基于UDP统计指印混合模型的VoIP流量识别方法

丁要军[1,2] 蔡皖东[1] 姚 烨[1]
(西北工业大学计算机学院 西安 710129)[1] (咸阳师范学院信息工程学院 咸阳 712000)[2]

摘 要 针对VoIP加密负载流量识别的难题,提出一种基于UDP统计指印混合模型的VoIP流量识别方法,以提高VoIP流量的识别精度和分类稳定性。该模型改进了统计指印模型中基于单一的网络流相异度来判定流量类别的方法,将UDP流的统计特征与网络流的统计指印相异度结合以共同训练一个支持向量机分类模型,把基于分类阈值点的分类转换到基于多维特征的高维空间中的分类面的分类,综合运用包层次和流层次统计特征,降低了因网络不稳定造成的统计特征偏差对分类模型精确度的影响。实验结果表明,该模型对VoIP流量的分类精确度达到97%以上,与统计指印模型和支持向量机模型相比分类稳定性更好。

关键词 统计指印,VoIP,流量分类,支持向量机,互联网
中图法分类号 TP393 **文献标识码** A

VoIP Traffic Identification Based on UDP Statistical Fingerprinting Mixture Models

DING Yao-jun[1,2] CAI Wan-dong[1] YAO Ye[1]
(Department of Computer, Northwestern Polytechnical University, Xi'an 710129, China)[1]
(Department of Information Engineering, Xianyang Normal University, Xianyang 712000, China)[2]

Abstract Because it is difficult to identify encrypted VoIP traffic, we proposed a UDP statistical fingerprinting mixture models to enhance the accuracy and stability of VoIP traffic identification. We used the statistical features of UDP flow along with the anomaly score of traffic flow in which the statistical fingerprinting model is used to identify a traffic flow to train a Support Vector Machine(SVM) classification model, and used a hyperplane of high-dimensional space instead of a threshold point to classify the traffic. Because we use both the packet level features and flow level features in our mixture models, the impact of the deviation of traffic features which is caused by the instability of network will be decreased. The results of our experiment show that the precision of VoIP traffic is over 97% in our model, and our model is more stable compare with the statistical fingerprinting model and Support Vector Machine(SVM).

Keywords Statistical fingerpinting, Voice over internet protocol, Traffic classification, Support vector machine, Internet

1 引言

VoIP(Voice over Internet Protocol)技术是一种以IP电话为主并推出相应增值业务的技术。VoIP采用的是计算机通信的分组化、数字化传输技术,先对语音数据进行压缩编码处理,然后把数据按IP等相关协议打包,数据包通过IP网络传输到接收地之后再重新串起来,经过解码解压恢复成原来的语音信号。与传统的语音业务相比,VoIP能在同样带宽条件下使通话数量成倍增加,可以实现低成本的语音传送、传真等传统电信业务。目前国内应用较为广泛的VoIP技术主要有Skype和QQ语音等,VoIP电话作为一种新的业务,有着自身的特点以及传统业务所无法比拟的长处,并已成为Internet应用领域的一个热点。因此,对VoIP业务流的分类和识别就显得更加重要。

Baset等人[1]对Skype协议进行了详细分析,发现Skype协议使用动态端口而且负载部分完全加密,传统的基于端口和DPI(Deep Packet Inspection)的协议识别方法已基本失效。近几年,基于机器学习的协议识别技术发展迅速,取得了很多成果。Moore等人[2]提出了基于大量传输层特征的朴素贝叶斯模型的流量分类方法,该方法提取了传输层的248个统计特征,使用实际流量数据对模型进行训练,对常用协议有很好的分类效果。徐鹏等人[3]提出了基于SVM的流量分类方法,该方法能有效降低冗余属性的干扰,而且不依赖于贝叶斯方法中的先验概率,有很好的分类准确率和稳定性,但目前这方面的研究并未考虑VoIP流量的识别。Crotti等人[4]使用统计指印方法实现对HTTP、POP3、SMTP等常用协议的识别,该方法基于TCP流的前4个包的统计特征建立相应的统计指印,并通过计算TCP流的相异度来判定流的协议类别,但文献中并未研究UDP指印的构造和应用。Bonfiglio等人[5]使用卡方检验和朴素贝叶斯的方法实现对Skype协议流量的识别,这也是目前在Skype流量识别方面最有效的方法,但该方法是根据Skype协议的编码方式和包头特征提出的,

