

基于相关向量机的在线网络流量分类方法

夏靖波,柏 骏,赵小欢,吴吉祥

(空军工程大学 信息与导航学院, 西安 710077)

摘要:在研究、分析分类结果预测概率及其对分类准确率影响的基础上,提出了一种新的基于相关向量机(RVM)的在线网络流量分类方法:首先,利用RVM对网络流量分类,输出分类结果预测概率;对于置疑区间 $[0.1, 0.9]$ 内的网络流,采用“端口号+深度数据包检测(DPI)”相结合的方法重新进行识别;对于预测概率处于 $[0, 0.1]$ 和 $[0.9, 1]$ 区间的分类结果则完全采纳。实验表明:该方法的整体分类准确率能达到98%左右,且实时性较好。

关键词:计算机应用; 流量分类; 相关向量机; 流量特征; 置疑区间

中图分类号:TP393 **文献标志码:**A **文章编号:**1671-5497(2014)02-0459-06

DOI:10.13229/j.cnki.jdxbgxb201402029

Online network traffic classification using relevant vector machine

XIA Jing-bo, BAI Jun, ZHAO Xiao-huan, WU Ji-xiang

(Information and Navigation College, Air Force Engineering University, Xian 710077, China)

Abstract:Based on the research and analysis of probabilistic classification and its influence on the overall accuracy, a new online traffic classification method is proposed. First, the Relevant Vector Machine (RVM) is used to classify traffic flows and output probabilistic classification. Then, the flows, whose classification probability is in doubting interval $[0.1, 0.9]$, are re-identified by using port & Deep Packet Inspection (DPI). If the predicted probability is in the interval $[0, 0.1]$ and $[0.9, 1]$, the classification is totally accepted. Experiment studies demonstrate that the proposed method can reach the overall accuracy of 98%, and perform well in online network traffic classification.

Key words: computer application; traffic classification; relevant vector machine; traffic features; douting interval

0 引言

实时流量分类技术能够对在线网络流量按照应用类型分类,对于网络管理、流量控制以及网络相关研究具有重要意义。目前主要的流量识别、分类技术有4种,分别是端口识别技术^[1]、深层数

据包检测技术(DPI)^[2-3]、基于网络行为的流量分类技术^[4]和基于机器学习的流量分类技术^[5-7]。上述方法中,除了基于网络行为的流量分类技术,其他方法均可应用于在线网络流量的分类,但仍存在诸多不足之处。端口识别技术随着端口跳变技术、伪装技术的出现已不再可靠;DPI则由于需

收稿日期:2012-11-08.

基金项目:陕西省科技计划自然基金重点项目(2012JZ8005);全军军事学研究生课题项目.

作者简介:夏靖波(1963-),男,教授,博士生导师.研究方向:通信网络管理与评估. E-mail:jbxiad@sina.com

通信作者:柏骏(1985-),男,博士研究生.研究方向:网络管理与信息安全. E-mail:peking1985-2005@163.com

要对每个数据包进行深度检测,时间、空间复杂度较高,在高速网络中难以满足实时性要求;基于机器学习的流量分类方法具有分类准确率高的优点,但需预先训练分类模型且测试时间相对较长,难以在高速网络中胜任在线分类的任务。

相关向量机(RVM)^[8]是一种新的机器学习方法,与支持向量机(SVM)具有相同的决策形式,通过引入稀疏贝叶斯学习理论,不仅使其具备了 SVM 避免过学习的优点,且极大地减少了核函数的计算量,还克服了 SVM 存在的稀疏性不强、计算量大、核函数必须满足 Mercer 条件以及需人为凭经验确定参数等缺点。因此,与 SVM 相比,RVM 更稀疏,测试时间更短,更适用于在线分类。另外,RVM 利用概率模型来解释数据中的噪声,使得 RVM 具备了预测结果概率的能力,通过概率预测可进一步加深对分类结果的理解。目前 RVM 已被广泛应用于语音识别^[9]、高光谱影像分类^[10]等领域中,取得了良好的分类效果。

本文将 RVM 应用于在线网络流量分类,通过引入置疑区间概念,研究并分析了预测概率对分类结果的影响,基于此,提出了一种新的基于 RVM 的在线网络流量分类方案。该方案以网络流开始阶段若干数据包组成的子流作为研究对象,采用 RVM 与“端口号+DPI”相结合的混合机制对网络流进行在线识别,具有分类准确率高、处理速度快的特点。

