基于 NetFlow 记录的高速应用流量分类方法

陈亮[1,2]，龚俭[1,2]

(1. 东南大学 计算机科学与工程学院，江苏 南京 210096；2. 江苏省计算机网络技术重点实验室，江苏 南京 210096)

摘 要：针对目前应用流量分类算法效率不高的现状，提出一种以 NetFlow 统计的 IP 流记录信息作为输入的高速应用流量分类(FATC, fast application-level traffic classification)算法。该算法采用基于简单相关系数的测度选择算法衡量测度变量间的相关关系，删除对分类无用或相互冗余的测度，而后使用基于 Bayes 判别法的分类算法将网络流量分至误判损失最小的应用类别中。理论分析及实验表明，FATC 算法在具有超过 95%的分类准确率基础上，极大降低了当前应用流量分类方法在训练和分类过程的时空复杂度，满足实时准确分类当前 10Gbit/s 主干信道网络流量的需求。

关键词：计算机系统结构；流量分类；NetFlow；相关系数；特征选择；Bayes 判别法

中图分类号：TP393 **文献标识码：**B **文章编号：**1000-436X(2012)01-0145-08

Fast application-level traffic classification using NetFlow records

CHEN Liang[1,2], GONG Jian[1,2]

(1. College of Computer Science and Technology, Southeast University, Nanjing 210096, China;
2. Jiangsu Province Key Laboratory of Computer Networking Technology, Nanjing 210096, China)

Abstract: In order to improve the performance and reduce the resources usage of application-level traffic classification, a novel fast application-level traffic classification(FATC) algorithm using IP flow record from NetFlow as input was presented. FATC adopted metric selection algorithm based on correlation coefficient to measure the correlation among flow metric variables, and deleted the irrelevant or redundant metrics, then used Bayes discrimination to classify network traffic to the application category that of smallest misjudge loss. The theoretical analysis and experimental results show that, with more than 95% accuracy, the FATC algorithm greatl reduces the time and space complexity of current application-level traffic classification algorithms during the training and classification processes, and can work efficiently on 10Gbit/s backbone network in real time.

Key words: computer system architecture; traffic classification; NetFlow; correlation coefficient; feature selection; Bayes discrimination

1 引言

实时准确地识别 Internet 流量所使用的应用层协议是网络 QoS、网络流量和用户行为等监控的前提和基础，在网络性能管理、计费管理、流量工程和入侵检测等研究领域有着重要作用。然而由于包

收稿日期：2010-05-06；**修回日期：**2010-10-18

基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2009CB320505）；国家科技支撑计划课题基金资助项目（2008BAH37B04）

Foundation Items: The National Basic Research Program of China (973 Program) (2009CB320505); The State Scientific and Technological Support Plan Project of China (2008BAH37B04)

括各种P2P协议在内的越来越多的应用不遵守默认端口约定或使用动态端口通信等原因，早期以IANA中注册的常用端口号区分应用协议流量的方法准确率已低于50%[1,2]，严重影响分析结果的可信性。依据报文负载内容识别应用协议的方法在主干网络带宽增长到10Gbit/s以上后是一个巨大的技术挑战，且该种方法无法处理流量加密的情况。

因此，自2004年开始基于行为特征识别应用流量的方法逐渐成为国际上研究的热点。这类方法首先归纳出各应用交互过程中在流/主机上表现出的不同行为特征，并以此为依据判别待分类流量所使用的应用协议。由于尚处于起步阶段，目前基于行为识别应用协议的方法不能精确识别单一的应用协议，而只能将流量分至大致的应用类别中。所谓应用类别，是对具有类似功能或行为的应用层协议的抽象概括，如BitTorrent、eDonkey等应用协议都属于P2P应用类别。

基于行为识别应用协议的方法可分为事先无训练集和有训练集2类，分别对应数理统计中的聚类分析和判别分析。使用聚类算法方面，A. McGregor[3]和Jeffrey Erman[4,5]等人分别使用EM和AutoClass等方法考虑流之间的相似性将流量分组，而后利用端口号或负载检查的方法分析其准确性。但聚类方法不能解释为什么流量会进行这样的分类，因此只能使用在对分类没有先验知识、没有训练集时，对类别进行初步探索上。判别算法方面，Thomas Karagiannis[6]等人分析应用类别在空间维上的行为特征(端口分布、链接数等)，构造主机交互关系图，并以此识别贡献流量的主机正在使用的应用协议类别。但该方法须对流量进行一定的累积，不仅有滞后性，而且在高速主干网络下，如何有效地存储流量，快速构造及匹配图本身就是一个仍待解决的问题。M. Roughan[7]和Sebastian Zander[8,9]等人基于k-NN和C4.5等机器学习方法，利用应用流在时间维上的传输特征(流长、持续时间等)将流量分至4~8个应用类别，然而这些早期方法的精度都不够高。

