基于隐马尔科夫模型的 P2P 流识别技术

许博，陈鸣，魏祥麟
(解放军理工大学 指挥自动化学院，江苏 南京 210007)

摘 要：为了实时、准确地识别多种 P2P 应用流，提出了基于隐马尔科夫模型(HMM, hidden Markov model)的 P2P 流识别技术。该技术利用分组大小、到达时间间隔和到达顺序等特征构建流识别模型，采用离散型随机变量刻画 HMM 状态特征；提出了能同时识别多种 P2P 应用流的架构 HMM-FIA，设计了 HMM 的状态个数选择算法。在校园网中架设可控实验环境，使用 HMM-FIA 识别多种 P2P 流，并与已有识别方法进行比较，结果表明采用离散型随机变量能降低模型建立时间，提高识别未知流的实时性和准确性；HMM-FIA 能同时识别多种 P2P 协议产生的分组流，并能较好地适应网络环境变化。

关键词：对等方到对等方；有限状态机；流识别；隐马尔科夫模型
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2012)06-0055-09

Hidden Markov model based P2P flow identification technique

XU Bo, CHEN Ming, WEI Xiang-lin
(Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: To identify various P2P flows accurately in real-time, a hidden Markov model (HMM) based P2P flow identification technique was proposed. This approach made use of packet size, inter-arrival time and arrival order to construct flow identification model, in which discrete random variable was used to depict the characteristics of HMM state. A framework called HMM-FIA was proposed, which could identify various P2P flows simultaneously. Meanwhile, the algorithm for selecting the number of HMM state was designed. In a controllable experimental circumstance in the campus network, HMM-FIA was utilized to identify P2P flows and was compared with other identification methods. The results show that discrete random variable can decrease the model constructing time and improve the time-cost and accuracy in identifying unknown flows, HMM-FIA can correctly identify the packet flows produced by various P2P protocols and it can be adaptive to different network circumstance.

Key words: peer to peer; finite state machine; flow identification; hidden Markov model

1 引言

因特网中新型网络应用不断涌现，流量成分日趋复杂，这使得优化网络结构、维护网络安全、强化网络管理以及理解网络行为面临挑战。基于端口识别流的方法简单、高效，在传统应用流识别中发挥了重要作用[1,2]。但 P2P 应用广泛采用动态端口、端口跳变和端口伪装等技术，使该识别方法不再可行[3]。基于特征字的 P2P 流识别方法[4, 5]具有识别准确率高、可在线处理等优点，其缺点是需要深度检

收稿日期：2010-09-15；**修回日期**：2011-01-04
基金项目：国家高技术研究发展计划(“863”计划)基金资助项目(2007AA01Z418)；江苏省自然科学基金资助项目(BK2009058)；国家自然科学基金资助项目(61072043)
Foundation Items: The National High Technology Research and Development Program of China (863 Program) (2007AA01Z418) The Natural Science Foundation of Jiangsu Province (BK2009058); The National Natural Science Foundation of China(61072043)

测分组应用层负载，无法识别采用信息加密传输的协议，且提取应用协议特征字比较困难。基于运输层统计信息和行为特征的启发式识别方法是当前的研究热点。这类方法利用流的属性、统计特征以及行为特征，按照启发式规则对流进行分析，达到分类和识别的目的[6,7]。但此类方法大多都无法高效、实时地识别特定协议产生的分组流。

本文提出了一种基于隐马尔科夫模型(HMM)识别 P2P 流的新方法，将 HMM 中的隐状态序列与分组流观察序列对应起来，并采用离散型随机变量刻画状态特征，有效地提高了识别 P2P 流的实时性和准确性。实验结果表明此方法能准确、快速地识别多种 P2P 应用协议产生的分组流。

2 相关工作

Mena 等人首先将流统计特征应用于流识别，他们利用分组大小和到达时间间隔的统计特征识别出实时音频流 [8]。A. McGregor 等人使用 EM 算法[9]，利用运输层统计信息，如分组大小、流字节数、连接持续时间等特征对流进行聚类并产生分类器。M. Roughan 等人提出从不同层面(单个分组、流、连接)提取流统计特征的方法[10]，采用最近邻和线性分类方法，利用连接持续时间和平均分组大小产生分类器，并将观测到的流划分为 4 种不同类型。Zander 等人使用 AutoClass 聚类算法(EM 算法的一种实现)，选择最佳聚类属性子集并建立分类器[11, 12]。J. Erman 等人采用 K-均值聚类算法(K-means)和基于密度的空间聚类算法(DBSCAN)产生分类器[13]。L. Bernaille 等人依据分组大小和方向，使用 K-Means、Gaussian Mixture 和 spectral clustering 算法构造分类器 [14]。M. Crotti 等人提出协议指纹 (protocol fingerprinting)概念[15]，由学习过程计算出协议指纹中的协议掩模(protocol mask)，在分类算法中遍历分类引擎中的已知协议集，计算观测到的流相对于各协议掩模的异常估值，据此进行分类。上述方法根据训练样本产生的分类器对未知流进行分类，但它们均无法准确识别产生分组流的应用协议，且只适应于特定网络环境或对特定应用有较高识别准确率。C.Wright 等人提出使用 HMM 识别分组流[16]，他们首先单独考虑分组大小和到达时间间隔特征，之后将分组大小与到达时间间隔进行矢量量化[17]，采用量化后的一维特征识别分组流，该方法没有考虑不同特征之间的相关性。Alberto 等人利用分组大小和到达时间间隔构造二维向量特征，提高了识别准确性，但他们采用连续型随机变量，降低了识别的实时性，同时缺乏不同 HMM 之间的相似度比较[18]。

