基于SVM概率输出的P2P流媒体识别法

陈 伟 兰巨龙 张建辉 杜锡寿
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 P2P流媒体占用大量带宽,且容易传播病毒,有必要对其进行识别。分析了Abacus方法的不足,提出一种基于SVM概率输出的P2P流媒体识别法P-Abacus。P-Abacus将待识别样本属于已知应用可能性的大小反映在概率输出上。对输出结果进行排序,根据最大概率,判决样本是属于最大概率类应用还是未知应用,或是需要进一步判断。若需进一步判断,则通过计算前两大类构建SVM概率输出的差值,来判断样本是属于其中的一类,还是未知应用。由于SVM概率输出包含大量可用信息,使得P-Abacus具有更好的识别效果。实验表明,P-Abacus比Abacus具有更高的识别率和更低的误判率,且时间开销增加有限。

关键词 P2P流媒体,识别,SVM,概率输出,端点
中图法分类号 TP393 **文献标识码** A

P2P Streaming Media Recognition Method Based on SVM Probabilistic Output

CHEN Wei LAN Ju-long ZHANG Jian-hui DU Xi-shou
(National Digital Switching System Engineering & Technological Research Center, Zhengzhou 450002, China)

Abstract P2P streaming media has taken up a lot of bandwidth, and is prone to spread the virus, so is required to be identified accurately. This paper analyzed the shortcomings of the method P-Abacus, and proposed a kind of P2P streaming media recognition method P-Abacus based on SVM probabilistic output. P-Abacus can express it in probabilistic output, which reflects the extent of the sample belonging to known applications. We ordered the output, and according to maximum probability, made a judgement on that whether the sample belongs to class of maximum probability or unknown, or needs a further judgement. If a further judgement is needed, we calculated the probabilistic output difference of the SVM built between the two largest classes, and made sure that whether the sample belongs to one of the two largest classes, or unknown. Thus P-Abacus has a better recognition effect, because probabilistic output contains more information that can be utilized. Experiments show that P-Abacus has a higher recognition rate and a lower false positive rate than Abacus, and has a limited increase of time overhead.

Keywords P2P streaming media, Recognition, SVM, Probabilistic output, Endpoint

1 引言

随着用户数量的急剧增长,网络中P2P流媒体[1]的流量迅速增加,这已影响到传统C/S模式业务的正常开展。尽管互联网带宽一直在扩容,电子商务、电子政务等关键业务仍受到带宽资源不足的影响。此外,P2P流媒体节点加入与退出的随意性,也使得攻击者能够很容易地传播有害内容,且该攻击方式成本低,危害范围广泛。因此,对网络服务提供商(ISP)而言,有必要对P2P流媒体流量进行合理的管控,而能够精确识别P2P流媒体流量则是有效管控的前提,也是当前网络流量识别领域的研究重点与难点。

目前,国内外针对P2P[2]流量以及P2P流媒体[3]流量的粗识别(Coarse Identification)研究较多,识别技术相对成熟,而针对P2P流媒体流量的精细化识别(Fine-Grained Identification)则研究较少。P2P流媒体由于协议私有、版本更新快等特性,使得基于关键字匹配的深度包检测法[4](Deep Packet Inspection,DPI)难以适用。基于机器学习的深度流检测法[5](Deep Flow Inspection,DFI)尽管不需要对载荷特征进行检测,就能够识别加密应用,但由于P2P流媒体中同时存在的控制流与数据流统计特征差异较大[6],因此该方法识别效果不佳。与此同时,大量针对P2P流媒体的流量测量[7,8]表明:P2P流媒体稳定运行时,流量主要由UDP协议产生,且集中在客户机一个或两个端点(endpoint)之上,其端点特性反映出P2P流媒体应用的总体特性。因此,已有的3种识别方法都是针对UDP端点进行的识别研究:1)KISS法[9],它运用类开方测试法从应用层负载的前N个字节提取统计特征,结合多分类SVM实现了流量的识别。KISS本质上是一种包检测法,其对加密应用无法识别,且计算复杂度高。2)PSD法[10],它对端点处的下行流量进行包长分布统计,结合SVM实现了流量的识别。由于它对包长的统计精确到1字节的粒度,因此其识别效果极佳。但是当应用软件版本更新或网络环境改变时,其包长分布特征将变化较大,识别的准确率会急