1 RVM 原理

在监督学习中,给定一组输入样本 $X = \{x_n\}_{n=1}^N$,其中 N 为样本的个数,对应的目标输出 $t = \{t_n\}_{n=1}^N$,在回归问题中, t_n 可以是任意值,在分类问题中, t_n 是类别标号(二元分类时可以是 0 或 1)。

RVM 采用了与支持向量机相同的决策形式:

$$\begin{aligned} t_n &= y(x_n; \omega) + \epsilon_n = \\ &\sum_{n=1}^N \omega_n K(x, x_n) + \omega_0 + \epsilon_n \end{aligned} \quad (1)$$

式中: $K(x, x_n)$ 为选用的核函数; $\{\omega_n\}_{n=0}^N$ 为不同的权重; ϵ_n 为噪声,假设其服从均值为零、方差为 τ^2 的高斯分布。

引入了贝叶斯概率模型来解释噪声对预测结果的影响,这样不仅能很好地解决 SVM 中误差

参数难以确定的问题,还获得了预测结果概率的能力,这也是 RVM 的核心概念。贝叶斯概率为

$$p(t_* | t) = \int p(t_* | w, \sigma^2) p(w, \sigma^2 | t) dwd\sigma^2 \quad (2)$$

为了使 w 中大部分元素为 0,参数 w 服从零均值的高斯先验分布:

$$p(w, \alpha) = \prod_{n=0}^N N(\omega_n | 0, \alpha_n^{-1})$$

这里每一个权值 ω_n 都独立地对应一个参数 α_n 。

根据贝叶斯准则进行推理,有:

$$\begin{aligned} p(t_* | t) &= \\ \int p(t_* | w, \sigma^2) p(w | t, \alpha, \sigma^2) p(\alpha, \sigma^2 | t) dwd\alpha d\sigma^2 &\approx \\ \int p(t_* | w, \partial_{MP}, \sigma_{MP}^2) p(w | t, \partial_{MP}, \sigma_{MP}^2) dw & \end{aligned} \quad (3)$$

用拉普拉斯方法将 $p(w | t, \partial)$ 近似为高斯分布,有:

$$\partial_i^{\text{new}} = \frac{\gamma_i}{\mu_i^2}, (\sigma^2)^{\text{new}} = \frac{\|t - \Phi\mu\|^2}{N - \sum_{i=0}^N \gamma_i}, \gamma_i = 1 - \partial_i \sum_{i,i}$$

通过不断地迭代,以逼近 ∂_{MP} 及 σ_{MP}^2 。其中大部分 ∂_i 趋近无限大,对应的 ω_i 则为零;少量的 ω_i 稳定趋于有限值,对应的 x_i 即为相关向量。

应用回归方法中解决分类问题的 Logistic 模型:

$$p(t_n = 1 | \omega^\top \varphi(x_n)) = \frac{1}{1 + \exp(-\omega^\top \varphi(x_n))}$$

有结果预测概率:

$$p(t | \omega) = \prod_{i=1}^N \sigma[y(x_n; \omega)]^{t_i} \{1 - \sigma[y(x_n; \omega)]\}^{1-t_i} \quad (4)$$

RVM 判别准则为:如果 $p_i = \frac{1}{1 + \exp(-y)} < 0.5$,则 $t_i = 0$;反之,则 $t_i = 1$ 。

2 在线网络流量分类方案

2.1 实验数据

为了便于对比、分析,本文采用 2 个实验数据集。一个是 Moore 等人在文献[11]中所用的实验数据集——Moore_Set,主要用于置疑区间的说明;另一个数据集采自某大学网络工程实验室,记为 IN_Set,主要用作在线分类方案的验证。

在文献[12]中有关于 Moore_Set 的详细说明,在此不再赘述。由于 Moore_Set 以完整的

TCP流作为流量样本,忽略了目前大量存在于网络中的UDP数据包,尤其是占用大量带宽的P2P应用,其UDP数据包甚至占到P2P数据包的60%左右。为此,本文按照五元组,以间隔T(T=120 s)为流结束标志,采集TCP/UDP双向网络流作为流量样本,并按照NetMate^[13]中定义的44种网络流量属性特征提取流量特征,产生数据集IN_Set。