3 基本思想

P2P 流识别实际上是根据收集的流信息识别出某种 P2P 网络应用协议的过程。网络协议可使用有限状态机描述，但协议有限状态机位于节点内部，其状态及状态间的转移特征均被节点屏蔽，只有在节点之间交互的分组才能被外部观察到。HMM 是一个二重马尔科夫随机过程，包括了能够输出观测值的随机过程和具有状态转移概率的马尔科夫链，其状态通过观测序列的随机过程表现出来。HMM 包含 2 层，一个可观察层和一个隐藏层。可观察层是待识别的观察序列，在 P2P 流识别中表现为节点间交互的多个连续分组，隐藏层是一个马尔科夫过程，即运行在节点内部的一个有限状态机，其中每个状态转移都带有转移概率。

在 HMM 中，对于一个随机事件，有一个观察值序列 $O=\{v_1,v_2,\cdots,v_N\}$，并隐含着一个状态序列 $S=\{s_1,s_2,\cdots,s_M\}$。本文提出的识别方法，采用分组大小和到达时间间隔作为构造观察序列 O 的二维向量，与 C.Wright 等人提出的方法[16,17]相比，更多的保留了不同特征之间的关联关系，提高了识别准确率；使用相同 P2P 协议交互的节点，协商阶段通常具有共同的状态转移序列[15]，假定观察序列 O 中的每个值 v_i 由 S 中唯一确定的状态 s_j 产生，对任意观察序列 O，与其对应的状态序列 S 是确定的，且采用离散型随机变量刻画 HMM 的状态特征，与 Alberto 等人的方法[18]相比，降低了模型的计算复杂度，提高了识别 P2P 流的实时性；此外，提出比较不同 HMM 之间相似度的指标，以及通过增加模型的状态个数降低误判率的方法。基于 HMM 识别 P2P 流的思想如图 1 所示。

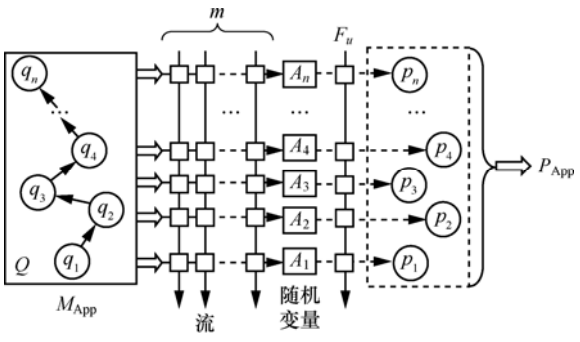


图 1 基于 HMM 的 P2P 流识别思想

假定协议 App 对应的有限状态机为 M_{App}，其有限状态集合为 $Q=\{q_1,q_2,q_3,\cdots,q_n\}$，可将 Q 看作 HMM 的隐状态集合，状态间的转移概率由 M_{App} 决定，且状态处于 q_i 时发出的分组特征集合表示为 $V_i=\{v_{i1},v_{i2},\cdots,v_{ic}\}$，将出现每个特征的概率记作 $P(v_{ij}\,|\,q_i)=b_{ij}$ $(1\leqslant \mathrm{j}\leqslant c)$，使用离散型随机变量 A_i 表示状态处于 q_i 时的分组特征分布。当网络通信质量较好，无分组丢失和失序时，HMM 样本中的分组对应的状态是相同的，由协议有限状态机和分组到达顺序唯一确定。在上述假定下，为 HMM 选取 m 个训练样本，使用 F 表示 m 条样本流集合，其中任意流 F_{App} 表示为 $F_{\text{App}}=\{f_1,f_2,f_3,\cdots,f_r\}$，$F_{\text{App}}$ 中任意分组 $f_j\in F_{\text{App}}$ 由 Q 中对应的唯一状态 q_i 产生，因此，样本集合 F 为每一个状态 $q_i(i\leqslant r)$ 提供了 m 个训练样本，可用来获得与 q_i 对应的随机变量 A_i 的分布函数。对未知的分组流 F_u，可求得其中每个分组 f_i 由对应状态产生的概率 p_i，最终得到 F_u 由 App 产生的概率 P_{App}，进而可根据 P_{App} 判定分组流 F_u 是否属于 App。