到稿日期:2012-11-07 返修日期:2013-03-12 本文受国家高技术研究发展计划(863)项目(2009AA01Z424),陕西省教育厅科研计划项目(12JK0933),咸阳师范学院专项科研基金项目(12XSYK068,10XSYK308,07XSYK280)资助。

丁要军(1980-),男,博士生,讲师,主要研究方向为网络与信息安全、流量分类,E-mail:dingyj@mail.nwpu.edu.cn;蔡皖东(1955-),男,教授,博士生导师,主要研究方向为网络安全与信息对抗。

只能识别特定版本的 Skype 协议流量，无法识别 QQ 语音、MSN 语音等其它 VoIP 流量，有一定的局限性。

从文献[1]中可知，Skype 使用 TCP 协议传输信令信息，而使用 UDP 和 TCP 共同传输媒体数据信息，因为传输媒体数据的 UDP 数据包特征明显，所以本文从 VoIP 流量中的 UDP 包特征和流特征入手建立 VoIP 流量分类模型。

2 相关概念

2.1 基本概念

网络流：在一段时间间隔内，具有相同源 IP、目的 IP、源端口、目的端口、传输层协议的网络报文序列称为网络流，这 5 个属性也称为五元组。

流统计特征：以流为单位计算出来的统计特征包括流的报文大小的期望和方差、报文到达时间间隔的期望和方差等。

2.2 统计指印

统计指印(Statistical FingerPrinting)是一个用矩阵形式表示的图像，矩阵的行和列分别代表数据包的两个特征。下面简单介绍建立统计指印的方法。

建立统计指印使用网络流的 3 个统计特征：流中包的到达顺序、流中包的大小 s 和流中数据包的到达时间间隔 Δt。用一个矩阵来表示网络流 $\vec{x}$，如式(1)所示：

$$\vec{x}=\begin{pmatrix} s_1 & \cdots & s_r \\ \Delta t_1 & \cdots & \Delta t_r \end{pmatrix} \tag{1}$$

式中，r 表示流 $\vec{x}$ 中包含 r 个数据包。现在假设选取 n 个流来构造第 i 个数据包的指印，则得到如下矩阵 F：

$$F=\begin{pmatrix} (s_1,\Delta t_1)^1 & (s_1,\Delta t_1)^2 & \cdots & (s_1,\Delta t_1)^n \\ (s_2,\Delta t_2)^1 & (s_2,\Delta t_2)^2 & \cdots & (s_2,\Delta t_2)^n \\ \cdots & \cdots & \cdots & \cdots \\ (s_r,\Delta t_r)^1 & (s_r,\Delta t_r)^2 & \cdots & (s_r,\Delta t_r)^n \end{pmatrix} \tag{2}$$

式中，一列代表一个网络流，共有 n 列，代表 n 个网络流，第 i 行代表每个流的第 i 个数据包，第 i 行的所有包用来构建流的第 i 个包的指印。

同一协议下的包大小和时间间隔存在一定的规律，所有的点会落在一个相对稳定的区域，在同一点上落的点较多时，这个点上的灰度值会比较高。由于存在网络丢包和拥塞等情况，统计指印通常存在一定量的噪声，使用核函数对指印进行高斯过滤可以有效地消除噪声。

2.3 支持向量机

支持向量机方法[6]是一种基于统计学习理论的机器学习方法，该方法将分类问题转化为在特定约束条件下的寻找最优超平面的二次寻优问题，有效提高了分类模型在小样本情况下的分类准确率和稳定性。下面简单介绍支持向量机实现二分类的原理。

给定一组独立同分布的样本点：

$$(x_1,y_1),\cdots,(x_n,y_n),x_i\in R^m,y_i\in\{-1,+1\} \tag{3}$$

式中，x_i 是指样本向量，y_i 是指样本所属类别，正例用 $+1$ 表示，反例用 -1 表示。

SVM 的目标是在高维空间上寻求一个最优分类面：

$$w^{\mathrm{T}}x+b=0 \tag{4}$$

最优分类面不仅能将两类样本分开，而且能使两类样本到最优分类面的距离最大。考虑可能有一些样本不能被分类面正确分类，引入松弛变量 $\xi_1,\xi_2,\cdots,\xi_n$ 以及惩罚因子 C，将最优分类面的求解转化为有约束的二次规划问题：