IN_Set数据集包含2个数据集,第1个数据集捕获于2012年5月22日14:30~17:30,记为IN_Set1;第2个数据集捕获于2012年9月7日14:30~17:30,记为IN_Set2。两个数据集都用端口号、DPI技术与人工处理相结合的方法对数据集进行应用识别,对于无法识别的流量标记为未知,实验数据中不包括未知流量。数据集的统计信息如表1、表2所示。

表1 IN_Set1数据集的统计信息

Table 1 Statistics of IN_Set1

Type of flow	Num of flow	Percent/%
WWW	69101	48.02
FTP	9176	6.38
MAIL	7429	5.16
DataBase	5648	3.93
P2P	40063	27.84
GAME	938	0.65
SERVER	241	0.17
Unknown	11307	7.85
Total	143903	100.00

表2 IN_Set2数据集的统计信息

Table 2 Statistics of IN_Set2

Type of flow	Num of flow	Percent/%
WWW	72906	49.08
FTP	8253	5.56
MAIL	7134	4.80
DataBase	6479	4.36
P2P	39457	26.57
GAME	1058	0.72
SERVER	461	0.31
Unknown	12769	8.60
Total	148517	100.00

2.2 置疑区间

对于二元分类,RVM不仅可以获得二值输出,还能得到结果的概率预测。预测概率越趋近

于0或1,其预测的准确性也就越高。而对于某一区间的预测值,其预测准确性有明显下降,分类结果存在较大的不确定性,其分类正确性值得怀疑,可称之为置疑区间。

为了研究置疑区间范围及其对分类准确性的影响,本文实验分别以Moore_Set和IN_Set数据集为实验对象。在第1组实验中,将entry01作为训练集,其余9个数据集作为测试集;在第2组实验中,分别将IN_Set1、IN_Set2均分为两组,前一组作为训练集,后一组作为测试集。利用CSF算法选取流量特征,对训练集随机抽样,建立分类模型,两组实验结果都取均值,如图1所示。

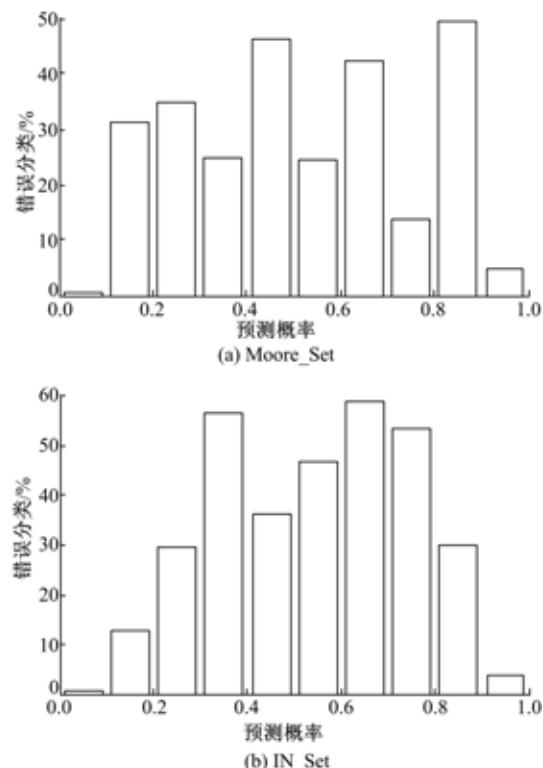


图1 分类错误率在预测概率上的分布

Fig. 1 Errors rate distribution on P-RVM

从图1可以看出,当分类结果预测概率p在[0,0.1]区间时,预测错误概率最低(在Moore_Set中错误率仅为0.3%,在IN_Set中为0.6%),即分类准确率最高;在[0.9,1]区间内,预测错误率也相对较低(在Moore_Set中错误率为4.85%,在IN_Set中为3.74%)。而在此之外的概率区间,预测错误概率明显上升。在Moore_Set数据中,有3个区间段的预测错误率达到了40%以上;在IN_Set数据中,有4个区间段的预测错误率达到了40%以上,其中[0.3,0.4]、[0.6,0.7]和[0.7,0.8]区间的错误率更是在