4　基于 HMM 的流识别模型

4.1　建立 HMM

给定协议 App，将其对应的 HMM 表示为 H_{App}，集合 $F=\{F_1,F_2,\cdots,F_m\}$ 表示由 App 产生的 m 条分组流，将作为 H_{App} 的训练样本集。与以往工作不同，本文采用由五元组（源宿 IP 地址、源宿端口号和传输层协议）流规范和 64s 超时定义的双向流作为训练样本，这样最大限度地保留了协议特征，并且 F_i 中分组无丢失、无乱序、无重传，防止造成 H_{App} 的训练误差。选择分组长度(PL, packet length)、分组到达时间间隔(IAT, inter-arrival time)和分组到达顺序作为构建 H_{App} 的特征，并将流 F_i 中第 k 个分组表示为 $g_{ik}=\pm\big(b_i(k),d_i(k)\big)^{\mathrm{T}}$，其中 $b_i(k)$ 表示第 k 个分组的长度，$d_i(k)$ 表示第 k 个分组的到达时间间隔，且 $d_i(k)=10\lg(I_k/1\mu\mathrm{s})$，$I_k$ 为第 k 个分组实际到达时间间隔，“+”表示分组由源 IP 地址发往目的 IP 地址，“−”表示反方向分组，则 F_i 可表示为 $F_i=\big(g_{i1},g_{i2},\cdots,g_{ir},\cdots\big)$。集合 $S=\{s_1,s_2,\cdots,s_r\}$ 为 Q 的子集，表示 H_{App} 的 r 个隐状态，对任意状态 s_j，其训练样本中对应的分组集合表示为 $G_j=\{g_{ij}:1\leqslant i\leqslant m\}$。采用 a_{ij} 表示 H_{App} 的状态由 s_i 转移到 s_j 的概率，$\boldsymbol{A}=\{a_{ij}\}$ 表示状态转移概率矩阵。H_{App} 的状态转移与分组序列相关，若能捕获到完整的分组序列，则状态转移概率为

$$a_{ij}=\begin{cases}1, j=i+1\\ 0, \text{其他}\end{cases}$$

但在被动测量时，分组可能会丢失，无法获得完整的分组序列，使得状态无法按序转移。将分组丢失的概率设为 q，并假定连续丢失的概率为 0，则 H_{App} 中的状态转移情况如图 2 所示。

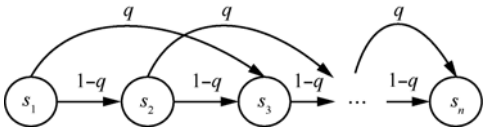


图 2　存在分组丢失的状态转移概率

相应的状态转移概率为

$$a_{ij}=\begin{cases}q^{\alpha}(1-q)^{1-\alpha},\alpha=j-i-1 \ \text{且}\ i<j\leqslant i+2\\ 0, j\leqslant i \ \text{或}\ j>i+2\end{cases}$$

其中，q 的取值可以通过网络测量来估计。

以往工作采用连续型随机变量构建模型参数，通常计算量较大，不利于实时识别和处理。本文采用划分区间的方式建立离散模型，这样可以有效降低计算复杂性，提高系统实时处理能力。分别使用 $pl_i([b',b'']_k)$ 和 $iat_i([d',d'']_k)$ 表示 H_{App} 处于状态 s_i 时，PL 落在区间 $[b',b'']_k$ 的概率和 IAT 落在区间 $[d',d'']_k$ 的概率，简写为 $pl_i(k)$ 和 $iat_i(k)$，则有：

$$pl_i(k)=\mathrm{P}(PL\in[b',b'']_k\,|\,s_i)$$

$$iat_i(k)=\mathrm{P}(IAT\in[d',d'']_k\,|\,s_i)$$

H_{App} 的状态处于 s_i 时分组长度区间集合记为 $\{[b',b'']\}_i$，区间个数记为 K_{i1}。$\{[b',b'']\}_i$ 可根据分组集合 G_i 来获得，即将 G_i 中 m 个分组长度由小到大排列，若相邻 2 个分组长度之差小于门限 ε_{PL}，则它们属于同一区间，并使用 $L[b',b'']_k$ 表示样本值落入区间 $[b',b'']_k$ 的个数。同理可获得分组到达时间间隔区间集合 $\{[d',d'']\}_i$，区间个数记为 K_{i2}，其门限为 ε_{IAT}，落入区间 $[d',d'']_k$ 的分组个数为 $L[d',d'']_k$。给定训练样本集合 F，$|F|$ 表示样本个数，则有：

$$pl_i(k)=\frac{L[b',b'']_k}{|F|} \tag{1}$$

$$iat_i(k)=\frac{L[d',d'']_k}{|F|} \tag{2}$$

使用 $\boldsymbol{B}^{pl}=\{pl_i(k):1\leqslant i\leqslant r,1\leqslant k\leqslant K_{i1}\}$ 和 $\boldsymbol{B}^{iat}=\{iat_i(k):1\leqslant i\leqslant r,1\leqslant k\leqslant K_{i2}\}$ 分别表示 PL 和 IAT 的观察概率矩阵。使用 $\pi=\{\pi_i:1\leqslant i\leqslant r\}$ 表示初始状态概率分布，则有：

$$\pi_i\begin{cases}1-q,i=1\\q,i=2\\0,\text{其他}\end{cases}$$

通过上述分析，可使用四元组 $\{\boldsymbol{A},\boldsymbol{B}^{pl},\boldsymbol{B}^{iat},\pi\}$ 表示 H_{App}。

4.2 计算未知流对应 HMM 的概率

对未知流 $F_u=\{g_1,g_2,\cdots,g_r,\cdots\}$，可用 $\lambda_{\text{app}}=\text{P}(F_u\mid H_{\text{App}})$ 表示 App 产生 F_u 的概率。如果捕获到 F_u 的前 r 个完整分组，则 H_{App} 的状态转移序列为 $S=s_1s_2s_3\cdots s_r$，因此

$$\lambda_{\text{app}}=\text{P}(F_u\mid H_{\text{App}})=\text{P}(F_u\mid S,H_{\text{App}})=\prod_{i=1}^{r}\text{P}(g_i\mid s_i,H_{\text{App}})$$