到稿日期:2011-12-26 返修日期:2012-04-24 本文受973计划项目(2012CB315900)资助,E-mail:709800535@qq.com。
陈 伟(1987—),男,硕士生,主要研究领域为互联网流量识别。

剧下降。3)Abacus 法[11,12]，它对端点处下行流量的包个数及字节数进行分布统计，结合多分类 SVM 及巴氏距离(Bhattacharyya Distance，BD)进行应用识别。该方法选取的特征鲁棒性好，识别的准确率高。但由于 SVM 采用硬判决的方式，未能考虑识别为不同应用可能性的比较，易产生误判。此外，不同应用的 BD 拒绝门限，应根据训练集的密集程度而定，尽管文献[12]找到较合适的统一值，但是当训练集发生改变或应用种类较多时，单一的拒绝门限 R 将不再适用。

本文针对 Abacus 的识别部分进行改进，提出了一种基于 SVM 概率输出的 P2P 流媒体识别法 P-Abacus。P-Abacus 将 SVM 的输出表示为属于不同应用的概率，概率能够反映待识别样本距离训练应用的远近程度，可以使判决门限得到统一。此外，P-Abacus 的二次判决也能够有效降低误判率。实验结果表明，P-Abacus 比 Abacus 具有更高的识别率和更低的误判率，且计算复杂度增加有限。

2 Abacus

2.1 基本原理

Abacus 采用机器学习的方法对 P2P 流媒体进行识别，其识别部分主要包括：端点特征提取、SVM 识别、BD 判决。

1)端点特征提取

Abacus 分别基于包个数和字节数对端点的分布特征进行统计。假定 ΔT 时间内，有 K 个端点向待识别端点 E 传递数据包，传递的包个数为 $\{P_1,\cdots,P_i,\cdots,P_K\}$。定义 $P+1$ 个区间 $\{I_0,\cdots,I_i,\cdots,I_P\}$，其中 $I_0=(0,1]$，$I_P=(2^P,\infty)$，其它 $I_i=(2^{i-1},2^i]$。若第 i 个端点传递的包个数 P_i 落在区间 I_j 内，区间 I_j 的计数值 n_j 增加 1(n_j 的初始值为 0)，得到 $P+1$ 维计数值 $(n_0,\cdots,n_P)$，其中 $\sum_{i=0}^{P}n_i=K$。对 $(n_0,\cdots,n_P)$ 进行归一化，得到 $(\bar{n}_0,\cdots,\bar{n}_P)$ 为基于包个数的统计特征，其中 $\bar{n}_i=n_i/K$。同样，针对字节数进行端点分布特征统计，得到特征 $(\bar{n}_0,\cdots,\bar{n}_B)$。将 $(\bar{n}_0,\cdots,\bar{n}_P)$ 与 $(\bar{n}_0,\cdots,\bar{n}_B)$ 联合，最终得到 $B+P+2$ 维的 Abacus 特征，即为一个 signature。

2)SVM 识别

SVM[13]一般用于二元分类，它在训练集中寻找最优分类面，得到判决函数 $f(x)=\mathrm{sgn}\{w^*x+b^*\}$。通过 Lagrange 变换，判决函数得到其对偶表达式 $f(x)=\mathrm{sgn}\{\sum_{i=1}^{l}a_i^*y_i(x_ix)+b^*\}$。当训练集中样本线性不可分时，通过引入核函数，SVM 将特征空间向高维映射，与内积运算进行融合，最终得到判决函数 $f(x)=\mathrm{sgn}\{\sum_{i=1}^{l}a_i^*y_iK(x_i,x)+b^*\}$。为了实现多元分类，Abacus 采用一对一方法(One-Against-One)为任意两个类构建超平面，训练得到 $k*(k-1)/2$ 个二元 SVM 分类器。测试时，得票最多(Max Wins)的类，即判为测试样本所属的类。

