doi:10.3969/j.issn.1000-436x.2014.07.020

未知网络应用流量的自动提取方法

王变琴[1]，余顺争[2]
（1. 中山大学 东校区教学实验中心，广东 广州 510006；2. 中山大学 信息科学与技术学院，广东 广州 510006）

摘 要：提取未知网络应用特征时需要获得其流量数据，但在网络工程中，采集的未知应用流量往往是几种应用流量的混合，如何将未知混合流量进行分离，按照应用进行归类是现有方法没有解决的问题。基于此提出一种基于载荷信息的流量聚类方法，该方法通过对报文载荷的部分字节编码，采用扩展的 ROCK 算法对未知混合流量进行分离，按照不同应用进行归类。实验结果表明，与基于会话行为特征（一种流量统计特征）的流量聚类方法相比，这种方法具有较高的精确度。

关键词：流量分类；会话行为特征；载荷；ROCK 算法
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)07-0164-08

Automatic extraction for the traffic of unknown network applications

WANG Bian-qin[1], YU Shun-zheng [2]
(1. Education & Experiment Center, East Campus, Sun Yat-sen University, Guangzhou 510006, China;
2. School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China)

Abstract: The features of unknown network applications can be extracted using its traffic data. However, the sample traffic in network engineering is usually a mixed traffic generated by several unknown applications. The separation of the mixed traffic by applications an unsolved problem presently. A clustering method for traffic classification was proposed based on payload information. The proposed method can firstly encode certain bytes of message payload, then separate and classify the unknown mixed traffic using an extended ROCK algorithm. The experiment results reveal that compared with the clustering method based on statistics character of traffic, the proposed method has higher accuracy.

Key words: traffic classification; behavioral features of session; payload; ROCK algorithm

1 引言

网络流量的准确识别和分类是提供差异性服务质量（QoS, quality of service）保障、入侵检测（intrusion detection）、流量监控（traffic monitoring）及计费管理（accounting）等方面的基础。然而，面对网络应用的快速发展，传统的端口应用识别法逐渐失效，基于流（flows）统计特征的流量分类方法[1,2]不能进行应用的精确识别，利用载荷特征（signature）的应用识别方法[3,4]简单、准确度较高，在实际中被广泛应用，然而如何自动获取载荷特征是这种方法面临的主要问题，目前已有一些研究成果[5~11]。Byung-Chul 等人[5]提出一种基于 LCS（longest common substring）的载荷特征提取方法——LASER；Wang 等人[6]提出利用序列比对方法提取载荷特征；Ye 等人[7]提出一种基于子树（subtree）结构的特征生成系统——AutoSig；赵咏等人[8]提出一种基于语义信息的协议特征自动发现方法——TPCAD；刘彬等人利用关联规则算法提取载荷特征。所有这些方法所需要的数据源都是同种应用协议的流量。获取已知应用的流量较为容易，可以通过特定的网络应用进程生成或从混合流量

收稿日期：2013-03-07；**修回日期**：2013-05-25
基金项目：国家自然科学基金资助项目(61202271)；广东省自然科学基金资助项目(S2012040007184)；国家自然科学基金-广东联合基金资助项目(U0735002)
Foundation Items: The National Natural Science Foundation of China (61202271); The Natural Science Foundation of Guangdong Province (S2012040007184); The Key Program of NSFC-Guangdong Joint Funds (U0735002)

中获取（对于端口固定的应用，利用端口从中提取其流量；对于非固定端口的应用，则可以利用支持这种应用流量识别的系统或工具提取其流量）。如果要将这些载荷特征自动提取方法扩展到未知网络应用，则需要获得单一未知应用的流量，然而在实际中获取的未知流量通常是一种或几种未知应用协议流量的混合。因此，首先需要将混合流量进行分离、归类，使得每类中的流量属于同一种应用，目前这方面的研究极少。Zhang 等人[12]提出利用分组的五元组信息对会话进行聚类，但由于五元组和应用类别没有直接关系，使得这种方法不太可靠；在协议逆向工程（protocol reverse engineering）研究领域，学者利用聚类方法[13,14]对报文进行归类，用于同类报文格式的提取，这些研究还只是一些初步探索。

本文利用聚类方法对未知混合应用流量进行分离，首先选择会话行为特征（一种流量统计特征）对混合流量进行聚类。实验评估发现，在不同应用会话行为特征相似情况下，这种方法不能有效分离不同应用流量。为此提出一种基于载荷信息的流量聚类方法，该方法通过对载荷中的字节进行编码，利用不同于传统距离函数的 Jaccard 系数作为相似性度量对流量进行聚类。实验评估表明，这种基于载荷信息的方法比利用会话行为特征的流量聚类方法更为有效。