若未能完整捕获 F_u 中前 r 个分组，则与分组序列对应的状态序列 S 未知，此时，将所有可能状态序列 S 产生 F_u 的概率相加，即可得到 λ_{app}，则有：

$$\lambda_{\text{app}}=\sum_{S}\text{P}(F_u\mid S,H_{\text{App}})\text{P}(S\mid H_{\text{App}})$$

但上式计算复杂度过高，且随着状态个数的增加成指数增长，不适合实时在线处理，可使用向前—向后算法来求解 λ_{app}。

定义前向变量 $\alpha_k(j)$ 为

$$\alpha_k(j)=\text{P}(g_1g_2\cdots g_j,s_i\mid H_{\text{App}})$$

即对于 H_{App}，出现分组序列 $g_1g_2\cdots g_j$ 以及出现分组 g_j 时状态正好处于 s_i 的概率，如下建立 $\alpha_k(i)$ 满足的递归关系：初始阶段，观察到 F_u 第一个分组 $g_1=+\left(b(1)_{f_1},d(1)_{h_1}\right)^{\text{T}}$，$b(1)_{f_1}$ 表示 PL 落入区间 f_1，$d(1)_{h_1}$ 表示 IAT 落入区间 h_1，则有：

$$\alpha_1(i)=\pi_i\ pl_i(f_1)\ iat_i(h_1),\quad i=1,2$$

当观察到 F_u 的第 j+1 个分组后，

$$\alpha_{j+1}(b)=\left[\sum_{i=1}^{r}\alpha_j(i)a_{ib}\right]pl_b(f_{j+1})iat_b(h_{j+1}),$$
$$1\leqslant b\leqslant r-1,\ 1\leqslant j\leqslant r$$

当观察到 F_u 的第 r 个分组后，即可得到：

$$\lambda_{\text{app}}=\sum_{i=1}^{r}\alpha_r(i)$$

5 P2P 流识别架构

5.1 HMM-FIA 架构描述

通常，多种 P2P 应用流混合在一起，要准确识别每一条流所属应用，需要为每种待识别应用构造独立的 HMM，假定有 n 种 P2P 应用 $\{\text{App}_1,\text{App}_2,\cdots,\text{App}_n\}$，每种应用对应的识别模型表示为 $\{M_1,M_2,\cdots,M_n\}$。本文提出了利用多个 HMM 识别多种 P2P 应用流的架构 HMM-FIA (HMM based P2P flow identification architecture)，如图 3 所示。

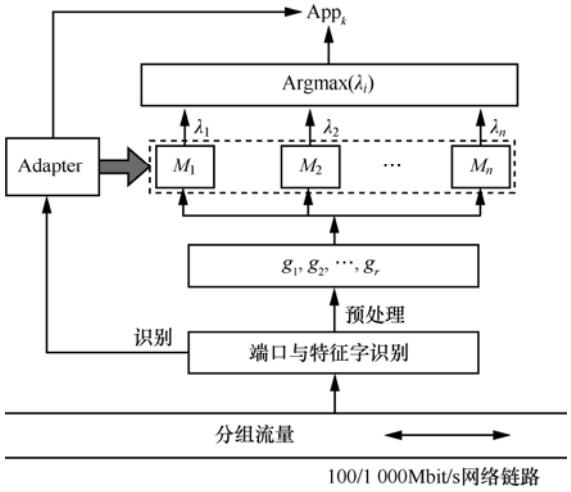


图 3 基于 HMM 的 P2P 流识别架构

文献[19]表明，因特网中非 P2P 流量平均占因特网总流量的 30%以上，其中大部采用固定端口通信，可使用端口号识别其所属应用，BitTorrent 和 eDonkey 流量占 P2P 总流量的 80%以上，未加密比例分别超过 73%和 83%，可使用特征字方法有效识别。因此，在 HMM-FIA 中设置端口与特征字识别模块，能快速识别多种非 P2P 流和未加密的 P2P 流。HMM-FIA 采用被动方式捕获链路中的分组，首先交给端口和特征字识别模块处理，若未能识别，则

将流中前 r 个分组预处理结果 $F_u = g_1 g_2 \cdots g_r$ 提交给 n 个并行的 HMM，分别计算 M_i 产生 F_u 的概率 λ_i $(1 \leqslant i \leqslant n)$，并将 F_u 归类为产生最大概率值的 HMM 所对应的应用。为适应流特征随网络环境的不同而有所变化，在 HMM-FIA 中设置适配器 Adapter，若端口与特征字识别模块能够识别分组流，则由 Adapter 更新对应 HMM 的参数 $\boldsymbol{B}^{pl}$ 和 $\boldsymbol{B}^{iat}$。实验结果表明，采用 Adapter，能使 HMM-FIA 适应新的网络环境，有效提高 HMM 识别未知流的准确性。

5.2　状态个数

在 HMM-FIA 中，为了同时识别多种 P2P 应用，需要多个特定的 HMM，表示为 $M = \{M_1, M_2, \cdots, M_n\}$。若 2 个模型之间具有较高的相似度，则会出现误判。在 HMM 中，状态与特定分组相对应，状态个数决定了识别分组流的实时性、复杂性和准确性。本文提出了通过增加 HMM 状态个数，降低模型之间相似度的算法，称为状态个数选择算法(SNSA, state number select algorithm)。

5.2.1　相关定义

定义 1　平行状态对。给定 2 个 HMM，M_i 和 M_j，在流序列 $F_u = g_1 g_2 \cdots g_r$ 中，任意分组 $g_k \in F_u$ 对应 2 个模型的状态 (s_{ik}, s_{jk}) 称之为平行状态对。