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \tag{5}$$

满足：$y_i(w\cdot x_i+b)\geqslant 1-\xi_i$ (6)

$$\xi_i\geqslant 0,i=1,\cdots,n$$

式中，常数 $C>0$ 称为“惩罚因子”，它在分类器的复杂度和经验风险之间进行权衡。

为求解式(5)和式(6)中的二次规划问题，引入 Lagrange 算子 $\alpha_i,i=1,\cdots,l$，并定义：

$$w(\alpha)=\sum_{i=1}^{l}\alpha_i y_i x_i \tag{7}$$

将二次规划问题转化为对偶问题：

$$\max W(\alpha)=\sum_i\alpha_i-\frac{1}{2}w(\alpha)\cdot w(\alpha) \tag{8}$$

满足

$$\alpha_i\geqslant 0,\sum_i\alpha_i y_i=0 \tag{9}$$

最终求得分类判别函数为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i y_i(x_i\cdot x)+b) \tag{10}$$

在最优分类面中采用适当的核函数就可以实现某一非线性变换后的线性分类，而计算复杂度却没有增加。引入核函数后判别函数转换为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b) \tag{11}$$

式中，$K(x_i,x)$ 为核函数。

3 构造 UDP 统计指印

3.1 原始指印

与 TCP 协议不同，UDP 协议没有 3 次握手，也没有严格的流划分机制。Li Wei 等人[7]提出 UDP 包的通信间隔时间大于 60s 的划分为一个流，这是目前效果较好的 UDP 流划分方法。

VoIP 协议主要使用 UDP 来传输语音等多媒体信息，一次语音通话构成一个 UDP 流，UDP 流通常都较长。因为单位时间内采集的 UDP 流个数远远小于 TCP 流的个数，指印算法的精度对训练集数量的依赖性比较大，如果训练集的数量不充分，那么算法的识别精度将会大大降低。因此，提出一种改进的统计指印构造方法。在 UDP 流个数有限的情况下，使用所有 UDP 流中的所有 UDP 包来建立一个指印，而不是多个指印。这样一方面降低了指印建立的算法复杂度，另一方面使指印的落点更加密集，增强了指印的抗噪声干扰能力。我们使用纯净的 VoIP 流量数据建立了 UDP 指印，包的落点比较集中，如图 1 所示。

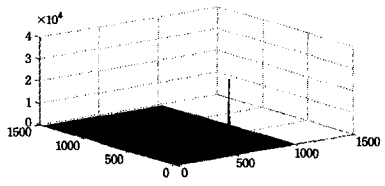


图 1 VoIP 流量的 UDP 指印三维图

对于坐标中的点，根据数据包的特征值来确定其所处的位置，根据网络数据包大小的可能取值，s 轴的取值范围从 40 到 1500 字节，取一个字节为单位长度，而 Δt 的取值范围从 10^{-7} 到 10^3 s 之间，取 0.01s 为单位长度。若有 m 个数据包的 $(s_i,\Delta t_i)$ 落在了某一坐标处，则这个坐标点的权值就定义为

m。如此，就构造出了 UDP 流中所有数据包的指印特征模型，这个特征模型可以用一个 1461×1001 的矩阵表示。

3.2 基于核函数的指印去噪

由于网络环境的复杂性，构造出来的 UDP 指印图像包含一定量的噪声，为进一步提高算法的鲁棒性，将对构造的指印图像进行高斯过滤。高斯核函数如式(12)所示：

$$K(\| X-X_C \|)=\exp\{-\| X-X_C \|^2/2*\sigma^2\} \tag{12}$$

式中，X_C 为核函数中心，σ 为函数的宽度参数，控制了函数的径向作用范围，相应参数的取值将在实验部分讨论。使用核函数对每一坐标点进行平滑处理，使它的取值与周围点的取值相关，并进行归一化，得到最终的协议指印 M。