目前，最全面准确的方法是Andrew W. Moore等人于2005年提出的[11]。该方法使用TCP流的248个测度值[12]，通过对称不确定性推导测度间相关关系并进行筛选，而后利用基于核密度估计的Naïve Bayes分类法将TCP流分至10个应用类别中。虽然该方法较之前研究更多地考虑了测度的选择和分布，但存在以下很严重的效率问题。1) 所选用的248个流测度过多，且其中一些计算过于复杂。2) 使用熵和对称不确定性(SU, symmetric uncertainty)作为两变量相关性的度量，计算变量取值概率、条件概率的时空复杂度都非常高，样本空间较大时分类器训练时间过长，而样本空间较小时不足以代表流量总体行为，影响识别精度。3) 使用核密度估计(KE, kernel estimation)需要当每一个新流到达时都对样本空间中的每一个样本计算一次密度函数，开销非常大。由于上述缺点，虽然其实验表明方法准确率超过90%，但不能用于实时环境下，更不可能在线处理10Gbit/s以上的主干带宽流量。

国内目前对通用应用层流量分类的研究还处于匹配应用协议特征串的阶段[13]。基于行为识别流量的方法目前只针对于P2P流量的发现[14,15]，这些研究不仅通用性不好，准确率不高，而且都没有放在实际环境中进行识别率及性能的测试与分析。

故从发展现状看，目前基于行为特征的应用流量分类算法在精度和速度上都达不到令人满意的效果。尤其先前各算法在效率上无法实时处理吉比特以上的信道流量，并且各算法输入均为信道原始报文首部，前期报文采集、组流、测度计算的开销甚至远超过算法本身的计算开销。因此为了提高应用流量分类的效率，满足10Gbit/s以上高速主干网络管理和安全监测的需要，必须在保证足够准确率的前提下降低当前应用流量分类前期工作及算法本身各阶段的时空开销，以较以往研究更为简单有效的计算方法处理高速流量。

Cisco公司提出的NetFlow[16]是目前实际主要使用的IP流测量系统，已实现在多种路由器中，被业界厂商广泛支持。若能利用NetFlow已统计的流记录信息进行应用类别行为特征分析与流量分类，则不仅可以省略应用分类前期报文采集、组流、测度计算的时空开销，提高算法效率，而且基于Netflow流记录的标准性和广泛可用性，可以使其像SNMP一样支持现有的网络监控与管理应用，满足管理者全面了解网络活动方式，对各种业务流进行实时监测与管理的需求。

据此，本文提出一种以NetFlow记录统计信息作为输入的高速应用流量分类(FATC, fast application-level traffic classification)算法。算法分为基于简单相关系数的测度选择算法和基于Bayes多元判别分析的流量分类算法2部分。前者衡量

测度变量之间的相关关系，在实际分类之前选择出能揭示网络应用类别行为特征的测度，删除对分类无用及相互冗余的测度；后者以测度选择的结果作为分类的依据，将流量分类至误判损失最小的应用类别中。FATC 算法优点在于：1) 仅使用 NetFlow 统计的流测度作为判别的依据，不仅省略采集报文、组流、测度计算的巨大开销，且提高了方法的实用性；2) 使用相关系数作为变量间最本质的相关性判别依据，计算量小，且事前删除对判别无效或冗余的测度，优化后期分类过程；3) 使用 Bayes 判别法对应用流量分类，时间复杂度小，且实践证明：当样本空间足够大后，可以克服样本变量不服从多元正态分布的事实，使得基本的 Bayes 方法能够达到很好的效果；4) 算法具有超过 95%的分类准确率，且能实时处理当前 10Gbit/s 主干网络信道的流量。

2　高速应用流量分类(FATC)算法

2.1　基于相关系数的流测度选择算法

变量选择对判别方法的实施有着重要意义，过多的变量不仅影响判别方法的效率，无效或冗余的变量还会成为噪声影响判别方法的效果。因此，若能在实际流量分类前删除对分类无效的和相互冗余的测度，则不仅可以提高分类的精度和效率，还可以揭示出对流量分类有实际意义的测度，即那些能表示应用类别行为特征的测度。