定义 2　状态相似度 $p_{ik,jk}$。给定平行状态对 (s_{ik}, s_{jk})，PL 区间集合分别表示为 $\{[b', b'']\}_{ik}^{PL}$ 和 $\{[b', b'']\}_{jk}^{PL}$，它们对应的相交区间组成的集合表示为 $\{[b', b'']\}_{ik \cap jk}^{PL}$。定义状态 s_{ik} 与 s_{jk} 的 PL 相似度为 $p_{ik,jk}^{PL} = \max\{P(PL \in [b', b'']), [b', b''] \in \{[b', b'']\}_{ik \cap jk}^{PL}\}$，即 PL 落入状态 s_{ik} 和 s_{jk} 的相交区间诸概率中的最大值，如图 4 所示。同理，可定义状态 s_{ik} 与 s_{jk} 的 IAT 相似度 $p_{ik,jk}^{IAT}$。进而定义状态 s_{ik} 与 s_{jk} 的相似度为

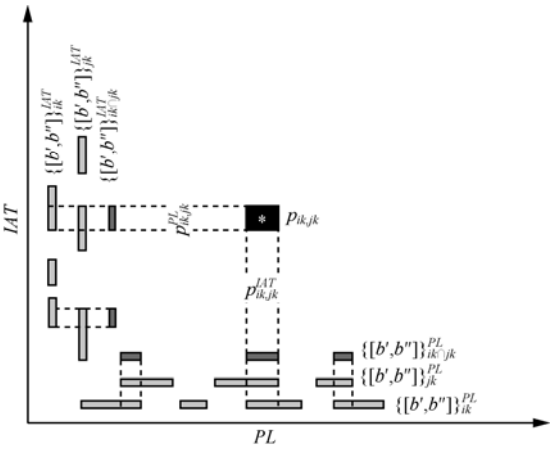


图 4　状态相似度

$$p_{ik,jk} = p_{ik,jk}^{PL} p_{ik,jk}^{IAT}$$

定义 3　HMM 相似度 $p_{i,j}$。给定 2 个模型 M_i 和 M_j，它们的平行状态对集合为 $\{(s_{i1}, s_{j1}), \cdots (s_{ir}, s_{jr})\}$，则 M_i 与 M_j 的相似度为

$$p_{i,j} = \prod_{k=1}^{r} p_{ik,jk}$$

其中，$p_{i,j}$ 反映了模型 M_i 与 M_j 之间的误判率。

5.2.2　状态个数选择算法

SNSA 的基本思想如下。

1) 设置状态个数范围 $[\alpha_{\min}, \alpha_{\max}]$，初始阶段每个模型的状态数量为 $\alpha_{\min}$，保证模型能够反映特定应用的流特征，并设置 HMM 相似度上限为 $p_{\max}$。

2) 若 M 中存在模型 M_i 与 M_j，它们的相似度 $p_{i,j}$ 大于 $p_{\max}$，若 M_i 的状态数量小于 $\alpha_{\max}$，则增加 1，若 M_j 的状态数量小于 $\alpha_{\max}$，则增加 1，并重新计算 M_i 与 M_j 的相似度。

3) 重复步骤 2)，直到 M 中各模型满足如下条件：$\forall i \forall j (M_i, M_j \in M \rightarrow p_{i,j} < p_{\max} \vee |M_i| = |M_j| = \alpha_{\max})$ 即 M 中任意 2 个模型之间的相似度小于 $p_{\max}$，或者状态数量均达到最大值。

SNSA 可描述如下：

//给定 HMM 集合 $M = \{M_1, M_2, \cdots, M_n\}$

//设置参数 $p_{\max}$，$\alpha_{\min}$，$\alpha_{\max}$

1)　foreach M_i in M do

2)　create(M_i, $\alpha_{\min}$);

//create 用来计算模型 M_i 的参数，其状态个数为 $\alpha_{\min}$

3) for i=0 to n do
4) for j=0 to n do
5) if $i \neq j$ do
6) while $p_{i,j} > p_{\max}$ do
7) if $|M_i| < \alpha_{\max}$ do create(M_i,$|M_i|+1$);
8) if$|M_j| < \alpha_{\max}$ do create(M_j,$|M_j|+1$);
9) if $|M_i| = \alpha_{\max}$ and $|M_j| = \alpha_{\max}$ do break;

上述算法在满足准确性(相似度 $p_{\max}$)的条件下，最小化了每个模型的状态数量，使模型之间的误判率小于 $p_{\max}$。算法的时间复杂度在最好情况下为 n^2，最坏情况下为 $n^2(\alpha_{\max}-\alpha_{\min})$，且模型建立阶段采用离线方式，不会影响识别效率。

6 实验及结论

本节选择 5 种流行的 P2P 应用，包括 eMule、BitTorrent、PPLive、SopCast 和 Skype 进行实验，分别为它们建立 HMM 以识别其分组流，并与已有识别方法进行比较。为描述方便，将这些 HMM 分别表示为 M_{eMule}、M_{BT}、M_{PPLive}、M_{SOP} 和 M_{Skype}。