2　相关研究

利用聚类方法对流量进行聚类的研究已有将近十年历史，目前研究主要是基于流量的各种统计特征的聚类，期望实现在线流量分类或识别[15~18]。Mcgregor 等人[15]利用流的统计特征和 EM（expectation maximization）算法将流聚成不同应用类型（Bulk transfer, single and multiple transactions and interactive traffic），这是聚类方法在本领域的最早尝试。Zander 等人[16]利用顺序前向选择（SFS, sequential forward selection）策略对流的统计特征进行优化，利用 AutoClass 算法（EM 算法的一种拓展，通过反复使用 EM 算法以便找到全局最优解）识别应用（FTP、Telnet、SMTP、DNS、HTTP、AOL Messenger、Napster、Half-Life）。Erman1 等人[17]则比较了 3 种聚类算法在流量分类时的性能，评估结果显示 *K*-means 和 DBSCAN 算法的性能均好于 AutoClass 算法，其中 DBSCAN 算法聚类效果最好。董仕等人[18]利用端口统计特征的差异对 P2P 流量进行聚类。

与现有方法不同，本文的目标是实现对未知混合流量的分离，以期获得同种应用流量的聚类，用于载荷特征自动提取前的数据预处理。这种方法的一般框架如图 1 所示，首先采集未知应用流量，这种流量由网络中部署的流量识别系统（例如 L7-filter）不能识别应用生成的混合流量，然后通过聚类分析获得同种未知应用流量，在此基础上，完成未知应用协议特征的自动提取。

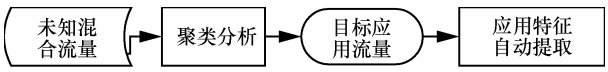


图 1　未知应用特征提取框架

通常对未知混合流量没有先验知识，只能根据流量自身内在特点、规律以及相似性原则对其进行归类。聚类分析（cluster analysis）是处理这类问题的一种有效方法，它根据“物以类聚”的规律，将数据集划分为若干簇（cluster），这些簇是事先未知的，其形成完全由数据驱动。本文进行聚类的对象是会话（与连接或双向流的概念相同），因其不仅包含通信双方在 2 个方向上独立的网络流特征，还包含 2 个单向流之间的关联特征，具有更强的特征描述能力。未知混合流量可以被处理成由会话组成的集合，据此本文研究的问题描述为：给定一个会话集 $S=\{s_1,\cdots,s_n\}$，定义一个映射 $f:S\to C$，其中第 i 个会话被映射到第 j 个簇 C_j 中，C_j 由所有被映射到该簇中的会话组成，即 $C_j=\{s_i \mid f(s_i)=C_j, 1\leqslant j\leqslant k, \text{且} s_i\in S\}$，每个簇 $C_j(1\leqslant j\leqslant k)$ 代表一种未知应用流量，问题的求解视为产生一个簇集 $C=\{C_1,\cdots,C_k\}$ 的过程。

聚类方法的性能与特征选择以及聚类算法有关。本文根据未知混合流量的特点，分别选择统计特征和载荷信息进行聚类，并通过实验评估不同聚类算法的性能。

3　基于会话行为特征的未知流量聚类方法

现有流量聚类方法大都是基于分类对象的统计特征。会话行为特征是指会话内报文（不含载荷）的各种统计特征。Moor 等人[19]对可能存在的多种候选特征进行了深入研究。在实际应用中，特征选择具有挑战性，需要结合实际情况。考虑到通常获

