一种流序列化的网络流量分类算法

李国元，李双庆，杨 铮
(重庆大学 计算机学院，重庆 400044)

摘 要：针对传统基于端口和有效负载的网络流量分类算法识别率低、分类算法复杂等问题，在分析了网络流量性质的基础上，提出流序列化方法。它将网络流分解成多个流原子，通过提取序列化网络流的特征向量并使用迭代最优化的聚类算法进行流量聚类，最终实现了网络流按不同行为模式聚类。该方法在实验环境中取得了良好的效果。

关键词：网络流；流序列化；迭代最优化；聚类
中图分类号：TP393 文献标识码：A

A network traffic classification algorithm based on flow serialization

LI Guo Yuan, LI Shuang Qing, YANG Zheng
(Department of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: Based on analysis of the properties of network traffic, this paper proposes a flow serialization method to alleviate the problems of low recognition rate and high implementation complexity associated with the traditional port-based or payload-based classification algorithms. The method divides network traffic into multiple flow atoms, extracts characteristic vector from the serialized flows, and finally assembles them into different behavior modes of clusters using iterative optimization algorithm. It works well and has good result in experimental environment.

Key words: network traffic; flow serialization; iterative optimization; cluster

网络流量分类技术是许多网络技术的基础，它关系到网络的控制、性能、安全、管理等多方面内容。通过对网络流量进行分类，可以为网络的运行和维护提供重要信息，对于网络性能分析、异常监测、链路状态监测、容量规划等发挥着重要作用。但是流量类型的多样性，流量特征的复杂性，以及网络新技术的出现(如P2P、数据包伪装及加密等技术)，使基于端口和负载分析的流量分类方法在时间、空间以及扩展性方面难以提高。一种更好的解决方法是使用机器学习的方法进行流量分类。为此本文提出一种基于聚类的无监督学习方法，通过对网络流量的采集、特征提取及模型选择，设计更具通用意义的分类器，从而揭示网络流量的一些内在规律和性质。

1 流量分类技术及现状研究

当前流量分类技术的一个发展方向是基于网络流量有效载荷的识别方法。它主要采用端口技术、包特征签名技术[1]以及会话分类技术[2]等。这种分类方法依赖于协议类型，需要对报文进行深度内容检查，对其有效载荷(payload)部分进行扫描，效率比较低下。它无法对加密的流量进行分类识别。更重要的是，它需要人工更新以识别不断涌现的网络新流量。

正是由于有效载荷分类方法的不足，最近，已有几种机器学习方法，如线性判别分析、参数化方法和基于朴素贝叶斯的核估计方法已经应用到网络流量分类中。如Moore[3]等通过对分类的流量数据集进行学习，对按照高斯分布的各类流量属性进行评估。每一个连接都会根据连接的条件概率，按其属性值进行分类。这种有监督学习的方法准确率很高，但不能发现新的应用，还需要预先知道分类的数量并标记训练样本集。而无监督学习的分类方法没有这些限制，参考文献[4]根据流的性质，对报文平均长度、到达间隔和持续时间等进行特征选择，并使用EM算法对每个报文属于哪一类进行概率计算。它适用于训练数据不是很充分的情况下对流量的模糊聚类。参考文献[5]提出基于统计特性的机器学习流量分类器框架，并在此基础上研究了特征选取方法。它使用顺序前进法(SFS)对选取的流量特征进行评估，有助于提高分类效果。参考文献[6]对各类聚类算法进行了比较，用以评估各类算法在网络流量分类中的性能。图1描述了使用机器学习方法对网络流量进行分类的过程。

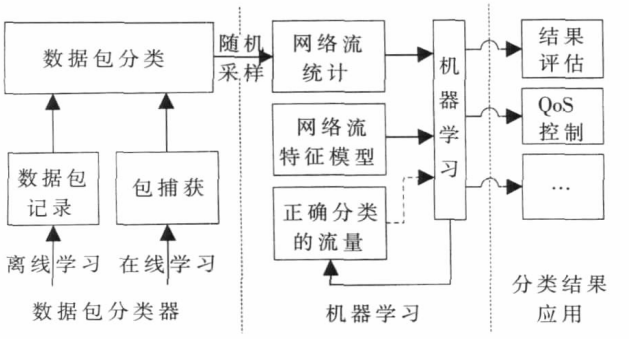

图 1 基于机器学习的流量分类

2 网络流性质及特征提取

2.1 网络流分析

定义 1：一条网络流定义为 5 元组：NFlow::=<SIP, DIP, SPort, DPort, Protocol>。

它是指在超时约束下的 2 个主机对之间应用进程的双向通信的报文集合，即包含相同的主机 IP 地址对，使用相同的协议(如 TCP、UDP、ICMP 等)，采用相同的进程端口对。网络流作为终端间数据交互的载体，以报文的形式穿梭在各个网络链接和存储转发的路由器之间。不同应用通过流的形式得以传输，从而在网络上形成各类异构流量。