50%以上。通常情况下,当错误率达到50%左右时,预测的正确性与猜测相当,可以认为该预测是无效的。

进一步研究发现,上述两组实验预测概率在[0,0.1]和[0.9,1]区间的分类结果占到总样本数的90.47%(Moore_Set)或94.58%(IN_Set),而在这两个区间内的分类准确率则高达98.79%(Moore_Set)或98.02%(IN_Set)。此外,在置疑区间[0.1,0.9]内的分类准确率仅为65.72%(Moore_Set)或40.12%(IN_Set),远不能满足分类的准确性需求。因此,本文定义区间[0.1,0.9]

为置疑区间。

2.3 在线网络流量分类方案

基于以上特点,为了获得更高的分类准确率,本文提出一种新的混合流量分类方法,其流程图如图2所示。方案由离线和在线两部分组成,前者基于离线数据包选取适当的流特征以及匹配特征,并利用RVM训练流量分类模型;后者在线捕获数据包形成网络流,通过RVM分类模型以及“端口号+DPI”实时识别网络流。当RVM预测概率在置疑区间时,将“端口号”和“DPI”联合使用,以提高分类准确性。方案在线部分步骤如下:

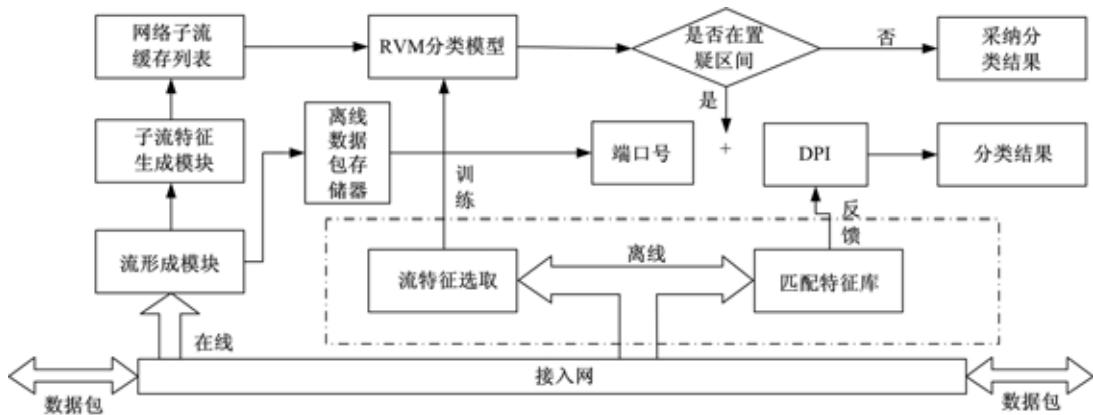


图2 基于RVM的在线网络流量分类方案

Fig. 2 Online network traffic classification using RVM

(1)在线捕获数据包,形成网络流并提取子流特征,将网络流前20个数据包存储于离线数据包存储器。

(2)将采集到的网络子流存储于缓存列表,当缓存满或一定时限后将列表内网络子流输入RVM分类模型,并清空缓存。

(3)RVM分类模型输出预测概率,判断预测概率是否在置疑区间[0.1,0.9]:如果是,按步骤(4)执行;否则,完全采纳分类结果。

(4)若该网络子流端口号在常用端口映射表中,且在匹配特征库中无相应特征字段,则按端口号标记应用类型;若在匹配特征库中有相应特征字段,按照特征字段标记应用类型;否则,标记为未知。

3 实验

3.1 实验数据预处理

3.1.1 流量特征选择

实时流量分类技术研究致力于发现一种在线识别、分类网络流量应用类型的方法。Netmate

中所述的流量特征将不能完全适用于实时流量分类,因为其中某些特征无法通过统计网络流前n个数据包获得。表3列举了可用于实时流量分类的主要流量特征,其最大特点是能随时通过在线计算得到,且时间、空间复杂度低。

利用CFS(Correlation-based feature selection)方法对数据集的流量特征进行筛选,选取与分类结果相关性最强且互相关最弱的5项流量特征,如表4所示。