得的未知流量是流量识别引擎无法识别的流量，只是实际聚合流量的部分数据，已经不再具有聚合流量的特征。因此关于多个会话之间的统计特征不再可靠。此外，实际中的流量识别系统，例如，L7-Filter，一般只保留会话的部分数据，而非完整会话数据。Bernaille1[20,21]等人提出利用会话的前 N 个分组长度和方向作为特征具有一定的根据，因为许多基于 TCP 协议的应用层协议（少数除外，例如 HTTP 协议）在完成 3 次握手后，进入应用层协议的协商过程，例如，SSL 协议密钥协商过程，FTP 协议认证密码过程等，这些不同协议有不同的协商过程，并且协商过程是双向的，因此还需要区分报文的不同方向。Bernaille1 等人提出 2 种特征选择方案：文献[20]中根据观察会话内报文的大小分布，把报文大小（字节数）映射为 4 个级别｛1, 2, 3, 4｝（其中，{1 =[0, 150], 2 =[150, 700], 3 =[700, 1300], 4 =[1300, 1500]}），并结合分组方向（利用符号表示分组方向：“+”代表客户端报文，“−”表示服务器端报文）将会话表示成 $\Sigma=\{\pm1,\pm2,\pm3,\pm4\}$ 上的一个序列，每个会话表示为 $S_i=(\upsilon_1,\cdots,\upsilon_{l_i})$，其中，$l_i$ 为序列长，$\upsilon_t\in\Sigma$ 代表序列的第 t^{th} 个分组，生成所有会话的集合 $S=\{s_i\}$，然后利用变换生成的序列进行聚类，本文称这种特征选择方案为会话映射策略；另一种特征选择方案[21]则将每个会话的前 N 个报文的精确长度与方向作为特征。上述这 2 种方案均符合未知混合流量特点，本文通过实验评估基于这 2 种特征的不同算法 k-means[22]、EM[23]和 DBSCAN[24]对未知混合流量的聚类。评估结果发现，当不同应用的会话行为特征比较相似时（例如 FTP、POP3 协议），基于会话行为特征的聚类方法则不能有效区分它们。

4 基于载荷信息的未知流量聚类方法

据研究，会话的前 m 个报文中包含载荷特征，并且多分布在报文开始和结束的固定位置处。特征由不同字节序列组成，通过对字节编码，可以将其作为不同类别字符，不同协议会话报文载荷则包含不同字符组合。假设选取每个会话的前 m 个报文，每个报文的前 n 个字节，然后按照偏移对报文中的字节进行编码处理：规定报文首字节偏移为零，则该字节标注为 b_0，偏移为 i 的字节标注为 b_i，编码后的字节当成不同类别字符对待，这样会话 s 可以表示为 $s=m_1m_2m_3,\cdots,m_m$，其中 $m_i=b_0b_1b_2\cdots b_n$，通过度量会话间的相似度来区分不同协议的会话。

Guha 等人[25]提出一种面向类别数据的聚类算法——ROCK（robust clustering algorithm for categorical attributes using link），其突出贡献是采用基于全局信息的公共邻居（链接）作为评价数据对象间的相似性（similarity）度量标准，受此启发，提出一种基于载荷信息的 ROCK 流量聚类方法，引入的概念如下。

定义 1 2 个会话之间的相似度由 Jaccard 系数（coefficient）定义，令 s_i,s_j 为 2 个会话，则

$$\text{sim}(s_i,s_j)=\frac{s_i\cap s_j}{s_i\cup s_j},\quad 0\leqslant\text{sim}(s_i,s_j)\leqslant1 \tag{1}$$

当函数 $\text{sim}(s_i,s_j)=0$ 时，表明 2 个会话完全不同；$\text{sim}(s_i,s_j)$越大，表明会话 s_i 和 s_j 越相似，当 $\text{sim}(s_i,s_j)=1$ 时，表明 2 个会话相同。

例如，会话 s_1=“220-HeUSER a331 UsPASS a”，会话 s_2=“220 WeUSER a331 UsPASS b”，对应位置相同的字符有 21 个，则 $\text{sim}(s_1,s_2)=21/(21+2\cdot(24-21))=0.778$。

定义 2 设 θ 为给定相似度阈值，称 s_i,s_j 互为邻居，当且仅当

$$\text{sim}(s_i,s_j)\geqslant\theta,\quad 0\leqslant\theta\leqslant1 \tag{2}$$

定义 3 令 s_i,s_j 为 2 个会话，称 s_i,s_j 的公共邻居数目为其链接，即

$$\text{link}(s_i,s_j)=N_i\cap N_j \tag{3}$$

其中，N_i 和 N_j 分别为 s_i,s_j 的邻居表。

定义 4 2 个簇间的链接定义为

$$\text{link}(C_i,C_j)=\sum_{s_q\in C_i,s_r\in C_j}\text{link}(s_q,s_r) \tag{4}$$

定义 5 指导合并的 Goodness 函数定义为

$$g(C_i,C_j)=\frac{\text{link}(C_i,C_j)}{(n_i+n_j)^{1+2f(\theta)}-n_i^{1+2f(\theta)}-n_j^{1+2f(\theta)}} \tag{5}$$

其中，分子是 C_i、C_j 之间的实际链接数目。

函数 $\text{link}(C_i,C_j)$是 2 个簇之间的链接数目，n_i 和 n_j 分别是 2 个簇中的点数目。大簇有较多链接，用分母来归一化链接数。当相似性度量的阈值为 θ 时，在 C_i 中点对之间链接数的估计值为 $n_i^{1+2f(\theta)}$。函数 $f(\theta)$ 的具体形式取决于数据，但它满足一条性质，即在簇 C_i 中的每一个成员大约有 $n_i^{f(\theta)}$ 个邻居。显然，如果簇中的所有点相互邻接，则 $f(\theta)=1$，