目前只有文献[11]考虑了在实际分类之前对测度进行选择，但其采用的对称不确定性作为测度相关性依据需计算测度取值的概率和相互间的条件概率，方法时空复杂度都很高，训练及重训练分类器所用的时间开销太大。因此为了提高算法的效率，需采用计算过程更为简易的相关性计算方法。既然流测度(包括流所属的应用层协议类别)是随机变量，完全可以用经典统计分析中的简单相关系数来表示测度间的相关程度[17]：

$$r_{XY}=r_{YX}=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

其中，X 和 Y 为 2 个待研究流测度，(x_i, y_i) $(i=1,2,\cdots,n)$为两变量的 n 对观察值，$\bar{x}$ 和 $\bar{y}$ 分别为 n 个观察值的均值。$|r_{XY}|$取值在$[0, 1]$之间，数值越大表示相关程度越强，反之则越弱。令集合 $M=\{M_1, M_2, \cdots, M_n\}$为全部可选用的流测度组成的集合，$C$ 为流所属的应用层类别。若某流测度 M_i 与类别 C 之间的相关系数小于某事先设定的阈值，则认为该测度不能提供对分类有用的信息，属于无效测度；若两测度之间的相关系数大于另一阈值，则认为这 2 个测度相互冗余，需删除其中贡献较小的测度。据此，基于相关系数的特征选择算法如下：

```
Begin
for each M_i ∈ M do
        compute r_{M_i,C};
        if |r_{M_i,C}| < d_1                    // d_1 为有效测度选择阈值
                M = M − {M_i};
end
rank M_i in descending order;
for each M_i ∈ M do
        for each M_j ∈ M and M_j after M_i do
                compute r_{M_i,M_j};
                if |r_{M_i,M_j}| ≥ d_2          // d_2 为冗余测度选择阈值
                        M = M − {M_j};
                end
end
End
```

最终测度集合 $M=\{M_1,M_2,\cdots,M_m\}$只包含了能对分类提供有用信息且相互独立的测度。另外，测度选择算法中阈值 d_1 和 d_2 的不同取值会影响入选的测度，继而影响分类算法的准确性和效率。二者的设置依赖于经验和实验的效果，本文第 3 节中将进一步分析不同阈值取值对 FATC 算法准确率的影响。

2.2　基于 Bayes 判别分析的流量分类算法

在利用相关系数对流测度进行筛选的基础上，本节给出以最终集合 M 中的测度为分类依据的基于 Bayes 判别分析的应用流量分类算法。

多元统计分析的 Bayes 判别方法建立在 Bayes 准则的基础上，偏重于集群分布的统计特性，分类原理是假定训练样本数据的光谱空间服从某类分布，做出样本的概率密度等值线，确定分类，然后

通过计算待判别样本属于各类别的概率，将新样本归属于概率最大的一组。Bayes 判别方法由于需要对所研究的对象在抽样前已有一定的认识(先验分布)，且考虑误判后的损失，故判别精度往往高于其他线性判别方法[17]。

令应用类别总数为 k，则 Bayes 判别方程为

$$h_i(x)=\sum_{j=1,\ j\neq i}^{k} q_j\, p_j(x) C(i\mid j) \tag{1}$$

$$h_t(x)=\min_{1\le i\le k} h_i(x) \tag{2}$$

式(1)中 q_j 为第 j 类别的先验概率，$p_j(x)$为待判别对象 x 属于第 j 类别的概率，$C(i|j)$称为损失函数，表示本应属于第 j 类别的对象误判给第 i 类别的损失：当 $i=j$ 时，有 $C(i|j)=0$；当 $i?j$ 时，有 $C(i|j)>0$。显然式(1)是对损失函数依概率的加权平均，即 $h_i(x)$ 为把 x 判给第 i 类别的损失期望。式(2)表明以误判损失最小作为分类的依据，即使得 $h_i(x)$最小的 i 即是对象 x 应属的类别号。

原则上说，考虑损失函数更为合理，误判损失 $C(i|j)$可以根据网络管理的不同需求来设置。如若当前较为关注 P2P 流量情况，则可将 P2P 流误判给其他类别的损失相应增大。由于本文公平考虑各应用类别，此处假定各种误判的损失皆相等，即

$$C(i\mid j)=\begin{cases}0, & i=j\\ 1, & i\neq j\end{cases}$$

则判别方程简化为

$$h_t(x)=\min_{1\le i\le k} h_i(x)=\min_{1\le i\le k}\sum_{j=1,\ j\neq i}^{k} q_j\, p_j(x)$$