6.1 实验环境及方法

实验环境如图 5 所示，校园网通过 100Mbit/s 光缆与 CERNET 连接，测试主机的 CPU 为 Intel Core2，主频 2.33GHz，内存大小为 2GB。在校园网内设置 eMule、SopCast、Skype、PPLive 和 BitTorrent 等 P2P 应用客户端，在测试阶段随机运行，人工记录运行信息，如通信开始时间、通信次数、下载文件大小等。设置交换机镜像端口，运行 Tcpdump 捕获分组，得到包括前 50byte 应用层数据在内的分组信息。为了比较分析 HMM 的识别能力，使用人工结合 Tcpdump 捕获的分组信息和 P2P 客户端运行信息，基于端口、特征字和 P2P 流特征，以离线方式识别分组流所属应用，并假定人工分析的结果是正确的。

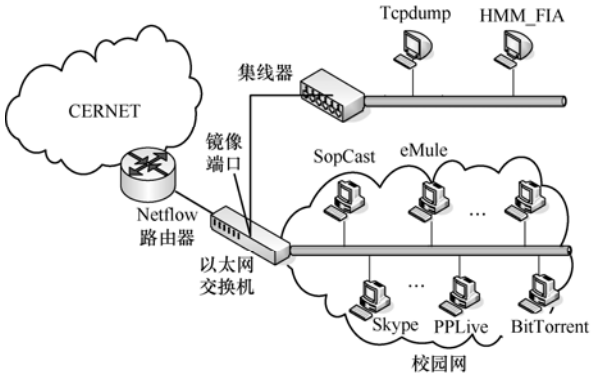


图 5　实验环境

6.2 创建 HMM

在校园网中采集样本流并进行人工识别，将已识别的流分为训练样本和离线测试样本，其统计信息如表 1 所示。训练阶段，设置 $p_{\max}$=0.06，$\alpha_{\min}$=6，$\alpha_{\max}$=20，ε_{PL}=10，ε_{IAT}=15，同时，训练样本中丢弃了分组个数小于 $\alpha_{\max}$ 的流。

P2P 应用通常同时采用 TCP 和 UDP 2 种协议，训练样本中既包含 TCP 流，又包含 UDP 流。分别采用混合与分离 2 种方式为 5 种应用建立 HMM，使用 SNSA 计算每个 HMM 状态数量，结果如表 2 所示。

表 1　样本统计信息

P2P 应用	训练集		测试集	
	流	字节	流	字节
eMule	1 580	598M	1 985	780M
BitTorrent	2 210	947M	1 412	582M
PPLive	2 820	792M	2 102	620M
SopCast	2 190	522M	1 308	348M
Skype	493	365M	243	185M

表 2　HMM 状态数量

模型	混合	分离	
		UDP	TCP
M_{eMule}	16	8	10
M_{BT}	16	8	10
M_{PPLive}	14	10	8
M_{SOP}	14	10	8
M_{Skype}	14	8	10

可见，混合 TCP 流与 UDP 流，使得 HMM 中每个状态特征区间增大，导致状态数量增多。如果分别为 TCP 流与 UDP 流建立独立的识别模型，将有效减少 HMM 的状态数量，提高识别效率。在

SNSA 中，HMM 的状态个数与 $p_{\max}$ 密切相关，表 3 显示了 $p_{\max}$ 取不同值时，TCP 流对应的 HMM 状态个数以及时间开销。

表 3　HMM 状态个数随 $p_{\max}$ 的变化情况

模型	$p_{\max}$						
	0.2	0.1	0.08	0.06	0.04	0.02	0.01
M_{eMule}	6	8	8	10	14	17	20
M_{BT}	6	8	8	10	14	17	20
M_{PPLive}	6	7	8	8	10	14	18
M_{SOP}	6	7	8	8	12	14	18
M_{Skype}	6	8	8	10	12	17	20
时间开销/s	28.47	33.65	35.14	41.73	53.92	78.01	103.90

可见，随着识别准确率的提高，HMM 的状态数量迅速增加，同时，SNSA 的时间开销也呈线性增长。

利用上述模型离线识别测试样本流，考查 SNSA 通过增加 HMM 的状态数量，控制识别多种不同 P2P 流的准确率，结果如表 4 所示。每个单元格中的数值表示第一列中的应用被识别为第一行中应用的比例，可以看到，HMM 具有较高的识别准确率，能有效区分不用协议产生的分组流。

表 4　测试样本流识别结果

	eMule	BitTorrent	PPLive	SopCast	Skype
eMule	95.0%	3.1%	0.2%	0.1%	1.6%
BitTorrent	3.5%	94.9%	0.3%	0.1%	1.2%
PPLive	0.3%	0.5%	94.3%	4.5%	0.3%
SopCast	0.2%	0.1%	5.3%	94.2%	0.2%
Skype	2.4%	1.4%	0.4%	0.2%	95.6%