进而得出簇 C_i 中的点对之间的连接数 n_i^3。实验中，选择 $f(\theta)=(1-\theta)/(1+\theta)$。

ROCK 算法的核心思想为：给定会话集 S、期望得到的簇数 k，计算任意两点间的相似度，进而求出两点间的链接数，每次合并 g 值最大的 2 个簇集，直到达到给定的簇数 k 为止。与距离函数相比，链接不全局信息的度量，其过程描述见算法 1。

算法 1　procedure cluster(S, k)

输入：

S　// $S=\{s_1,s_2,\cdots,s_n\}$会话集

k　// 期望簇数目

输出：

C　// $C=\{C_1,\cdots,C_k\}$ 簇集

begin

1) link : = compute_links(S)　// 计算两样本间共同邻居数

2) for each $s \in S$ do

3)　$q[s]$:= build_local_heap(*link*,s)　// 初始时，对每个簇创建局部堆 $q[s]$

4)　Q := build_global_heap(S,q)　// 创建全局堆 Q

5) while size(Q) > k do {

6)　u := extract_max(Q)　// 抽取 Q 中 $g(u,\max(q[u]))$最大的簇 u

7)　v := max($q[u]$)　// 抽取 u 的局部堆 $q[u]$中 $g(u, v)$最大的簇 v

8)　delete(Q,v)

9)　w := merge(u,v)　// 合并簇 u 和簇 v 为簇 w

10)　for each $x \in q[u] \cup q[v]$ do { // 完成包含簇 u,v 的局部堆与创建簇 w 的局部堆

11)　link[x,w]:= link[x,u] + link[x,v]

12)　delete($q[x]$,u); delete($q[x]$,v)

13)　insert($q[x]$,w,$g(x,w)$); insert($q[w]$,x,$g(x,w)$)

14)　update(Q,x,$q[x]$)

15)　}

16)　insert(Q,w,$q[w]$)

17)　deallocate($q[u]$); deallocate($q[v]$)　// 释放簇 u 与簇 v 的局部堆 $q[u]$、$q[v]$

18) }

end

算法分析：第 1)步由函数 compute_links(S)计算 2 个样本间共同邻居 link，其过程描述如算法 2 所示；第 2)步和第 3)步创建局部堆，最初各个样本各自为一个簇，对于每个簇 i，创建对应的局部堆 $q[i]$，对于簇 j，若 link[i,j]不为 0，则 $q[i]$中包含簇 j，并且计算簇 i 与簇 j 的优良度 $g(i,j)$，按 $g(i,j)$的降序排列；第 4)步创建一个包含所有簇的全局堆 Q，计算 $g(j,\max(q[j]))$（$\max(q[j])$为与簇 j 最优合并的簇，即 $q[j]$的第一个元素），同样按 $g(j,\max(q[j]))$降序排列，全部簇加入后，此时，Q 的第一个元素簇 j 与 $q[j]$ 的第一个元素为最佳合并的簇；第 5)~18)步不断地合并簇，直到簇的数目为 k 或者剩下簇间的 link 等于 0，其中，第 6)步抽取 Q 中 $g(u,\max(q[u]))$最大的簇 u；第 7)步对应抽取 $q[u]$中 $g(u,v)$最大的簇 v；第 8)步由于簇 u 与簇 v 合并，删掉对应 Q 中的簇 u 与簇 v；第 9)步合并簇 u 与簇 v 并创建簇 w；第 10)~15)步对于每个局部堆中包含簇 u 或簇 v 的簇 x，都要在它的局部堆 $q[x]$中用 w 代替，并按优良度排序，创建簇 w 的局部堆 $q[w]$，第 16)步 $q[w]$创建完成，并将 w 插入到 Q 中；第 17)步释放簇 u 与簇 v 的局部堆 $q[u]$、$q[v]$继续循环。

算法 2　procedure compute_links(S)

begin

1) Compute nbrlist[i] for every point i in S　// 计算样本 i 相似度得其邻居列表 nbrlist[i]