3.3 计算 UDP 流与指印的相异度

给定一个 UDP 流 F，我们用 $\vec{x}=(x_1,\cdots,x_r)$ 来表示这个数据流，结合朴素贝叶斯分类的方法定义相异度变量 S：

$$S(\vec{x}|\omega_t)=|\log_{10}\prod_{i=1}^{r}p(x_i|\omega_t)/r| \tag{13}$$

式中，$p(x_i|\omega_t)$ 表示第 i 个数据包属于类别 ω_t 的条件概率，r 表示从待检测数据流中挑选的包的个数，r 的取值大小对最终的识别准确率有一定的影响，将在实验部分讨论。$p(x_i|\omega_t)$ 的取值通过协议掩模 M 来计算：

$$p(x_i|\omega_t)=M(s_i,\Delta t_i) \tag{14}$$

对于 $p(x_i|\omega_t)$ 为空值的情况，我们在计算时用一个很小的数 10^{-300} 来代替它。

性质 1 相异度 S 的取值范围为[0,1]，S 越趋近于 0 时，则 UDP 流 F 属于类别 ω_t 的概率就越大；S 越趋近于 1 时，则 UDP 流 F 属于类别 ω_t 的概率就越小。

证明：因为 $y=|\log_{10}x|$ 在区间 $0<x<1$ 上是递减函数，所以 S 关于 $\prod_{i=1}^{r}p(x_i|\omega_t)$ 在区间[0,1]上为递减函数。

由朴素贝叶斯的条件独立假设可知：

$$p(x|w_t)=\prod_{i=1}^{r}p(x_i|\omega_t) \tag{15}$$

由贝叶斯定理可知：

$$p(w_t|x)\propto p(x|w_t) \tag{16}$$

则性质 1 成立。

3.4 阈值选取

由性质 1 可知，判定一条 UDP 流是否属于 VoIP 流量，最好的方法是分别建立 VoIP 和非 VoIP 流量的统计指印，分别计算与两种指印的相异度并判定 UDP 流的类别。建立所有的非 VoIP 协议的指印比较困难，文献[4]中只建立了正例协议的指印，并设定了一个相异度的阈值 T_{acc} 来判定流量的协议类别：

$$\hat{\omega}(\vec{x})=\begin{cases}\omega_t, & S<T_{acc}\\ \omega_r, & \text{其它}\end{cases} \tag{17}$$

如式(17)所示，若 S 小于阈值，则数据流属于类别为 ω_t 的协议流，否则属于其它协议的数据流。

文献[4]中的阈值的取值方法是从训练集中选取属于类别 ω_t 的一个子集，阈值 T_{acc} 的选取使得在式(17)的判定方法下子集中 99%的样本属于类别 ω_t，1%的样本属于类别 ω_r。

4 UDP 指印混合模型

4.1 统计指印算法的不足

从统计指印算法的分析可以看出，算法主要存在两方面的不足：

(1)基于阈值的判定方法的可靠性和精确度很大程度上依赖于指印的建立和阈值的选取，而指印的建立完全依赖于训练数据集。文献[4]中的实验表明，当训练集的数量不够充足或者阈值的选取不够精确时，算法的准确性无法保障。

(2)网络流量的特征并不稳定，容易受到网络拥塞等外在因素影响，包到达时间间隔等特征也会有一定的偏差，单纯地依靠包层次特征会造成特征描述的偏斜。

4.2 指印混合模型

为弥补文献[4]中算法的不足，提出一种 UDP 指印的混合模型，将通过 UDP 指印计算出来的流的相异度与 UDP 流的其它统计特征结合，共同训练一个 SVM 分类模型并使用训练好的 SVM 分类模型来判定流的类别。

算法首先使用 VoIP 流量的包层次特征训练一个 UDP 统计指印模型，然后通过指印模型将训练集中网络流的包特征转换为流层次上的相异度。最后，将相异度与约减后的 UDP 流特征结合，共同训练一个 SVM 分类模型，实现基于多层次特征的流量分类。算法的实现过程如图 2 所示。