在校园网中进行多次实验，使用 Tcpdump 记录完整的分组信息，并保存在流文件中，下面取 2 组不同时间段进行的典型实验数据进行分析，人工分析实验数据的结果如表 5 所示，其中“其他”包括 HTTP、FTP、DNS、Flash、FTP、ICMP、IMAP、MSN、POP、QQ、SMTP 等协议和未知流量，作为干扰数据，其中以 HTTP 和 FTP 为主，约占“其他”流量的 74%以上。为了考查 HMM-FIA 的准确性与运行时间的关系，以及 Adapter 对识别准确率的影响，分别采用 4 种不同 HMM 系统识别 P2P 流：1)PL-HMM，采用文献[18]提出的方法为每个应用协议构建 HMM，训练样本与 HMM-FIA 相同，与本文不同之处在于，PL-HMM 采用连续型随机变量刻画状态特征，并且采用独立的分组流作为训练样本；2)HMM-SIM，取消了 HMM-FIA 中的端口与特征字识别模块和 Adapter，考查网络环境的变化对 HMM 的影响；3)HMM-FIAN，不使用 Adapter，在识别过程中保持状态参数不变；4)HMM-FIA，采用 Adapter 在识别过程中调整状态参数。此外，为减少误判率，设置参数 ε_{err}=0.1，若流 F_u 对应的识别结果 λ 小于 ε_{err}，则将 F_u 视为未知流，使用 4 个 HMM 系统识别 T1 的结果如表 6 所示。可以看出，HMM-FIAN 的假阳性与假阴性均低于 PL-HMM，说明 HMM-FIAN 能够更好地保留协议特征，提高识别准确率。在识别结果中，假阳性普遍较高，这是由于网络中存在多种非 P2P 应用流，且它们的特征与已有 P2P 流相近所导致。HMM-FIAN 采用端

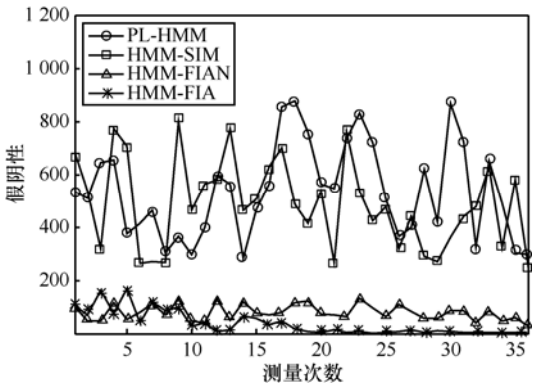
表 5　Tcpdump 捕获校园网流量统计信息

流文件	开始时间	周期	流							字节						
			合计	eMule	SopCast	Skype	PPLive	BT	其他	合计	eMule	SopCast	Skype	PPLive	BT	其他
T1	May 16 09:00	3 h	130.6K	22.1K	8.7K	0.41K	15.3K	50.1K	33.99K	47.3G	8.37G	2.07G	0.32G	4.2G	19.24G	13.1G
T2	May 15 18:00	2 h	128.4K	19K	9.5K	0.43K	16.1K	53.5K	29.87K	45.04G	7.19G	2.27G	0.36G	4.41G	22.8G	8.01G

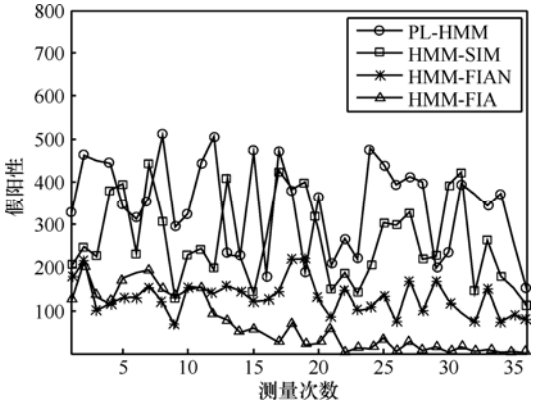
表 6　不同 HMM 系统识别 T1 的结果

系统	eMule			SopCast			Skype			PPLive			BT			时间/s
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN	
PL-HMM	15.7K	7.1K	6.4K	6.53K	2.9K	2.17K	290	2.7K	120	10.8K	6.8K	4.5K	36.57K	19.3K	12.54K	572
HMM-SIM	16.96K	6.2K	5.14K	7.16K	2.45K	1.54K	354	2.2K	56	12.2K	5.7K	3.1K	40.59K	17.6K	9.51K	346
HMM-FIAN	19.74K	1.4K	2.36K	7.9K	1.2K	0.8K	354	0.5K	56	13.85K	0.9K	1.45K	45.4K	2.9K	4.7K	228
HMM-FIA	17.03K	0.8K	1.2K	6.57K	0.3K	0.4K	295	0.2K	23	11.25K	0.6K	0.67K	38.28K	1.4K	2.9K	247

口和特征字识别模块，能有效识别具有端口特征的非 P2P 流和具有负载特征的 P2P 流，减少了假阳性与假阴性的比例。HMM-FIA 启用 Adapter，根据特征字识别结果更新 HMM 参数，进一步提高了识别准确率。这是因为，在不同的网络环境中，随着 P2P 客户端软件的版本和类型的变化，分组大小会有所改变，另外，分组的到达时间间隔也会因排队时延和处理时延有所变化。因此，在 HMM-FIA 中使用 Adapter 能较好地适应网络环境变化，提高识别准确率。从表 6 还可以看到，HMM-SIM 的时间开销远小于 PL-HMM，说明使用离散型随机变量描述状态特征，并将 HMM 的隐状态序列与分组到达序列相对应，能有效提高识别效率，减少时间开销。HMM-FIAN 与 HMM-FIA 采用了端口与特征字识别方法，进一步减少了 HMM 处理分组流的数量，节省了识别时间。进一步统计了 4 个系统每 5min 识别 BitTorrent 流发生假阳性与假阴性的次数，结果如图 6 所示。可见使用端口与特征字识别模块，能够有效减少假阳性与假阴性数量，此外在 HMM-FIA 中假阳性与假阴性数量随着系统运行不断减少，说明适配器 Adapter 根据识别结果不断更新 HMM 参数，有助于 HMM 适应被测试网络环境，有效提高识别准确率。