2) Set link[i,j] to be zero for all i,j

3) for i := 1 to n do {

4)　N := nbrlist[i];　// N 为 i 邻居列表

5)　for j := 1 to $|N|-1$ do　// $|N|$为 i 邻居列表邻居个数

6)　for i := j+1 to $|N|$ do

7)　link[$N[j]$,N[i]] := link[$N[j]$,$N[i]$] + 1 // 邻居对 $N[j]$与 $N[i]$

8) }

end

该算法的时间复杂度和空间复杂度分别为 $O(n^2\log n)$ 和 $O(n^2)$。2 个会话之间的链接数目 link 由它们共同的邻居数目来定义，聚类算法的目标是将具有更多链接的点聚成簇，属于层次聚类方法（hierarchical method），其相似度度量采用的是链接数目。与距离度量相比，使用链接度量提供一种全局的方法，因为两点之间的相似度也受其他点的影响。

5 实验评估

分别评估基于会话行为特征和基于载荷信息的未知流量聚类方法的性能。利用工具 Wireshark 在 CERNET 某小区网络出入口处采集包含载荷（payload）数据的真实网络流量，采集时间为 2009 年 1 月 8 日~15 日，分别随机选择 3 个不同时段（每个时段持续 5 min）数据，从中提取目前流行的 7 种应用（HTTP、FTP、SMTP、POP3、SSL、MSN、BT）的会话（如表 1 所示）进行测试，它们的会话数不等，这也符合未知混合流量的实际情况，因为在采样的时间段内常常仅有某种或某几种未知应用流量的出现，各种应用的流量分布不均，可能某一种或几种流量较多，而其他则较少。每种应用的数据平均分成 3 组分别进行 3 次测试，所得实验结果是其平均值。

表 1 样本会话集

应用	Session	Packet	Size/(Mbit·s^{-1})
HTTP	880	57 568	77.2
FTP	920	20 616	102.24
SMTP	652	32 884	60.32
POP3	512	6 000 072	72.8
SSL	788	24 018	44.70
MSN	556	52 460	18.52
BT	1184	297 594	44.7

实验机为 PC 机（CPU：Intel Core 2 Duo E6550 2.33 GHz，内存：0.99 GB），其操作系统为 Windows，ROCK 算法用 C++语言实现，其余算法利用 Weak3.6.5[26]测试。

聚类效果的评价指标：漏报率（FNR, false negative rate）和误报率（FPR, false positive rate），它们可表示为

$$FNR=\frac{FN}{TP+FN}\times 100\% \tag{6}$$

$$FPR=\frac{FP}{FP+TN}\times 100\% \tag{7}$$

其中，*TP*(true positive)是被预测为正的正样本；*TN*(true negative)是被预测为负的负样本；*FP*(false positive)是被预测为正的负样本；*FN*(false negative)是被预测为负的正样本。

5.1 会话行为特征的流量聚类方法评估

一般会话的前 4 个报文正好处于应用协议的协商阶段，这个阶段的报文序列是预定义好的，不同应用之间有差异，因此，选择会话建立阶段的前 4 个报文大小和方向作为特征，分别利用 *k*-means、DBSCAN 和 EM 3 种聚类算法进行测试。

实验 1 比较 2 种会话行为特征选择方案：1）会话的精确报文大小和方向；2）会话映射策略。选择会话行为特征差异较为显著的 3 种协议的流量进行测试，其结果如表 2 所示。

从表 2 可知，对于第 1 种特征选择方案，3 种算法都能产生较好的聚类效果，相比之下，DBSCAN 算法的聚类效果最好，具有低漏报和误报，虽然它忽略某些样本，但从数据预处理的角度看，只是减少样本数，不大会影响应用特征的提取结果。第 2 种特征选择方案的聚类效果普遍比第 1 种稍差，这是由于将报文大小映射到有限的区间，屏蔽了协议报文的某些差异，使其不能精确反映协议之间的区别。

理论上，*k*-means 和 DBSCAN 算法的时间复杂度分别为 $O(n)$、$O(n^2)$，而 EM 是基于统计模型的聚类算法，其时间复杂度一般比 DBSCAN 高，实验获得的 3 种算法的耗时长短为：*k*-means (16 ms)< DBSCAN(50 ms)<EM (64 ms)，这与理论分析结果是吻合的。对于同一种算法，2 种特征选择策略的运行时间几乎一样。由于数据量不大，测出的耗时可能有误差，但也能反映 3 种算法在同一数据集上的时间差异。

表 2 2 种不同特征选择方法比较

应用协议	*FNR*(方案 1/方案 2)			*FPR*(方案 1/方案 2)		
	k-means	EM	DBSCAN	*k*-means	EM	DBSCAN
SSL	1.82/1.82	0.00/0.00	0.00/0.00	9.85/12.12	7.88/11.94	6.73/7.73
FTP	0.00/6.06	0.00/0.00	0.00/0.00	6.61/7.74	0.00/5.79	0.00/0.00
HTTP	30.0/30.3	30.0/30.3	17.02/17.02	0.00/0.00	0.00/0.00	0.00/0.00