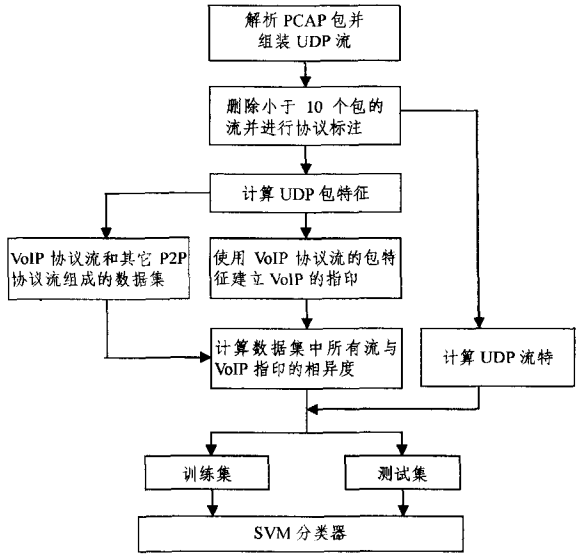


图 2 UDP 指印混合模型分类示意图

与文献[4]中的方法相比，UDP 指印混合模型的优势主要体现在两个方面：

(1)使用 SVM 分类模型取代了基于阈值的判定方法，分类过程也从单独的基于分类阈值的点转换到了基于多维特征的高维空间中的分类面。

(2)综合运用了包层次统计特征和流层次统计特征，降低了因网络不稳定造成的单个包特征或流特征的偏差对分类模型精确度的影响。

5 实验结果与分析

实验平台使用一台个人计算机，CPU 为 Intel Pentium-4 2.80G Hz，内存为 1G Bytes，操作系统为 Windows XP。使用 matlab7.0[8] 实现 UDP 指印的构造和去噪，采用 T. Joachims 的 SVM-Light[9] 实现 SVM 分类，核函数选用 RBF(Radial Basis Function)。

5.1 数据预处理

5.1.1 数据采集

在实验室局域网的网关出口使用流量捕获卡采集流量，

流量捕获过程中在不同的计算机上运行相应的程序。为获取纯净的 VoIP 数据，我们在固定的时间段、固定的计算机上只运行 Skype 软件和 QQ 语音。最终采集了一周时间的流量，共计 62GB。

5.1.2 特征提取

采集的网络流量数据以 PCAP 文件的形式存放，使用 C# 语言编写报文解析程序，首先根据五元组{源 IP，源端口，目的 IP，目的端口，传输层协议}完成 UDP 流的组装，然后挑选流长度大于 10 个包的流并以流为单位分别提取流的统计特征和流中各个包的特征，具体特征如表 1 和表 2 所列，其中最后一个特征是流的协议类别。

表 1 提取的包特征

序号	简称	描述
1	Pkt1_len	The first packet's length of flow
2	Pkt2_len	The second packet's length of flow
3	Pkt3_len	The third packet's length of flow
4	Pkt4_len	The fourth packet's length of flow
5	Pkt5_len	The fifth packet's length of flow
6	Pkt6_len	The sixth packet's length of flow
7	Pkt7_len	The seventh packet's length of flow
8	Pkt8_len	The eighth packet's length of flow
9	Pkt9_len	The ninth packet's length of flow
10	Pkt10_len	The tenth packet's length of flow
11	Pkt1_inter	The first packet's inter-arrival time
12	Pkt2_inter	The second packet's inter-arrival time
13	Pkt3_inter	The third packet's inter-arrival time
14	Pkt4_inter	The fourth packet's inter-arrival time
15	Pkt5_inter	The fifth packet's inter-arrival time
16	Pkt6_inter	The sixth packet's inter-arrival time
17	Pkt7_inter	The seventh packet's inter-arrival time
18	Pkt8_inter	The eighth packet's inter-arrival time
19	Pkt9_inter	The ninth packet's inter-arrival time
20	Pkt10_inter	The tenth packet's inter-arrival time
21	Class	Type of flow

表 2 提取的流统计特征

序号	简称	描述
1	Num_pkts	Number of packets seen on both directions
2	Min_pbyte_clnt	Minimum payload bytes seen(client to server)
3	Min_pbyte_serv	Minimum payload bytes seen(server to client)
4	Max_pbyte_clnt	Maximum payload bytes seen(client to server)
5	Max_pbyte_serv	Maximum payload bytes seen(server to client)
6	Ini_pbyte_clnt	Payload bytes sent from client to server before the first packet coming back
7	Max_csct_pkts_clnt	Maximum number of consecutive packets(client to server)
8	Serv_port	Server Port
9	Clnt_port	Client Port
10	Class	Type of flow

包特征的选取主要根据建立 UDP 指印的需要，选取流的前 10 个包的包大小和到达时间间隔。流特征的选取主要参照文献[7]中给出的 UDP 流统计特征，使用 FCBF(Fast Correlation-Based Filter)[10]算法筛选得到。

5.1.3 协议标注

VoIP 流量大多采用负载加密技术，因此标注起来比较困难。在流量采集过程中，在固定时间段、固定 IP 地址的计算机上只运行 VoIP 类软件，包括 Skype 语音和 QQ 语音，然后根据时间和 IP 地址从采集的流量中挑选出 VoIP 类流量，其它协议流量的标注使用 *L7-fileter*[11]。*L7-fileter* 是基于应用层特征字符串来标注协议，对于负载部分没有完全加密的协议识别效率很高。最终标注的协议类别和相应的流个数如表 3 所列。