(a) 假阴性随时间变化趋势



(b) 假阳性随时间变化趋势

图 6　假阳性与假阴性随时间变化情况

7　结束语

本文从协议有限状态机角度出发，提出了基于 HMM 的 P2P 流识别技术，采用分组大小、到达时间间隔和到达顺序作为构建 HMM 的特征，并详细介绍了模型的建立和识别过程。为了同时识别多种 P2P 应用流，提出了基于 HMM 的 P2P 流识别架构 HMM-FIA，并设计了状态个数选择算法。实验结果说明了本文提出的方法能准确、高效地识别多种 P2P 应用流，并且能较好地适应网络环境变化。

下一步工作将研究多种识别方法与 HMM-FIA 的融合，进一步提高 HMM-FIA 在高速网络环境下的运行效率。

参考文献：

[1] SEN S, WANG J. Analyzing peer-to-peer traffic across large networks[A]. Proceedings of ACM SIGCOMM Internet Measurement Workshop[C]. Marseilles, France, 2002.

[2] MOORE D, KEYS K, KOGA R, *et al*. The CoralReef software suite as a tool for system and network administrators[A]. Proceedings of the 15th USENIX Conference on Systems Administration[C]. San Diego: USENIX Association, 2001.

[3] KARAGIANNIS T, BROIDO A, BROWNLEE N, *et al*. Is P2P dying or just hiding[A]. Proceedings of the IEEE Globecom 2004[C]. Dallas, Texas, 2004.

[4] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network applications[A]. Proceedings of the 6th Passive and Active Measurement Workshop[C]. Berlin, 2005.

[5] SEN S, SPATSCHECK O, WANG D. Accurate, scalable in-network identification of P2P traffic using application signatures[A]. The 13th

International Conference on World Wide Web[C]. New York, 2004.

[6] LI W, 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.

[7] CHEN L, GONG J. Analyzing the characteristics of application traffic behavior based on chi-square statistics[J]. Journal of Software, 2010, 21(11): 2852-2865.

[8] MENA A, HEIDEMANN J. An empirical study of real audio traffic[A]. Proceedings of IEEE INFOCOM 2000[C]. Israel, 2000.

[9] MCGREGOR A, HALL M, LORIER P, *et al*. Flow clustering using machine learning techniques[J]. Lecture Notes in Computer Science, 2004, 3015: 205-214.

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

[11] ZANDER S, NGUYEN T, ARMITAGE G. Self-learning IP traffic classification based on statistical flow characteristics[A]. Proceedings of the 6th International Workshop on Passive and Active Network Measurement[C]. Boston, Masschusetts: Springer-Verlag Berlin, 2005.

[12] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[A]. Proceedings of the IEEE 30th Conference on Local Computer Networks[C]. Washington, DC, 2005.

[13] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms[A]. Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data[C]. Pisa, Italy, 2006.

[14] BERNAILLE L, TEIXEIRA R, SALAMATIAN K. Early application identification[A]. Proceedings of the 2006 ACM CoNEXT conference[C]. Lisboa, Portugal, 2006.

[15] CROTTI M, DUSI M, GRINGOLI F, *et al*. Traffic classification through simple statistical fingerprinting[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(1): 5-16.

[16] WRIGHT C, MONROSE F, MASSON G. HMM profiles for network traffic classification[A]. Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security[C]. New York, 2004.

[17] WRIGHT C, MONROSE F, MASSON G. Towards Better Protocol Identification Using Profile HMMs[R]. Technical Report JHU-SPAR051201, Johns Hopkins University, 2005.

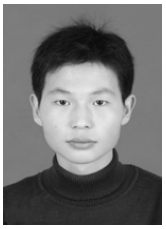
[18] ALBERTO D, WALTER D, ANTONIO P, *et al*. Classification of network traffic via packet-level hidden Markov models[A]. GLOBECOM 2008[C]. New Orleans, 2008.

[19] SCHULZE H, MOCHALSKI K. Internet study 2008/2009[EB/OL]. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009.

作者简介：

许博（1980-），男，甘肃兰州人，解放军理工大学讲师，主要研究方向为网络分布式计算和 P2P 流量识别等。

陈鸣（1956-），男，江苏无锡人，解放军理工大学教授、博士生导师，主要研究方向为网络测量、网络体系结构、网络管理等。



魏祥麟（1985-），男，安徽砀山人，解放军理工大学博士生，主要研究方向为网络测量、流量异常检测、对等网络性能优化等。

（上接第 54 页）

YI X P, TANG Y X, HAO S H. Optimal linear detection algorithm of MIMO with distributed transmit antennas[J]. Journal of Electronics. 2009, 37(12):2694-2699.

[8] 张贤达. 矩阵分析与应用[M]. 第 1 版.北京:清华大学出版社, 2004.102-104.
ZHANG X D. Matrix Analysis and Applications[M]. The first edition. Beijing:Tsinghua University Press, 2004.102-104.

作者简介：

窦冬冬（1986-），男，河南商丘人，解放军信息工程大学研究生，主要研究方向为分布式 MIMO 信号检测技术。

刘军博（1984-），男，江苏连云港人，解放军信息工程大学研究生，主要研究方向为虚拟 MIMO 资源管理技术。

王大鸣（1971-），男，辽宁大连人，博士，解放军信息工程大学副教授，主要研究方向为无线移动通信。

李兆训（1968-），男，山东沂南人，硕士，解放军信息工程大学副教授，主要研究方向为无线移动通信。