注：参数设置：*k*-means (*k*=3)、EM (*k*=3)、DBSCAN(*Eps*=0.23, *MinPts*=6)，DBSCAN 会自动丢弃一些自认为是噪声的样本，在计算漏报率与误报率时，实际统计的是参与聚类的样本个数，而非全部样本个数。

实验 2　评估不同数据集上各种算法的性能。在实验 1 的数据集中分别混入 MSN、BT 以及 POP3 协议的流量来评估 3 种算法的聚类效果。MSN 协议比较特别，它有 3 种连接，分别是连接到 DS 服务器，获得 NS 服务地址；连接到 NS 服务器，用于身份验证、得到个人信息、上线通知，或者发送、接受聊天请求；连接到 SB 服务器，用于发送或接收聊天消息。3 种连接的会话形式不同，实验数据只取其连接到 DS 服务器的会话来做聚类处理。测试结果如表 3~表 5 所示。

表 3　增加 MSN 流量时不同聚类算法的聚类效果

应用协议	*FNR*			*FPR*		
	k-means	EM	DBSCAN	*k*-means	EM	DBSCAN
SSL	1.82	1.82	0.00	3.48	0.00	0.00
FTP	0.00	98.00	0.00	4.21	0.00	0.00
HTTP	30.30	0.00	0.00	0.53	6.23	0.00
MSN	7.25	8.70	0.00	5.35	32.62	0.00

注：参数设置：*k*-means (*k*=4)、EM (*k*=4)、DBSCAN(*Eps*=0.048, *MinPts*=6)

表 4　增加 BT 流量时不同聚类算法的聚类效果

应用协议	*FNR*			*FPR*		
	k-means	EM	DBSCAN	*k*-means	EM	DBSCAN
SSL	1.82	1.82	0.00	3.08	0.00	0.00
FTP	0.00	98.00	0.00	13.59	0.00	3.91
HTTP	30.30	0.00	0.00	0.00	3.80	0.00
BT	31.75	1.59	4.93	5.35	32.62	0.00

注：参数设置：*k*-means (*k*=4)、EM (*k*=4)、DBSCAN (*Eps*=0.085, *MinPts*=6)

表 5　增加 POP3 流量时不同聚类算法的聚类效果

应用协议	*FNR*			*FPR*		
	k-means	EM	DBSCAN	*k*-means	EM	DBSCAN
SSL	1.82	1.82	0.00	4.5	0.00	0.00
FTP	99.00	98.00	98.00	0.00	0.00	0.00
HTTP	21.21	0.00	0.00	0.53	7.41	0.00
POP3	1.47	5.88	0.00	38.50	30.48	42.66

注：参数设置：*k*-means (*k*=4)、EM (*k*=4)、DBSCAN (*Eps*=0.07, *MinPts*=6)

从结果可知，当混入 MSN 协议或 BT 协议的流量时，相比其他 2 种算法，DBSCAN 算法最好，具有低误报和低漏报，但当混入 POP3 协议的流量时，DBSCAN 也同样不能区分 POP3 和 FTP 协议的流量，它将绝大部分的 FTP 协议流量误报为 POP3 协议的流量，这缘于 2 种协议交互过程产生的会话行为特征比较相似。

5.2　基于载荷信息的 ROCK 流量聚类方法评估

分别取会话前 4 个报文（不包括 TCP 返回的 ASK 报文）的前 6 个字节拼接成 24 个字节的字节序列作为特征，利用 ROCK 算法对混合流量进行聚类。

ROCK 算法对于离群点（outliers）的处理有 2 种方法：1）在聚类前得到每对会话的 *link* 值时，通过设定阈值来去除没有或较少邻居的样本；2）通过设定一个大于实际簇数 k 值来实施聚类，在聚类之后再去除离群点。本实验采用第 1 种方法（丢掉 *link*=0 的点）。

ROCK 算法有 2 个重要参数：相似度阈值 θ 和聚类簇数 k。在实现算法时，当簇没有邻居或剩下的簇数目等于 k 时，停止聚类，因此最后的聚类数目大于等于 k。由于实际的混合流量中包含未知流量的种类有限，因此在设置簇数 k 时，可以设定一个稍大的 k 值来去除一些离群点来获得较纯的簇；θ 对聚类结果的影响将通过实验评估。