表3 在线网络流量特征

Table 3 Features of online traffic flow

流量特征	TCP/UDP	时间复杂度	空间复杂度
前(后)向数据包个数	TCP/UDP	O(n)	O(1)
前(后)向数据包总(最小、最大、平均)长度	TCP/UDP	O(n)	O(1)
前(后)向数据包平均(最小、最大)到达间隔	TCP/UDP	O(n)	O(1)
子流持续时间	TCP/UDP	O(n)	O(1)
前(后)向标识位(PSH、URG)个数	TCP	O(n)	O(1)

表4 CFS 算法选取的特征子集
Table 4 Features selected by CFS

特征	特征描述
dstport	目的端口号
minFpktLen	最小前向报文长度
meanFiat	平均前向报文到达间隔
duration	子流持续时间
bUrgCnt	后向报文中有紧急标志位的包的个数

3.1.2 标准化处理

通过前期实验发现,若直接使用原始数据进行训练、分类,其分类效率低下,准确性不高且相关向量较多。究其原因,主要有两点:首先,在流量特征的统计过程中不同特征采用不同测量单位,各变量大小在数值上差异很大,直接使用原始数据进行分类可能导致信息丢失并引起数值计算的不稳定;其次,大部分流量样本数据过度集中而少量数据偏离度较大,导致 RVM 难以产生有效相关向量以区分不同类型数据。因此,需对流量特征向量做预处理。

首先考虑采用线性归一化方法对其进行标准化处理,但实验结果发现该方法并没有有效解决少量数据偏离度过大的问题。然而,利用对数函数进行标准化,数据中的零元素以及极小元素将会使得数据之间的偏离度更大。因此,出于减小数据偏离度、尽量合理分布样本数据的考虑,本文采用如下方法

$$y = \left(n \cdot \frac{x - x_{\min}}{x_{\max} - x_{\min}} \right)^{\frac{1}{n}} \quad (5)$$

对数据进行标准化处理。实验表明,当 $n=13$ 时,即可达到令人满意的分类准确率,同时相关向量数量亦可接受。

3.2 结果及分析

3.2.1 子流的选取对分类的影响

对于某一类样本,其类召回率记为 $recall_i = \frac{TP_i}{TP_i + FN_i}$; 对于所有样本,有整体准确率:

$$OverallAccuracy = \frac{\sum_{i=1}^m TP_i}{\sum_{i=1}^m (TP_i + FN_i)} \quad (6)$$

整体准确率为相关领域研究人员广泛采纳,它反映了分类模型正确预测样本数在预测总数中的比例。本文主要采用整体准确率作为评价工具。

Bernaille 在文献[14]中指出,利用网络流前 m 个数据包的流量特征即可达到区分流量应用类型的目的。为了选取合适的 m ,本文以 IN_Set 数据集为实验对象,将 IN_Set1、IN_Set2 分为 IN_Set1a、IN_Set1b、IN_Set2a 和 IN_Set2b 四组。其中 IN_Set1a 和 IN_Set2a 作为训练集;IN_Set1b 和 IN_Set2b 作为相应的测试集,分别进行两组实验。取 $m=10, 11, \dots, 20$, 提取网络流前 m 个数据包并产生流样本,对训练集随机抽样,建立分类模型,两组实验结果都取平均值,实验结果如图 3 所示。

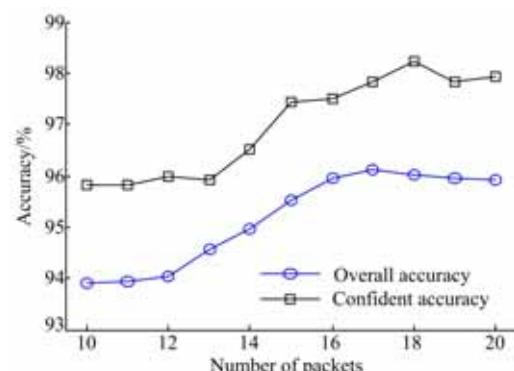


图3 整体分类准确率和可信准确率

Fig. 3 Overall and confident accuracy

图 3 中,“Overall accuracy”表示利用 RVM 分类网络子流的整体分类准确率,“Confident accuracy”表示分类结果预测概率在 $[0, 0.1]$ 和 $[0.9, 1]$ 区间内的可信准确率。从图 3 可以发现,整体分类准确率稳步上升,并在 $m=17$ 时取得最高整体分类准确率 96.12%,随后稍有下降并趋于稳定;可信准确率呈抖动上升,在 $m=18$ 时达到最高,即 98.25%。为了在随后的混合分类方案中取得更高的分类准确性,本文选取 $m=18$ 作为网络子流的数据包数。

