoost 组合流量分类方法能够在一定程度上克服单一分类算法过于依赖特定数据分布的缺点,算法具有更好的实用性和鲁棒性。

4　结　论

随着互联网规模的不断扩大和新兴业务的持续涌现,互联网的可控可管性越来越差,精确的网络流量分类是实现网络可控可管的关键,同时流量分类对于网络性能、网络安全、网络计费、网络规划等也具有重要的作用。基于机器学习的流量分类方法是近年来网络流量分类领域的研究热点之一,由于传统的单一机器学习算法都需要构建基于特定假设的某种模型,不能满足复杂多变的网络流量的分类要求,本文采用基于 AdaBoost 的组合流量分类方法组合决策树、关联规则和贝叶斯共 5 种方法进行流量分类,算法在 1% 训练样本的情况下能够正确分类出测试集中 98.92% 的网络流,算法的分类效果优

于常用的单一流量分类算法,同时 AdaBoost 组合流量分类方法能够在一定程度上克服单一分类算法对于小样本网络流分类效果差的问题,算法对于待分类数据的分布要求低,具有更广泛的实用性和更好的鲁棒性。

然而 AdaBoost 组合流量分类方法同样存在小样本网络流分类效果低于大样本网络流的问题,这是由于网络流分布的不均衡以及多分类 AdaBoost 算法仅考虑错分代价总和最小而不区分不同类型代价的差异所致。为进一步提高小样本网络流的分类效果,下一步可考虑引入重抽样技术使得训练样本集的不平衡度降低,或者对于某些关键的少数类引入较大的加权系数使得这些类被错分时产生较大的代价来改进 AdaBoost 组合分类方法。

参考文献:

[1] Nguyen T T, Armitage G. A Survey of Techniques for Internet Traffic Classification using Machine Learning[J]. IEEE Communications Surveys & Tutorials, 2008,10(4):56 – 76.

[2] Kim H, Claffy K C, Fomenkov M, et al. Internet traffic classification demystified: Myths, caveats, and the best practices[C]// Proceedings of 2008 ACM CoNEXT Conference. New York: ACM, 2008:1 – 12.

[3] Li Wei, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks,2009,53(6):790 – 809.

[4] Callado A, Kelener J, Sadok D, et al. Better network traffic identification through the independent combination of techniques[J]. Journal of Network and Computer Applications, 2010,33(4):433 – 446.

[5] Dainotti A, Pescape A, Claffy K C. Issues and Future Directions in Traffic Classification[J]. IEEE Network, 2012,26(1):35 – 40.

[6] MooreA W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]//Proceedings of 2005 International Conference on Measurement and Modeling of Computer Systems. Banff, AB, Canada: ACM, 2005: 50 – 60.

[7] Schmoll C, Zander S. Network Measurement and Accounting Meter[EB/OL]. [2013 – 04 – 08]. http://sourceforge.net/projects/netmate – meter/.

[8] 范明, 昝红英, 牛常勇. 机器学习导论[M]. 北京: 机械工业出版社, 2009: 230 – 231.
FAN Ming, ZAN Hong-ying, NIU Chang-yong. Introduction to Machine Learning[M]. Beijing: China Machine Press, 2009:230 – 231.(in Chinese)

[9] Viola P, Jones M. Robust real-time face detection[J]. International Journal of Computer Vision,2004,57(2):137 – 154.

[10] 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.

[11] Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques[M]. 3rd ed. San Francisco: Morgan Kaufmann Publishers,2011:424 – 438.

作者简介:

赵小欢(1984—),男,湖北枣阳人,2009 年于空军工程大学获通信与信息系统专业硕士学位,现为博士研究生,主要研究方向为网络流量测量;

ZHAO Xiao-huan was born in Zaoyang, Hubei Province, in 1984. He received the M.S. degree from Air Force Engineering University in 2009. He is currently working toward the Ph.D. degree. His research concerns network traffic measurement.

Email:zxhzxh_2012@163.com

夏靖波(1963—),男,河北秦皇岛人,教授、博士生导师,主要研究方向为军事信息网络管理与安全;

XIA Jing-bo was born in Qinhuangdao, Hebei Province, in 1963. He is now a professor and also the Ph.D. supervisor. His research concerns military information network management and security.

连向磊(1981—),男,山东荣城人,硕士,工程师,主要研究方向为军事信息网络管理;

LIAN Xiang-lei was born in Rongcheng, Shandong Province, in 1981. He is now an engineer with the M.S. degree. His research concerns military information network management.

李巧丽(1983—),女,山东聊城人,硕士,工程师,主要研究方向为网络测量与管理。

LI Qiao-li was born in Liaocheng, Shandong Province, in 1983. She is now an engineer with the M.S. degree. Her research concerns network measurement and network management.