3)BD 判决

为了实现未知应用的判决，Abacus 引入巴氏距离(BD)对以上 SVM 的识别结果进行判断。$BD(p,q)=\sqrt{1-B}$ 定义了不同离散分布 p、q 之间的迥异程度，其中 $B=\sum_{k=1}^{d}\sqrt{(p(k)q(k))}$。显然，BD 值越小，$p$、$q$ 越相似。Abacus 基于 $(\bar{n}_0,\cdots,\bar{n}_B)$ 进行判断，当 SVM 识别出样本 n 属于应用 C 时，$BD(n,\bar{n}(C))$ 表示该样本所属应用与 C 应用的迥异程度，$\bar{n}(C)$ 表示训练集中 C 样本的分布中心。定义拒绝门限 R，若 $BD(n,\bar{n}(C))$ 小于 R，则接受 SVM 判决；否则，识别为未知应用。

2.2 问题分析

Abacus 采用 SVM 硬判决及 BD 拒绝门限的方式，存在以下 3 点不足：

1)SVM 的符号判决输出，无法比较识别为不同应用可能性的大小，其针对最优分类面附近的样本由于不能细致识别，易产生误判；

2)BD 判决只对 SVM 的识别结果进行接受与拒绝，在 SVM 判决时，应用 A 一旦被误判为应用 B，就无法改判；

3)对于训练集中的各种应用，由于密集程度不同，BD 判决的拒绝门限 R 应当不同，因此 Abacus 采用同样的拒绝门限不合理且难以确定。

3 P-Abacus

针对 Abacus 的不足，本文提出 P-Abacus 方法，它采用 SVM 概率输出的形式，将识别结果表示为属于不同应用的概率，通过概率判决，有效地解决了 Abacus 误判率高及单一拒绝门限的问题。P-Abacus 与 Abacus 的不同之处在于 SVM 的概率输出与概率判决。

3.1 概率输出

文献[12]中的实验表明，被误判的测试样本通常出现在 SVM 最优分类面附近，因此需要对该类样本仔细识别。Abacus 方法的 SVM 输出 $f(x)=\mathrm{sgn}\{\sum_{i=1}^{l}a_i^*y_i(x_ix)+b^*\}$，它不能反映待识别样本距离最优分类面的远近程度，这是导致误判的主要原因。因此，P-Abacus 针对 SVM 的输出进行修改，得到概率输出形式[14]：$P(C_{+1}|x)=\frac{1}{1+e^{-g(x)}}$ 及 $P(C_{-1}|x)=\frac{1}{1+e^{g(x)}}$，其中 $g(x)=\sum_{i=1}^{l}a_i^*y_i(x_ix)+b^*$。$P(C_{+1}|x)$ 表示识别为正类的概率，$P(C_{-1}|x)$ 表示识别为负类的概率。则一对一多分类时，样本 x 属于 C_i 类的概率为 $P(C_i|x)=\frac{\sum_{j=1,j\neq i}^{N}P_{ij}(C_i|x)}{\sum_{k=1}^{N}(\sum_{j=1,j\neq k}^{N}P_{kj}(C_k|x))}$，其中 $P_{ij}(C_i|x)$ 表示第 i 类与第 j 类 $(i\neq j)$ 构成的 SVM 中计算得到属于 C_i 类的概率。最终的概率输出为 $(P(C_1|x),\cdots,P(C_i|x),\cdots,P(C_N|x))$，其中 N 为训练集中应用种类的数目。概率输出能够反映待识别样本属于不同应用的概率，有效地对其距离最优分类面的远近程度进行归一化，这是概率判决中不同应用使用共同判决参数的前提。

3.2 概率判决

训练时，对有监督数据集中所有已知应用(训练集包含的应用)及未知应用(训练集未包含的应用)的实例概率输出 $(P(C_1|x),\cdots,P(C_i|x),\cdots,P(C_N|x))$ 进行排序，得到 $(P(C_{(1)}|x),\cdots,P(C_{(i)}|x),\cdots,P(C_{(N)}|x))$，其中 $P(C_{(1)}|x)\geqslant P(C_{(2)}|x)\geqslant\cdots\geqslant P(C_{(N)}|x)$。