当 x 取定时由于 $\sum_{i=1}^{k} q_i\, p_i(x)$ 为常数，则

$$\min_{1\le i\le k}\sum_{j=1,\ j\neq i}^{k} q_j\, p_j(x)\Leftrightarrow\max_{1\le i\le k} q_i\, p_i(x)$$

故判别方程等价于

$$h_t(x)=\max_{1\le i\le k} q_i\, p_i(x)\text{。}$$

假设流对象 $X=(M_1,M_2,\mathrm{L},M_m)^{\mathrm{T}}$ 服从多元正态分布（3.5 节将通过实验说明只要样本空间足够大，就可以克服流测度不服从正态分布的事实），其中流属性 $M_1\sim M_m$ 对应于应用第 2.1 节的测度选择算法所得到的最终测度。X 的分布密度函数为

$$f(x)=(2\mathrm{p})^{-n/2}\left|S\right|^{-1/2}\exp\left\{-(x-m)'S^{-1}(x-m)/2\right\}$$

其中，先验概率 q_i、均值向量 m 和方差阵 S 可利用样本通过无偏估计得到：

$$q_i=n_i/n$$

$$\hat{m}_i=\frac{1}{n_i}\sum_{x_j\in C_i} x_j \tag{3}$$

$$\hat{S}_i=\frac{1}{n_i-1}\sum_{x_j\in C_i}(x_j-\hat{m}_i)\,(x_j-\hat{m}_i)' \tag{4}$$

其中，n 为样本空间大小，即总样本流个数。n_i 为属于第 i 类别的样本流个数。根据微分中值定理，有

$$p_i(x)\approx p_i(X\in(x,x+e))\approx f_i(x)e$$

由于 n 和 e 为定值，故判别方程可化为

$$h_t(x)=\max_{1\le i\le k}\ n_i\left|\hat{S}_i\right|^{-1/2}\exp\{-(x-\hat{m}_i)'\,\hat{S}_i^{-1}(x-\hat{m}_i)/2\}$$

其中，未知数只有待判别流对象 x。

据此，基于 Bayes 判别分析的流量分类算法如下：

Begin

compute $n[i]$, $\mu[i]$, $S[i]$ for each $i\in C$ according to formulas (3) and (4);

for each unclassificated flow X do

get value of metrics $X[1]$~$X[m]$ from flow record;

compute $h[i]=n[i]\left|S[i]\right|^{-1/2}\exp\left\{-(x-m[i])'S[i]^{-1}(x-m[i])/2\right\}$ for each $i\in C$;

let $h[t]=\max h[i]$ for each $i\in C$;

//t 即为流 X 属于的应用类别号

end

End

3 实验结果与分析

3.1 实验数据及流测度

为便于对比算法效果，本文采用和 Andrew 相同的实验 TRACE[11]：采集信道为一条吉比特全双工以太网，在一天内随机选取了 10 个持续时间约为 1 680s 的时间片，提取其中完整的 TCP 流，计算出每个 TCP 流的 248 种流测度及所属的应用层协议类别，作为 TRACE 中的记录。表 1 详细列举了 TRACE 中的各应用层协议类别及属于该类别的流数。

表 1　应用层类别及相应流数

应用类别	流数
BULK	11 539
WWW	328 091
MAIL	28 567
P2P	2 094
ATTACK	1 793
DATABASE	2 648
MULTIMEDIA	1 152
SERVICES	2 099
INTERACTIVE	110
GAMES	8
Total	377 526

Andrew 使用 248 种测度作为可用测度集合，从简单的 TCP 端口号至复杂的傅里叶变换。这不仅要求系统采集所监听网络上的每个报文并组流，且需占用很大的资源计算这些测度。而本文提出的 FATC 算法的可用测度仅限定为 NetFlow V5 统计可得的(如表 2 所示)，不仅省去采集报文、组流、计算测度的前期工作，降低了系统开销，而且简化后期测度选择算法和流量分类算法的输入，使分类更高效。

表 2　可用流测度集合

测度	描述
pkts	流内报文数
bytes	流内字节数
duration	流持续时间（单位：s）
s_port	服务器端口（流中低位端口号）
IAT	流内平均报文间隔时间：duration/pkts
pkt_size	流内平均报文大小：bytes/pkts
pps	pkts/duration
Bps	bytes/duration