表 3 实验数据的协议类别及数量

协议	流条数	包个数
Skype	2846	361643
QQ 语音	2554	283342
Thunder	2975	433792
PPLive	2564	242901

因为实验是基于 UDP 流的特征来识别 VoIP 流量，所以实验数据主要选取使用 UDP 传输数据的协议，如迅雷、PPLive 等。

5.1.4 生成数据集

实验选择 Skype 和 QQ 语音流量作为正例，以 Thunder 和 PPLive 协议作为反例，将数据流分为 VoIP 流量和非 VoIP 流量。

从标注好的流量中挑选 66%作为训练集，剩余的流量作为测试集，为观察算法在数据空间上的稳定性，我们将测试集随机分成 4 个子集：Set1，Set2，Set3，Set4。

对训练集中的包特征分布进行了统计，如图 3 和图 4 所示。

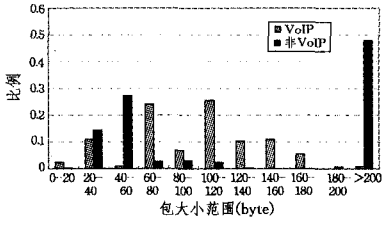


图 3 训练集中包大小分布图

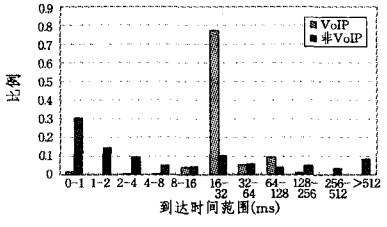
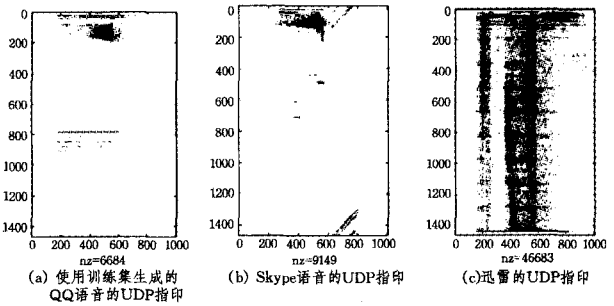


图 4 训练集中包到达时间间隔分布图

从图中可以明显看出，两种类别的流量在包大小和包到达时间间隔分布上差异比较明显，这也验证了使用 UDP 指印的可行性。

下面对比分别使用训练集中的 Skype 流量和迅雷流量构造的 UDP 指印，如图 5 所示。



(a) 使用训练集生成的 QQ语音的UDP指印　(b) Skype语音的UDP指印　(c)迅雷的UDP指印

图 5

从图中可以看出，QQ 语音和 Skype 语音的 UDP 指印相似度较高，而迅雷的 UDP 指印与前两种协议的指印差别明显。

5.2 算法性能比较

5.2.1 算法性能评估参数

使用 Overall Accuracy、Recall 和 Precision 3 个指标来评价算法的精确度，使用 Training Time 和 Testing Time 来评价算法的效率。

(1)Overall Accuracy：所有类别中被正确分类的样本数占所有样本总数的百分比。

(2)Precision：对某一类别 A，被正确分类为类别 A 的样本数占所有被分类为 A 的样本数百分比。

(3)Recall：对某一类别 A，被正确分类为类别 A 的样本数占类别 A 真实所包含样本数的百分比。

(4)Training Time：算法使用训练集完成分类模型训练所需要的时间。

(5)Testing Time：分类模型完成对测试集的测试所需要的时间。

5.2.2 混合模型在不同测试集上的表现

混合模型中选取 10 个包来计算相异度，SVM 分类器中选择径向基函数作为核函数分别在 4 个测试集上测试 UDP 指印混合模型的精确度并使用 Precision 和 Recall 加以评价，实验结果如表 4 所列。

表 4 UDP 指印混合模型的 VoIP 流量分类精确度

测试集	Precision(%)	Recall(%)
Set1	98.31	97.68
Set2	97.89	97.86
Set3	98.95	98.91
Set4	98.86	98.83