由于测试样本的最大输出概率 $P(C_{(1)}|x)$ 是一个随机变量(定义其为 Y)，可以针对已知应用及未知应用分别求其期望 μ_1、μ_2 及方差 σ_1^2、σ_2^2。假定已知应用及未知应用的最大输出概率 $P(C_{(1)}|x)$ 分别服从期望为 μ_1、μ_2，方差为 σ_1^2、σ_2^2 的高斯分布，其概率分布示意图如图 1 所示。定义参数 α、β 分别

为 $\mu_2+\sigma_2 z_\tau$、$\mu_1-\sigma_1 z_\tau$（其中 $0<\beta<\alpha<1$，τ 值一般很小，z_τ 值查表可知）。则对样本 x 进行识别时，若 $P(C_{(1)}|x)\geqslant\alpha$，根据 $P\{\frac{Y-\mu_2}{\sigma_2}<z_\tau\}=1-\tau$ 可知未知应用的 FPR 为 τ，表明 x 属于 $C_{(1)}$ 类的概率足够大，直接判为 $C_{(1)}$ 类。若 $P(C_{(1)}|x)\leqslant\beta$，根据 $P\{\frac{Y-\mu_1}{\sigma_1}<-z_\tau\}=\tau$ 可知已知应用的 FNR 为 τ，表明 x 属于未知应用的概率足够大，可直接判为未知应用。概率判决的流程图如图 2 所示。

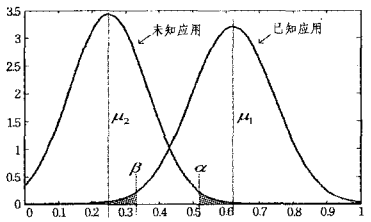

图 1　$P(C_{(1)}|x)$的概率密度函数

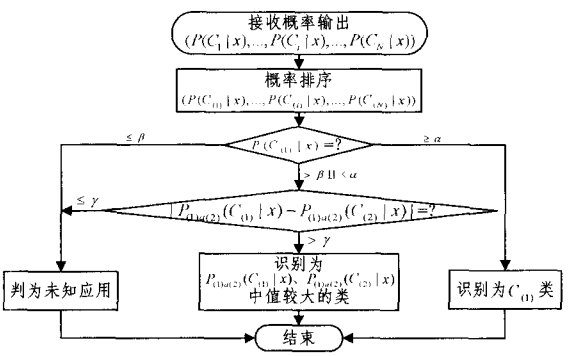

图 2　概率判决流程图

当 $\beta<P(C_{(1)}|x)<\alpha$ 时，则样本 x 在 SVM 的最优分类面附近，易产生误判，其可能属于概率较大的 $C_{(1)}$、$C_{(2)}$ 类或为未知应用，需要进行二次判决。针对以上满足 $\beta<P(C_{(1)}|x)<\alpha$ 的样本，计算 $|P_{(1)a(2)}(C_{(1)}|x)-P_{(1)a(2)}(C_{(2)}|x)|$ 的值，其为高斯随机变量。分别对已知应用及未知应用的该类样本求其期望及方差，得到高斯概率密度函数，如图 3 所示。两种应用的曲线交于 γ，而进行二次判决时，为了使误判率最小，如果待识别样本 x 的 $|P_{(1)a(2)}(C_{(1)}|x)-P_{(1)a(2)}(C_{(2)}|x)|$ 值小于等于 γ，表明属于前两类的概率相当，将其判为未知应用；如果大于 γ，则将其识别为 $P_{(1)a(2)}(C_{(1)}|x)$、$P_{(1)a(2)}(C_{(2)}|x)$中值较大的类。该步骤重新比较了属于 $C_{(1)}$、$C_{(2)}$ 类概率的大小，可以有效降低一次判决的误判率，此外，$P_{(1)a(2)}(C_{(1)}|x)$及 $P_{(1)a(2)}(C_{(1)}|x)$的值可以直接调用，几乎不增加计算的复杂度。