算法测试前期工作还包括将 Andrew 的 TRACE 转换至 NetFlow V5 流记录格式，其中 IP 地址、AS 号等 FATC 算法的无用字段可忽略。

3.2　算法准确率分析

首先给出 2 个评价算法准确率的标准。

定义 1　$$\text{总准确率} = \frac{\text{正确分类的总流数}}{\text{待分类的总流数}}$$

定义 2

$$\text{某类别准确率} = \frac{\text{正确分至该类的流数}}{\text{真正属于该类的待分类流数}}$$

算法测试时，本文在实验 TRACE 的 10 个时间片内任取 5 个作为训练集，另 5 个作为测试集，取该类组合共 $C(10, 5) = 252$ 组中随机 20 组实验后的均值作为最后结果。取 d_1=0.06，d_2=0.6(第 3.4 节将说明二者不同取值对 FATC 算法的影响)，则最终测度集合 M={s_port, pkt_size, IAT, duration}。FATC 算法准确率如表 3 所示。

表 3　算法准确率比较

应用类别	准确率/%	
	Andrew	FATC
BULK	82.25	88.38
WWW	99.27	97.20
MAIL	94.78	75.80
P2P	36.45	34.17
ATTACK	13.46	10.01
DATABASE	86.91	97.65
MULTIMEDIA	80.75	96.97
SERVICES	63.68	84.79
INTERACTIVE	0	24.68
GAMES	0	0
Total	96.29	93.25

由表 3 可见，FATC 算法准确率远高于基于端口分类流量的方法，在大多数类别上也高于 Andrew 所提出的流分类算法，然而总准确率略低于 Andrew 算法。造成差异的原因有 2 个：1）由于 WWW 类别流数量占据了流总数的 87%，对其略低的识别率将极大地影响总准确率；2）由于 FATC 算法的输入来自 NetFlow 流统计信息，相较于 Andrew 所用的 248 个测度，极大地减少了所提供的类别行为特征信息。但是，一方面如 3.5 节所示，随着训练集空间的增长 FATC 算法准确率上升，9 个时间片时准确率已为 95.7%，可以弥补缺少测度信息带来的不足；更重要的一方面，如 3.3 节所示，FATC 算法极大的降低了以往分类算法的时空复杂度，使得在可接受的精度损失下分类效率有极大的提高。表 3 还表明两算法对 GAMES、INTERACTIVE 和 P2P、ATTACK 的识别率都非常低。Andrew 并未对此现象作出解释。分析如下。1) 由于 INTERACTIVE 和 GAMES 2 种类别的流数非常少（如表 1 所示），不足以提供该类别的行为特征信息，造成这 2 类流

量识别率极低。2) 对于 P2P 和 ATTACK，由表 1 可知这 2 类应用的流数并不少。是由于这 2 种应用涵盖范围很广，各协议间行为差异较大，造成算法很难对其进行类别的行为特征归纳，致使判别出现偏差。更进一步的证据和处理方法将是下一步研究的重点。

3.3 算法时空效率分析

3.3.1 时间效率

训练算法中，Andrew 使用的 SU 算法需多遍扫描样本空间或内存空间以统计测度取值概率和条件概率，而 FATC 算法中的简单相关系数仅需单遍遍历样本空间。识别算法中，Andrew 使用的 KE 算法在每个新流到达时需对样本空间中的每个样本计算一次密度函数，而 FATC 算法只需计算应用类别数次的密度函数。故即使样本空间中只有 10^4 条流记录，分为 10 类，则使用 KE 的 Andrew 算法在分类过程的时间开销是 FATC 算法的 10^3 倍。

由上可见，为了提高 Andrew 算法的效率必须使用较小的样本集。而小样本空间不足以提供完全的行为分布信息，会使算法的结果产生很大的偏差。因此 Andrew 算法存在着性能—效率的矛盾。文献[11]表明，分别使用不足 25 000 条流记录训练并测试的情况下，其算法时间开销约 300s，而同样条件下 FATC 算法仅需 4s。更严重的是，实际使用时 Andrew 算法还需采集原始报文、组流、计算 248 个测度，这更使得该算法不可能应用于超过 1Gbit/s 的网络环境中。