实验 3　选择行为相似的 3 种协议的流量，评估基于载荷信息的 ROCK 流量聚类方法的性能，并测试不同 θ 值对结果的影响，同时与基于会话行为特征的 DBSCAN 方法进行比较，其结果如表 6 所示。

表 6　基于载荷信息的 ROCK 算法的流量聚类效果

应用协议	*FNR*			*FPR*		
	ROCK (θ=0.2)	ROCK (θ=0.4)	DBSCAN	ROCK (θ=0.2)	ROCK (θ=0.4)	DBSCAN
FTP	10.00	10.00	0.00	5.40	0.00	57.50
SMTP	99.00	4.30	99.00	0.00	0.00	0.00
POP3	3.30	3.70	42.90	39.30	0.00	2.40

注：参数设置：ROCK(*k*=3)、DBSCAN (*Eps*=0.11, *MinPts*=6)

由表 6 可知，当 θ=0.4，没有误报率；但当 θ=0.2 时，则存在一定程度的误报，说明 θ 对 ROCK 算法聚类结果有很大影响。选择小的 θ 可以减少样本的丢弃，但很有可能会引起误报，导致簇纯度降低，而选择大的 θ 会使邻居数减少，从而减小数据点间的链接数量，降低误报。基于会话行为特征的 DBSCAN 方法在最佳参数时的聚类效果明显差于基于载荷信息的 ROCK 方法，原因在于这 3 种协议具有相似的会话行为特征，但其载荷信息有差异。

理论上，ROCK 算法的时间复杂度为 $O(n^2 \log n)$，高于 DBSCAN 算法的时间复杂度

$O(n^2)$，实验中它们的实现方法不同，无法直接比较其效率。作为载荷特征自动提取方法的数据预处理，一般采用离线式处理方法，因此，方法的效率和效果相比，后者更为重要。

6 结束语

本文研究了未知混合流量中不同应用流量的自动归类问题。通过对会话的各种候选特征进行分析，最终选择会话的前 4 个报文大小和方向作为区分特征，利用适合于此特征的 3 种不同聚类算法（*k*-means、EM、DBSCAN）对来自真实网络上的混合流量进行聚类分析，实验结果表明，基于密度的 DBSCAN 聚类算法具有较低的漏报率与误报率，其效果明显好于其他 2 种算法。但在不同应用的会话行为较为相似时，这种基于会话行为特征的方法仍不能有效地分离不同应用的会话。为此本文提出一种基于载荷信息的 ROCK 流量聚类方法，通过对载荷中字节进行编码，利用 Jaccard 系数作为相似性度量，评估结果显示，这种方法具有较好的效果。然而，当它完全面对未知应用流量、在缺乏评判标准时，如何能够自适应选择参数，达到未知应用的自然分类仍需要进一步研究。

参考文献：

[1] BERNAILLE L, TEIXEIRA R, AKODKENKUT I, *et al*. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26.

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

[3] SEN S, SPATSCHECK O, WANG D M. Accurate, scalable in-network identification of P2P traffic using application signatures[A]. Proceedings of WWW2004[C]. New York, 2004.512-521.

[4] Application layer packet classifier for linux[EB/OL]. http://17-filter. Sourceforge.net,2012.

[5] PARK BC, WON YJ, KIM MS, *et al*. Towards automated application signature generation for traffic identification[A]. Proceeding of the IEEE/IFIP Network Operations and Management Symposium[C]. Salvador da Bahia, 2008.160-167.

[6] WANG Y, XIANG Y, ZHOU W L, *et al*. Generating regular expression signatures for network traffic classification in trusted network management[J]. Journal of Network and Computer Applications, 2011, 17: 1-9.

[7] YE M J, XU K, WU J P, *et al*. Autosig-automatically generating signatures for applications[A]. Proceeding of the IEEE Ninth International Conference on Computer and Information Technology[C]. Xiamen, China, 2009.104-109.

[8] 赵咏, 姚秋林, 张志斌等. TPCAD: 一种文本类多协议特征自动发现方法[J]. 通信学报, 2009, 30(10A): 28-35.
ZHAO Y, YAO Q L, ZHANG Z B, *et al.* TPCAD: a text-oriented multi-protocol inference approach[J]. Journal on Communications, 2009, 30(10A): 28-35.

[9] 刘兴彬, 杨建华, 谢高岗等. 基于 Apriori 算法的流量识别特征自动提取方法[J]. 通信学报, 2008, 29(12): 51-59.
LIU X B, YANG J H, XIE G G, *et al*. Automated mining of packet signatures for traffic identification at application layer with Apriori algorithm[J]. Journal on Communications, 2008, 29(12): 51-59.