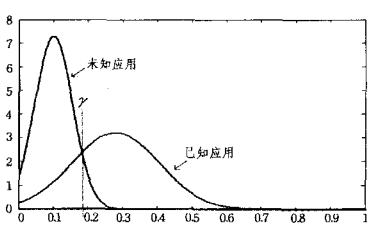

图 3　$|P_{(1)a(2)}(C_{(1)}|x)-P_{(1)a(2)}(C_{(2)}|x)|$的概率密度函数

4　实验设置

4.1　实验数据集

文献[12]表明，Abacus 需从 P2P 流媒体应用的全部流量提取特征，网络位置(Network Site，NS)的变化对识别结果影响也不大，鲁棒性较好，且本文不针对接入技术(Access Technology，AT)及频道流行度(Channel Popularity，CP)的影响进行研究。因此，可以直接在校园网出口处进行有监督的流量采集，并基于此数据集对 Abacus 与 P-Abacus 的识别效果进行对比。数据集的流量由国内外 7 种 P2P 流媒体应用(包含文献[12]中的 4 种)在收看 CCTV 1 时产生，其采集时间为 2011.11.06。关于数据集的具体描述如表 1 所列。

表 1　校园网数据集的采集信息

	signatures	packets	bytes	endpoints
PPLive	29k	16.9M	7.6G	12
PPStream	32k	21.2M	9.2G	15
QQLive	26k	17.2M	7.5G	13
UUSee	28k	16.4M	7.1G	10
SopCast	30k	14.7M	8.5G	11
TVants	27k	15.6M	8.1G	15
Joost	22k	6.2M	6.4G	13

4.2　分析工具与平台

本文利用 Java 平台进行数据集的处理，并借用数据挖掘工具 Weka-3.5.6[15] 进行研究。Weka 是由新西兰怀卡托大学 Witten 教授等人开发的开源工作平台。本文对其中的 LibSVM 算法进行修改，实现了 Abacus 与 P-Abacus 的功能。实验的仿真环境为一台普通 PC 机，其 CPU 为 Intel Core2 2.4GHz，内存为 DDR-667 2GB，运行 Windows XP 操作系统。

4.3　评价指标

本文在测试时，分别对已知应用及未知应用进行识别，得到 Abacus 与 P-Abacus 的识别效果对比。针对 signature 的识别效果，已知应用及未知应用的评价指标分别如表 2、表 3 所列。

表 2　已知应用的评价指标

编号	缩写	简单描述
1	TPR	应用 A 被识别为 A 的比率
2	Mis	应用 A 被识别为非 A 应用的比率
3	Unk	应用 A 被判为未知的比率

表 3　未知应用的评价指标

编号	缩写	简单描述
1	TNR	未知应用的正确识别率
2	FPR	未知应用被识别为已知的比率

5　实验及结果分析

本节对 Abacus 与 P-Abacus 的识别效果进行对比研究。为了实现识别功能，实验选用数据集中 PPLive、PPStream、QQLive、UUSee、SopCast 这 5 种应用用于训练，所有的 7 种应用用于测试。由于不针对参数灵敏度等进行研究，文章借鉴文献[12]，选用了径向基核函数(Radical Basis Function，RBF)，令训练集及测试集中各应用的样本数为 4000，P、B 分别为 8、14，ΔT 为 5s。

5.1　参数确定

1)Abacus 拒绝门限的确定

本文与文献[12]相比，训练集及测试集中应用的种类与规模发生了改变。为了使 Abacus 达到最佳识别效果，需要重新选择合适的拒绝门限 R。针对 5 类训练应用，在数据集中分别随机选取 8000 个样本，其中 4000 个用于训练，4000 个

用于已知应用的测试。又针对 TVants 和 Joost 应用，分别在数据集中随机选取 4000 个样本用于未知应用的测试。令 R 的值从 0.1 递增至 0.9 分别训练，则测试时得到已知应用的 TPR 及未知应用的 FPR。重复该训练及测试过程 10 次，求得不同门限 R 对应的 TPR 及 FPR 的均值，结果如图 4 所示。

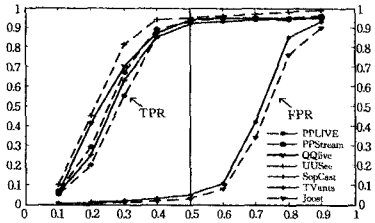

图 4　不同 R 值对应 TPR 及 FPR 的均值

从图 4 可以看出：当 R 的取值为 0.6 时，已知应用具有较高且较平稳的 TPR，但此时 TVants 的 FPR 达到 0.11；而当 R 的取值为 0.4 时，未知应用的 FPR 接近于 0，但 PPLive 的 TPR 只有 0.85。因此，综合考虑，选取 R 为 0.5，此时 TPR 最小为 0.92，FPR 最大为 0.05。