FATC 算法现每秒约能处理 18 000 条流记录。据华东(北)地区网络中心日常统计，地区主干到国家主干的 10Gbit/s 信道一天内的流数不足 800MB，即 FATC 算法理论上能在不到 12h 内处理完目前该 10Gbit/s 信道 24h 的流量。考虑到当前实验为读取硬盘上的 TRACE，速度较慢，实际使用直接从路由器接收 NetFlow 格式的流记录时 FATC 算法效率会有更为明显的提高，完全满足实时分类当前 10Gbit/s 主干网络流量的需求。

3.3.2 空间效率

样本存储空间：由于 Andrew 算法可用测度集合庞大，若样本数相同，则其所需的样本存储空间约为 FATC 算法的 30 倍(248/8)。因此使用同样的磁盘或内存空间，FATC 算法可以较 Andrew 算法多存储约 30 倍的样本流记录。

计算内存空间：在应用类别数目一定的情况下，Andrew 的 SU 算法在统计样本取值概率 $p(x_i)$和条件概率 $p(x_i|y_j)$时所需内存空间随样本数和测度数的增长而增长。同时由于 KE 算法在每个新流到达时需对样本空间中的每个样本计算一次密度函数，出于效率考虑显然应将每条样本记录都放在内存中。文献[11]表明使用全部测度，在样本空间不足 25 000 条流记录的条件下(仅 2/3 个时间片大小)，其内存使用达到 256MB。而 FATC 中的测度选择算法和分类算法所需内存空间不随样本数和测度数的增长而增长，仅需记录各应用类别样本的均值和方差，运行总内存不足 70kB，为 Andrew 算法的约 $1/10^4$。

3.4 测度选择阈值对算法的影响

任何测度选择算法的效果都和其筛选测度的阈值相关。基于简单相关系数的测度选择算法的效果好坏也很大程度上取决于 2 个参数的取值：有效测度选择阈值d_1和冗余测度选择阈值d_2。d_1取值过小会将某些对分类无效的测度引入分类算法中，d_2取值过大会将本身冗余的测度认定为彼此独立，二者不仅增加分类算法的计算复杂度，而且可能影响分类算法的效果；而d_1取值过大可能会淘汰掉某些对分类有用的测度，d_2 的取值过小会使本身互相独立的测度被认定成冗余而被删除，这更会极大地降低分类算法的准确率。

图 1 表现了 2 参数的不同取值对 FATC 算法准确率的影响。由图 1 可见 FATC 算法对 2 个参数取值的选择，即测度的选择要求很高。选择不适宜的测度将导致算法的准确率一直非常低(10%~20%)，而合适的参数取值则能够选中最能表现应用类别行为特征的测度，使算法准确率有很大提高(大于 90%)。另外，由测度选择算法可知图中d_1=0，d_2=1 的点为未对测度进行筛选，使用表 2 中所有测度进行流量分类的效果，其准确率只有约 25%。可见使用合适的方法在流量分类之前剔除杂音与冗余特征，不仅可以精简分类器的结构，同时也极大提高了分类器的准确率。然而就如何决定测度的取值，目前的研究还没有很好的方法，仍只能通过平时的经验和实验得出，这也是今后需要继续考虑的内容之一。