从表中可以看出，混合模型在不同测试集上的精确度都在 97%以上，在数据空间上的稳定性较好。

5.2.3 3 种模型的总体分类精度对比

为进一步评价 UDP 指印混合模型的性能，对 3 种模型在不同测试集上的总体分类精度进行对比。3 种模型分别为 UDP 指印混合模型、UDP 指印模型、SVM 模型，其中 UDP 指印模型需要确定分类阈值，根据文献[4]中的方法得到 T_{acc} 的值为 5.01，SVM 模型中选择径向基函数作为核函数，实验结果如图 6 所示。

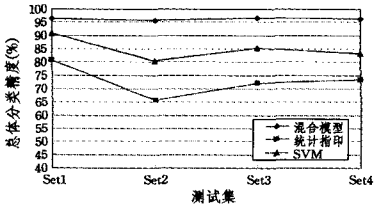


图 6 3 种模型的总体分类精度对比

从图中可以看出，混合模型在 4 个测试集上的总体分类精度都是最高的，并且精度波动较小，而其它两种模型的总体精度偏低，而且波动较大。主要是因为混合模型综合运用的包层次特征和流层次特征受某一个层次特征的干扰较小，稳定性更好。

5.2.4 模型的时间效率对比

实验中记录了 3 种模型的训练时间和 Set1 上的测试时间，如表 5 所列。

表 5 3 种模型的时间效率(CPU-seconds)

类别	统计指印	SVM	混合模型
Training Time	6.56	2.82	9.31
Testing Time	0.32	0.03	0.04

其中，统计指印模型的训练时间为 matlab 中生成 UDP 指印的时间和高斯过滤的时间，从表中可以看出混合模型的训练时间相对较长，包含了 UDP 指印的构造和 SVM 分类模型的训练，但测试时间与 SVM 模型相差不大。因为在测试集生成过程中已经计算了流的相异度，所以在分类测试过程中混合模型的计算复杂度与 SVM 相当。由于网络流量分类模型的训练是在离线条件下进行的，对训练时间的要求不高，因此混合模型在时间复杂度上是可行的。

结束语 近几年基于机器学习方法的流量分类成果较多[12]，但研究成果中大多使用 TCP 流的统计特征来训练分类模型，忽略了对 UDP 流统计特征的研究。以 VoIP 为代表的音频和视频应用协议同时使用 TCP 和 UDP 来传输数据，因此提出一种基于 UDP 流统计特征和包层次特征的混合分类模型，改进了统计指印模型，实验证明混合模型对 VoIP 流量具有很高的分类精度，与其它模型相比在数据空间上的稳定性更好。实验数据是在实验室局域网中采集的，VoIP 流量数据偏少，将在以后的工作中采集更多流量进行实验，以进一步优化分类模型。

参 考 文 献

[1] Salman A. Baset, Henning Schulzrinne. An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol[C]// Proceedings of the 2006 IEEE Infocom, IEEE'06. Barcelona, Spain, Apr. 2006

[2] Moore A W, Zuev D. Internet Traffic Classification using Bayesian Analysis Techniques[C]// Proceedings of the 2005 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems. New York, USA: ACM, 2005: 50-60

[3] 徐鹏，刘琼，林森. 基于支持向量机的 Internet 流量分类研究[J]. 计算机研究与发展，2009，46(3)：407-414

[4] Crotti M, Dusi M. Traffic Classification through Simple Statistical Fingerprinting[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(1): 5-16

[5] Bonfiglio D, Mellia M, Meo M. Revealing Skype Traffic: When Randomness Plays with You[C]// Proceedings of 2007 ACM SIGCOMM Computer Communication Review. New York, USA: ACM, 2007: 37-48

[6] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26(1)：32-42

[7] LI Wei, Canini M, Moore A W. Efficient Application Identification and the Temporal and Spatial Stability of Classification Schema[J]. Computer Networks, 2009, 53(6): 790-809

[8] Company of MathWorks. MATLAB [EB/OL]. http://www.mathworks.cn/products/matlab/, 2011-06-02

[9] Joachims T. SVM-Light[EB/OL]. http://svmlight.joachims.org/, 2011-06-02

[10] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlationbasedfilter solution[C]// Proceedings of the 20th International Conference on Machine Learning (ICML'03). 2003

[11] COMPANY of SOURCEFORGE. L7-filter[EB/OL]. http://l7-filter.sourceforge.net/. 2011-06-02

[12] 刘琼，刘珍，黄敏. 基于机器学习的 IP 流量分类研究[J]. 计算机科学，2010，37(12)：35-40