2）P-Abacus 参数的确定

为了使 P-Abacus 达到最佳识别效果，需要确定参数 α、β、γ 的值。实验中 P-Abacus 采用 Abacus 同样的方法来选取训练集及测试集，同时选取 τ 值。在模型的训练过程中，根据不同的 τ 值计算出 α、β、γ，测试时分别得到所有已知应用的平均 TPR、Mis 及未知应用的 TNR。重复该训练及测试过程 10 次，求得不同 τ 值所对应的 TPR、Mis 及 TNR 的均值（τ 选取了概率统计中常用的 5 个值），结果如图 5 所示。

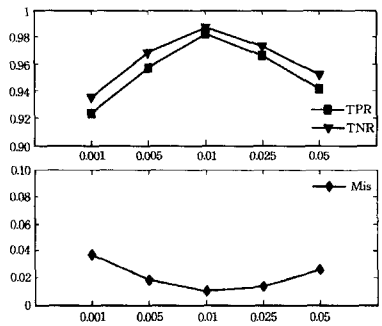

图 5　不同 τ 对应 TPR、Mis 及 FPR 的均值

从图 5 可以看出，选取 τ 为 0.01 时，P-Abacus 达到最佳识别效果，此时计算出 α、β、γ 的十次均值分别为 0.52、0.33、0.18。

5.2　Abacus 与 P-Abacus 的识别效果对比

实验分析从两个方面对比分析 Abacus 与 P-Abacus 的识别性能，即识别准确率和时间开销。

1）识别准确率

令参数 R 为 0.5，α、β、γ 分别为 0.52、0.33、0.18。实验选用相同规模的共同训练集及测试集，重新训练 Abacus 及 P-Abacus 的识别模型，测试时仍旧使用 10 折交叉验证的测试方式。其针对已知应用的评价指标如表 4 所列。

表 4　已知应用的评价指标

编号	缩写	简单描述
1	TP	应用 A 被识别为 A 的比率
2	Mis	应用 A 被识别为非 A 应用的比率
3	Unk	应用 A 被判为未知的比率

针对未知应用的评价指标如表 5 所列。

表 5　未知应用的评价指标

编号	缩写	简单描述
1	TN	未知应用的正确识别率
2	FP	未知应用被识别为已知的比率

采用两种识别法，重复进行识别模型的测试实验 10 次，对测试结果求均值，计算求得已知应用的识别结果如表 6 所列。

表 6　两种识别法对已知应用的识别准确率对比

	Method	%TP	%Mis	%Unk
PPLive	Abacus	92.36	4.60	3.04
	P-Abacus	97.18	1.65	1.17
PPStream	Abacus	93.45	3.76	2.79
	P-Abacus	97.93	1.12	0.95
QQLive	Abacus	94.22	3.23	2.55
	P-Abacus	98.42	0.96	0.62
UUSee	Abacus	95.01	2.87	2.12
	P-Abacus	99.84	0.16	0.00
SopCast	Abacus	92.78	4.46	2.76
	P-Abacus	97.86	1.09	1.05

未知应用的识别结果如表 7 所列。

表 7　两种识别法对未知应用的识别准确率对比

	Method	%TNR	%FPR
TVants	Abacus	95.10	4.90
	P-Abacus	98.56	1.44
Joost	Abacus	96.37	3.61
	P-Abacus	98.88	1.12

从表 6 中可以看到，针对已知应用的识别，P-Abacus 的 TP 能够达到 98%左右，而 Abacus 的 TP 只有 93%左右；P-Abacus 的 Mis 为 1%左右，Abacus 的 Mis 为 3%以上；P-Abacus的 Unk 为 1%左右，而 Abacus 的 Unk 则不到 3%。因此，由以上测试结果可以看出，P-Abacus 比 Abacus 具有更高的识别率以及更低的误判率，这表明 SVM 的概率判决比硬判决更加精细，并且 P-Abacus 中的二次判决能够对一次识别结果仔细判断，这是其误判率降低的主要原因。而从表 7 中可以看出，针对未知应用的识别，其识别率从 95%以上提高至98.5%，表明 P-Abacus 方法中使用的概率判决参数比 Abacus 中使用单一的拒绝门限 R 要更加合理。因此，综合以上分析，本文提出的 P-Abacus 方法在识别准确率方面比 Abacus 要好。