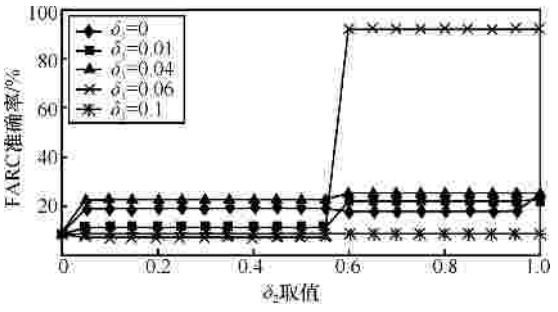

图 1 参数对 FATC 算法准确率的影响

3.5　训练集大小对算法的影响

图 2 显示了 FATC 算法准确率随训练集大小的变化情况。由图可见如下。1) 过小的训练集不能体现各应用流量总体分布的情况，导致使用不全面的信息训练判别算法，从而影响 FATC 算法分类的精度。随着训练集中样本数量的增加，训练集所能提供的流量分布信息增多，分类算法就越能根据已知的正确信息判断新流的所属类别，算法准确率不断上升。2) 当训练集大小超过 4 个时间片时，FATC 算法准确率的增加逐渐缓慢。此时再增加训练样本的效果并不明显。同时，较小的训练集不仅可以降低手动构造训练集所需的前期工作量，而且可以减少算法在训练及重训练过程的时间开销。因此实际中可根据所要求的准确率调整初始训练集大小，以较小的工作量得到所需的精度。当精度需求提高时，可以相应增大样本空间，以补充信息。3) 当训练集大小达到 8 个时间片时，FATC 算法准确率已超过 95%，9 个时间片时的准确率为 95.7%，非常接近 Andrew 所提出的算法。由此可见，只要训练样本空间满足一定大小，就可以破除 Bayes 判别中对样本正态分布的假设，达到 Andrew 使用 KE 算法相同的效果；另一方面，实验表明即使训练集包含 9 个时间片，FATC 算法在训练阶段的时间开销仍只有 12s，远小于训练集只包含 2/3 个时间片的 Andrew 算法，且不影响分类过程的时间复杂度。

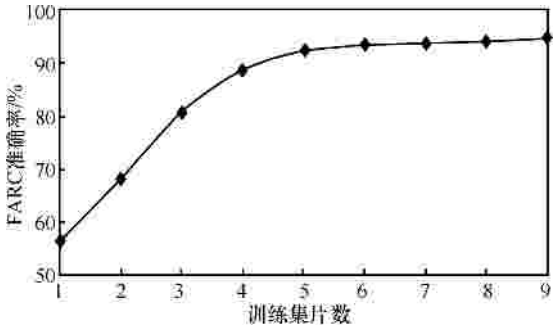

图 2　FATC 算法准确率随样本集大小变化曲线

3.6　流量行为变化对算法的影响

基于流量行为特征的应用流量分类算法都会面临网络流量行为随着时间推移发生变化的问题。其是由新应用协议的加入、网络管理策略的改变、用户习惯的转变等原因所造成，包括各应用类别流量在总体流量中所占比重的变化和流测度分布的改变 2 个方面。对 FATC 算法而言，前者改变判别方程中各类别的先验概率，后者影响各类别的均值和方差。故使用旧样本训练的分类器精度会随着时间的推移逐渐降低。表 4 为使用原样本进行训练，并使用 12 个月之后的另一组 TRACE 进行测试所得的 FATC 算法准确率，其中 3 个类别对应的 N/A 表示该测试 TRACE 中没有该类别的流量，同时 FATC 算法也未将任何其他类别的流量误分至该类别。由表可见虽然基本各类别准确率都稍有下降，但总体仍保持有较高的精度，总准确率在一年之后仍维持在 90%以上，只下降了不到 3%。实验结果表明 FATC 算法具有很强的时间适应性，可以长时间稳定的监测网络流量，在必要时只需稍加新样本进行重训练就可恢复算法原先的精度。

表 4　使用较晚采集的 TRACE 对算法的测试结果

应用类别	准确率/%
BULK	98.20
WWW	91.26
MAIL	61.84
P2P	25.15
ATTACK	N/A
DATABASE	88.06
MULTIMEDIA	N/A
SERVICES	87.05
INTERACTIVE	75.00
GAMES	N/A
Total	90.37

4　结束语

针对目前应用流量分类算法效率不高，不能满足主干网中流量监测需求的现状，本文提出一种以 NetFlow 统计信息作为输入，利用不同应用类别在交互过程中表现出的行为测度差异区分各应用类别流量的高速应用流量分类算法——FATC。算法使用多元数理统计中的简单相关系数作为测度间相关性依据，在分类之前选择对分类有效且彼此独立的测度，并以这些测度为依据使用 Bayes 判别法将流量分至误判损失最小的应用类别。相较于之前的研究，FATC 算法有以下改进。1) 首次使用 NetFlow 记录信息作为输入，消除了前期报文采集、组流、测度计算的开销，减少了输入数据量，且使算法更具实用性。2) 极大降低分类算法在训练及分类过程的时空复杂度，使算法具有极高的效率。理论分析和实验结果表明，FATC 算法具有超过 95%的分类准确率，在保持当前最全面准确的 Andrew 方法准确率的基础上，将其时空开销降低至少 10^3 倍，能实时稳定地分类当前 10Gbit/s 主干信道的流量。

下一步工作将深入地分析应用层协议分类中流测度的选择问题，进一步借鉴文献[10]和文献[18~20]中所述的流量统计属性揭示应用层流量分类与流记录详细程度之间的关系，研究流测度的种类、个数和应用类别分类粒度之间的对应关系，以及不同流测度对识别不同应用类别流量的重要程度，从而为当前流信息统计系统和网络监测系统的改进提供信息。

参考文献：

[1] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network applications[A]. Proc of PAM 2005[C]. oston, USA, 2005. 41-54.