[10] 龙文, 马坤, 辛阳等. 适用于协议特征提取的关联规则改进算法[J]. 电子科技大学学报, 2010, 39(2): 302-305.
LONG W, MA K, XIN Y, *et al*. Improved association rules algorithm for protocol signatures extracting[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(2): 302-305.

[11] 王变琴, 余顺争. 一种自适应网络应用特征发现方法[J]. 通信学报, 2013, 34(3): 127-137.
WANG B Q, YU S Z. Adaptive extraction method of network application signatures[J]. Journal on Communications, 2013, 34(3): 127-137.

[12] ZHANG M W, LIU D P. Scalable and accurate application signature discovery[A]. Proceeding of the IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application[C]. 2008.482-487.

[13] MA J, LEVCHENKO K, KREIBICH C, *et al*. Unexpected means of protocol inference[A]. Proceeding of the 6th ACM SIGCOMM Conference on Internet Measurement[C]. New York, NY: ACM, 2006: 313-326.

[14] 李伟明, 张爱芳, 刘建财等. 网络协议的自动化模糊测试漏洞挖掘方法[J]. 计算机学报, 2011, 34(2): 242-254.
LI W M, ZHANG A F, LIU J C, *et al*. Automatic network protocol fuzz testing and vulnerability discovering method[J]. Chinese Journal of Computers, 2011, 34(2): 242-254.

[15] MCGREGOR A, HALL M, LORIER P, *et al*. Flow clustering using machine learning techniques[A]. Proceedings of PAM'04[C]. Antibes Juan-les-Pins, France, 2004.205-214.

[16] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[A]. Proceeding of LCN'05[C]. Sydney, Australia, 2005.

[17] ERMAN J, ARLITT M, and MAHANTI A. Traffic classification clustering algorithms[A]. Proceedings of SIGMETRICS'06 (MineNet)[C]. Pisa, Italy, 2006.281-286.

[18] 董仕, 王岗. 基于UDP流量的P2P流媒体流量识别算法研究[J]. 通信学报, 2012, 33(12): 25-34.

DONG S, WANG G. Research on P2P streaming media identification based on UDP[J]. Journal on Communications, 2012, 33(12): 25-34.

[19] MOORE A W, ZUEV D. Discriminators for Use in Flow-based Classification[R]. Intel Research, Cambridge, 2005.

[20] BERNAILLE L, TEIXEIRA R, AKODKENOU I, *et al*. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26.

[21] BERNAILLE L, TEIXEIRA R, SALAMTIAN K. Early application identification[A]. Proceedings of CoNEXT'06[C], Lisboa, Portugal, 2006.

[22] JAIN A K, DUBES R C. Algorithms for clustering data[M]. Prentice-Hall, Inc, 1988.

[23] DUMPSTER A P, PAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[24] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial database with noise[A]. Proceedings of the International Conference on Knowledge Discovery in Databases and Data Mining[C]. Portland, Oregon, 1996.226-231.

[25] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes[J]. Information System, 2000, 25(5): 345-366.

[26] WEKA[EB/OL]. http://www.cs.waikato.ac.nz/~ml/weka/index.html.

作者简介：

王变琴（1963-），女，陕西蒲城人，博士，中山大学高级工程师，主要研究方向为网络安全与数据挖掘。

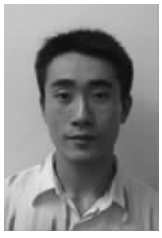
余顺争（1958-），男，江西南昌人，博士，中山大学教授、博士生导师，主要研究方向为计算机网络与网络安全。

（上接第 163 页）

[10] TANG X H, LI C, XIE R Q. Square attack on CLEFIA[J]. Journal of Electronics & Information Technology, 2009, 31(9): 2260-2263.

[11] 李超，孙兵，李瑞林．分组密码的攻击方法与实例分析[M]．北京：科学出版社，2010.
LI C, SUN B, LI R L. Block Cipher Attack Method and Example Analysis[M]. Beijing: Science Press, 2010.

作者简介：



潘志舒（1985-），男，江苏镇江人，解放军信息工程大学硕士生，主要研究方向为分组密码设计与分析。

郭建胜（1972-），男，河南沁阳人，解放军信息工程大学教授、博士生导师，主要研究方向为密码学和信息安全。

曹进克（1964-），男，河南偃师人，硕士，解放军信息工程大学副教授，主要研究方向为信息安全理论与技术。

罗伟（1987-），男，四川双流人，解放军信息工程大学硕士生，主要研究方向为分组密码设计与分析。