2）时间开销

由本文 SVM 的机理可知，SVM 识别法的结果由少数支持向量决定，同时文献[12]也表明识别法的主要时间开销取决于 SVM 中使用支持向量的个数，与其成正比。显然，P-Abacus 与 Abacus 方法，在采用共同的训练集时会使用相同的支持向量，因此在时间开销上，两种方法相差不大。尽管 P-Abacus 在进行二次判决时需重新比较属于前两类的概率，但是该计算过程可直接调用之前 SVM 多分类的结果，因此其时间开销增加有限。表 8 列举了两种识别法在每种应用实例数为 4000 时，模型的训练时间以及测试时间。

表 8　两种识别法的时间开销对比

	TrainingTime(s)	TestTime(s)
Abacus	581.6	368.7
P-Abacus	602.6	371.3

从表 8 可以看出，基于相同的训练集及测试集，P-Abacus 比 Abacus 的训练时间增加 3.6%，其测试时间只增加0.7%。

综合识别准确率以及时间开销两方面来考虑，P-Abacus 与 Abacus 相比，其识别精度从 93%增至 98%，尽管训练时间增加了 3.6%，但是识别时间却只增加 0.7%，因此从识别方法的实用性来考虑，P-Abacus 比 Abacus 更优。

结束语 本文针对 P2P 流媒体流量的精细化识别法进行了研究，对 Abacus 进行了改进，提出了一种基于 SVM 概率输出的识别法 P-Abacus。该方法能够将待识别实例属于已知应用的可能性大小反映在概率输出上，并结合概率判决对输出结果进行判断。由于 SVM 的概率输出包含更多的信息，同时，其概率判决中的二次判断也对最优超平面附近的实例进行了细致的识别，因此 P-Abacus 具有更好的识别效果。通过对 7 种较流行的 P2P 流媒体应用进行的实验测试表明，P-Abacus 比 Abacus 具有更高的识别率和更低的误判率，并且时间开销增加有限。

本文基于 P2P 流媒体应用的全部流量提取特征，因此 P-Abacus 方法同 KISS、PSD、Abacus 一样，只能部署在网络出口处，针对骨干链路上 P2P 流媒体流量的识别将是下一步的研究重点。

参 考 文 献

[1] 张艺濒，张志斌，赵咏，等. TCP 与 UDP 网络流量对比分析研究[J]. 计算机应用研究，2010，26(6)：2192-2197

[2] Perenyi M, Dang T D, Gefferth A, et al. Identification and Analysis of Peer-to-Peer Traffic[J]. Journal of Communications, 2006, 1(7): 36-46

[3] 雷蕾，沈富可. 基于连接特征的 P2P 流媒体应用的识别[J]. 计算机应用，2007，12：41-43

[4] Moore A, Papagiannaki K. Toward the accurate identification of network applications[C]// Proceedings of the Passive and Active Measurements Workshop. 2005: 41-54

[5] Nguyen T, Armitage G. A survey of techniques for internet traffic classification using machine learning[J]. IEEE Communications Surveys, 2008, 10(4): 56-76

[6] Liu Chao-bin, Yang Yue-xiang, Tang Chuan. A Classification Method of Unstructured P2P Multicast Video Streaming Based on SVM[C]// 2009 IEEE International Conference on Multimedia Information Networking and Security(MINES2009). 2009: 11

[7] Silverston T, Fourmaux O. P2P IPTV measurement: a comparison Study[M]. Preprint, 2006

[8] Alessandria E, Gallo M, Leonardi E, et al. P2P-TV Systems under adverse network conditions: a measurement study[C]// IEEE Infocom. Riode Janeiro, 2009

[9] Finamore A, Mellia M, Meo M, et al. KISS: Stochastic Packet Inspection[C]// Proceedings of the First International Workshop on Traffic Monitoring and Analysis (TMA). Aachen, Germany, 2009