3.2.2 在线流量分类方案分类结果

按照 2.3 节所述的在线流量分类方案对实验数据 IN_Set1 和 IN_Set2 进行分类,其分类准确率如表 5 所示。

从表 5 可以看出,方案取得了良好的分类效果,整体分类准确率达到了 98% 左右。另外还可以发现,在数据中所占比重较大的应用类型流量的识别准确率都相对较高,达到了 98% 左右,其中 WWW 应用类型流量更高,达到 99%;相反,在实验数据中所占比重较小的应用类型流量的识别准确率却相对较低,其中 GAME 流量的识别

率不到 80%。究其原因,第一,在 RVM 流量分类过程中此部分的流量分类准确率不高;第二,

GAME 类型流量由 UDP 包组成,难以通过“端口号+DPI”方法识别。

表 5 在线流量分类方案分类结果

Table 5 Accuracy of online network traffic classification

数据集	WWW	FTP	MAIL	DataBase	P2P	GAME	SERVER	Overall
IN_Set1	99.24	98.17	95.48	92.15	97.83	79.83	97.9	98.09
IN_Set2	99.15	96.28	96.35	93.52	97.46	70.63	95.17	97.83

4 结束语

RVM 不仅能提供二值输出,还能提供概率预测。本文分析、研究了分类结果预测概率及其对分类准确率的影响,发现在置疑区间(本文实验中为[0.1,0.9])内的分类准确率难以满足网络流量分类的准确性要求。因此,本文基于 RVM 的预测概率提出了一种混合的在线网络流量分类方案。该方案在线捕获网络流开始阶段的 18 个数据包作为网络子流,随后利用 RVM 训练的分类模型对网络子流进行分类,输出的预测概率若在置疑区间[0.1,0.9]内,则采用“端口号+DPI”相结合的方法重新进行识别,否则完全采纳。实验表明,该方案可实现在线的网络流量分类,且分类准确率较高。同时,在本文实验中遇到的一些问题还需进一步深入、系统研究,在线流量分类方案的工程实现将是下一步的工作重点。

参考文献:

- [1] Internet Assigned Numbers Authority (IANA)[EB/OL]. [2010-08-28]. <http://www.iana.org/assignments/port-numbers>
- [2] Moore A W, Papagiannaki D. Toward the accurate Identification of network applications[C]// Proc 6th Passive Active Measurement Workshop (PAM), Boston, MA, USA, 2005.
- [3] Sen S, Spatscheck O, Wand D. Accurate, scalable in-network identification of P2P traffic using application signatures[C]// Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters (WWW'04). New York, USA: ACM, 2004.
- [4] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark [C]// ACM SIGCOMM. Phila-delphia, PA, USA, 2005.
- [5] Dainotti A, Pescape A, Sansone C. Issues and future directions in traffic classification[J]. IEEE Network, 2012, 26(1):35-40.
- [6] Gu Cheng-jie, Zhang Shun-yi, Sun Yan-fei. Real-time encrypted traffic identification using machine learning[J]. Journal of Software, 2011, 6(6):1009-1016.
- [7] 孙知信, 张玉峰. 基于多维支持向量机的 P2P 网络流量识别模型[J]. 吉林大学学报:工学版, 2010, 40(5): 1298-1302.
- [8] Sun Zhi-xin, Zhang Yu-feng. P2P network traffic identification model based on MSVM[J]. Journal of Jilin University (Engineering and Technology Edition), 2010, 40(5): 1298-1302.
- [9] Tipping M. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001, 1(1): 211-244.
- [10] Yang Cheng-fu, Zhang Yi. Study to speaker recognition using RVM[J]. Journal of Electronic Science and Technology, 2010, 39(2):311-315.
- [11] Mianji F A, Zhang Y. Robust hyperspectral classification using relevance vector machine[J]. Geoscience and Remote Sensing, 2011, 49(6):2100-2112.
- [12] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]// Proc of the 2005 ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems, Banff, 2005.
- [13] NetMate [EB/OL]. [2012-07-03]. <http://sourceforge.net/projects/netmate-meter/>.
- [14] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2):23-26.