[2] KIM M S, WON Y J, HONG J W K. Application-level traffic monitoring and an analysis on IP networks[J]. ETRI Journal, 2005, 27(11): 22-42.

[3] MCGREGOR A, HALL M, LORIER P, *et al*. Flow clustering using machine learning techniques[A]. Proc of PAM 2004[C]. Antibes Juan-les-Pins, France, 2004. 205-214.

[4] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms[A]. Proc of ACM SIGCOMM Workshop on Mining Network Data 2006[C]. Pisa, Italy, 2006.281-286.

[5] ERMAN J, MAHANTI A, ARLITT M. Internet traffic identification using machine learning[A]. Proc of 49th IEEE Global Telecommunications Conference[C]. San Francisco, USA, 2006. 1-6.

[6] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[A]. Proc of ACM SIGCOMM 2005[C]. Philadelphia, USA, 2005.229-240.

[7] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[A]. Proc of ACM SIGCOMM IMC 2004[C]. Taormina, Italy, 2004. 135-148.

[8] ZANDER S, NGUYEN T, ARMITAGE G J. Self-learning IP traffic classification based on statistical flow characteristics[A]. Proc of PAM2005[C]. Boston, USA, 2005. 325-328.

[9] ZANDER S, WILLIAMS N, ARMITAGE G. Internet archeology: estimating individual application trends in incomplete historic traffic traces[A]. Proc of PAM 2006[C]. Adelaide, Australia, 2006. 205-206.

[10] LI M, ZHAO W. Representation of a stochastic traffic bound[J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(9): 1368-1372.

[11] MOORE A W, ZUEV D. Internet traffic classification using bayesian analysis techniques[A]. Proc of ACM SIGMETRICS 2005[C]. Banff, Canada, 2005. 50-60.

[12] MOORE A W, ZUEV D. Discriminators for Use in Flow-Based Classification[R]. RR-05-13, London: Intel Research, Cambridge, 2005.

[13] 刘元勋, 徐秋亮, 云晓春. 面向入侵检测系统的通用应用层协议识别技术研究[J]. 山东大学学报(工学版), 2007, 37(1): 65-69.
LIU Y X, XU Q L, YUN X C. Research on IDS-faced general-purpose application-level protocol identification technology[J]. Journal of Shandong University(Engineering Science), 2007, 37(1): 65-69.

[14] 宫婧, 孙知信, 顾强. 基于行为特征描述的 P2P 流识别方法的研究[J]. 小型微型计算机系统, 2007, 28(1): 48-53.
GONG J, SUN Z X, GU Q. Research of identification method based on P2P flow behavior characterization [J]. Journal of Computer Systems, 2007, 28(1): 48-53.

[15] 刘斌, 李之棠, 李佳. 一种基于流特征的P2P流量实时识别方法[J]. 厦门大学学报(自然科学版), 2007, 46(2): 132-135.
LIU B, LI Z T, LI J. A new method on P2P traffic identification based on flow[J]. Journal of Xiamen University(Natural Science), 20 46(2): 132-135.

[16] Cisco. cisco IOS netflow introduction[EB/OL]. http://www.cisco.com/warp/public/ 732/Tech/NetFlow, 2006.

[17] 朱道元, 吴诚鸥, 秦伟良. 多元统计分析与软件 SAS[M]. 南京: 东南大学出版社, 1999.
ZHU D Y, WU C O, QIN W L. Multivariate Statistical Analysis and SAS [M]. Nanjing: Southeast University Press,1999.

[18] LI M, LIM S C. Modeling network traffic using generalized Cauchy process[J]. Physica A, 2008, 387(11): 2584-2594.

[19] LI M. Change trend of averaged Hurst parameter of traffic under DDOS flood attacks[J]. Computers & Security, 2006, 25(3): 213-220.

[20] LI M. An approach to reliably identifying signs of DDOS flood attacks based on LRD traffic pattern recognition [J]. Computers & Security, 2004, 23(7): 549-558.

作者简介：

陈亮（1981-），男，江苏南京人，东南大学博士生，主要研究方向为网络行为学。

龚俭（1957-），男，上海人，博士，东南大学教授、博士生导师，主要研究方向为网络行为学、网络安全、网络管理。