[10] Li Jin, Zhang Xin, Zuo Xiao-liang, et al. Using Packet Size Distribution To Identify P2P-TV Traffic[C]// 2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery(PSD). Oct. 2010: 150-155

[11] Valenti S, Rossi D, Moe M, et al. Accurate, Fine-Grained Classification of P2P-TV Applications by Simply Counting Packets [M]. Papadopouli M, Owezarski P, Pras A, eds., TMA 2009. LNCS 5537, Berlin Heidelberg: Springer-Verlag, 2009: 84-92

[12] Bermolen P, Mellia M, Meo M, et al. Abacus: Accurate behavioral classification of P2P-TV traffic[J]. Computer Networks, 2011, 55(6): 1394-1411

[13] 李国正，王猛，曾华军. 支持向量机导论[M]. 北京：电子工业出版社，2004

[14] 吴朝晖，杨莹春. 说话人识别模型与方法[M]. 北京：清华大学出版社，2009

[15] Weka[OL]. http://www.cs.waikato.ac.nz/ma/weka

(上接第 30 页)

能力数据也有着迥异。本文的分析是基于一个平台整体的用户统计分析，如果将用户根据其工作归类后再分析，应该能得到一些更为准确的分析结论。另外对于单个威客平台 Elance 的数据分析有局限性，将全球另外两家平台数据收集起来加以横向比较，也是未来的工作方向之一。

参 考 文 献

[1] 王伟军，孙晶. Web2.0 的研究与应用[J]. 情报科学，2007，25(12)：1907-1931

[2] 刘锋. 威客(witkey)的商业模式分析[D]. 北京：中国科学院研究生院，2006

[3] AbleSky 国内威客网站调研报告[OL]. http://www.ablesky.com/viewCourseDetail.do? courseId=8223, 2011-11-25

[4] 2010 中国威客行业白皮书[OL]. http://www.iresearch.com.cn/report/, 2011-11-25

[5] Elance[OL]. https://www.Elance.com/, 2012-01-12

[6] oDesk[OL]. https://www.odesk.com/, 2012-01-12

[7] Freelancer[OL]. http://www.freelancer.com/, 2012-01-12

[8] 樊丽杰，王素贞，刘卫. 基于人类信任机制的移动电子商务信任评估方法[J]. 计算机科学，2012，39(1)：190-192

[9] Paul Resnick AT, Labs-Research T, Hill M, et al. Recommender Systems[J]. Communications of the ACM, 1997, 40(3): 56-58

[10] 李杨，韦伟，刘永忠. 一种基于 AHP 的信息安全威胁评估模型研究[J]. 计算机科学，2012，39(1)：61-64

[11] 吴明晖，曹梦筠，王海涛. 基于 FAHP 的综合评价系统及在软件外包中的应用[J]. 计算机系统应用，2008(11)：91-94

[12] 纪淑娴，胡培，程飞. 在线信誉管理中信用度计算模型研究[J]. 预测，27(4)：59-64

[13] Jin Song-he, Song Bao-wei, He Lei. Recommendation of Online Tasks Based on Witkey Mode Website[Z]. IFITA. 2009: 268-270

[14] 刘瑞芳，谢长生，谭志虎. 基于 CMM 的软件开发过程研究[J]. 计算机应用研究，2004(7)：133-134

[15] 尤薇佳，刘鲁，杨俊杰，等. 基于交易记录的欺诈识别[A]// 第二届网商及电子商务生态学术研讨会论文集[C]. 2009：178-182

[16] 朴春慧，韩旭芳，陈挥青. 威客电子商务作弊评价模型及算法研究[J]. 科学的实践与认知，2010，40(11)：124-130

[17] 朴春慧，韩旭芳，杨春燕. 基于 Web2.0 的威客电子商务作弊处理机制研究[J]. 情报杂志，2009(10)：124-128

[18] Openfire 开源项目[OL]. http://www.igniterealtime.org/projects/openfire/, 2012-01-08

[19] Arale 开源项目[OL]. http://flavio.tordini.org/arale, 2012-01-09

[20] 猪八戒威客网站[OL]. http://www.zhubajie.com/, 2012-01-15