流在数据交互的过程中，具有一定的性质，这正是表征一条网络流所属类别的关键。为更好地验证流在数据交互过程中的规律特性，通过抓包分析了使用 HTTP 和 FTP 下载过程中网络流数据包的分布情况（如图 2、图 3 所示）。图 2 中网络流很清楚地分成了两部分，包长约 1 500 B 的包是服务器发送给客户端的数据，而包长约 60 B 的包是客户端发给服务器数据包的 ACK 包。图 3 中数据包之间比较密集，但还是看出该网络流的特性，即服务器端发送的数据包要比客户端稍大。通过分析发现，不仅是 HTTP、FTP，其他的如 TELNET、SMTP 和 BT 等应用的网络流也都具有各自的分布特点。

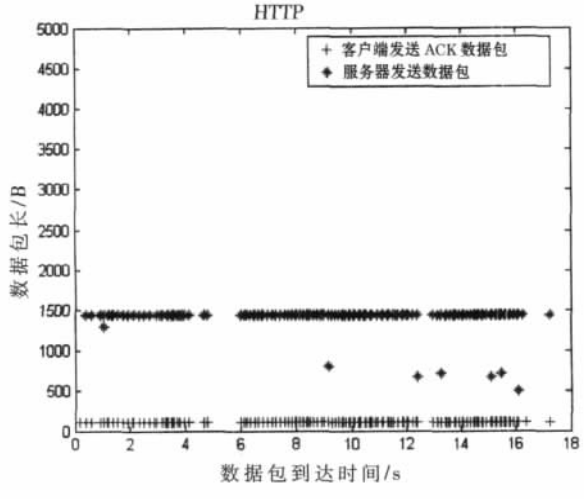

图 2 HTTP 下载数据包分布

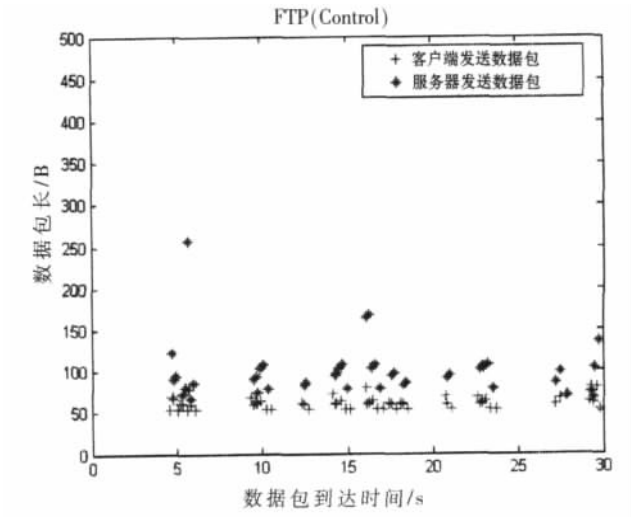

图 3 FTP 控制连接数据包分布

通过以上分析看出：网络流的特征不仅包含流中数据包自身的性质(如最大/最小包长)，还包含流的统计特性(如平均包长、连接周期及空闲时间等)，以及其在连接周期中数据包的分布情况。

2.2 网络流序列化

将网络流看作一连串数据包的序列，则每个数据包将按其到达目的主机的时刻分布在时间轴上。

定义 2：一个流原子定义为 2 元组：A::=<S, D>。**它表示沿时间轴方向上捕获通信双方发送数据包的大小。其中 S 表示连接的发起者发出数据包大小，而 D 表示流反向传输的数据包大小。则一条网络流可表示为一个流序列**$(A_1 \cdots A_n)$，**这里 n 是该网络流所含流原子的个数。例如上面 FTP 控制连接的流序列可表示为：**

((69,90),(67,95),(54,1095)(60,73)…)

网络流交互过程并不是完全对称的，有可能存在不能一一对应的情况。上面使用 HTTP 下载的例子中，客户端有时在收到 2 个数据包后才发送 1 个 ACK 包。流序列规定：在沿时间轴方向上，若连续 $N(N \geqslant 2)$ 个数据包流向相同，则在这 N 个数据包之间插入 $N-1$ 个 0 字节的反向数据包，以确保流原子的完整性，同时也不影响网络流性质。HTTP 下载的流序列片段为：

(…(0,1480),(54,1480),(0,1480),(54,1480)(54,1480)…)

2.3 特征提取

本文在网络流序列化基础上进行流特征提取。设集合 A、B 分别表示流序列中两个流向的数据包集合，即 $A=\{S_1, S_2 \cdots S_n\}$，$B=\{D_1, D_2 \cdots D_n\}$，其中 n 是流序列中流原子个数。在集合 A 中选取 $S_{Max}=\mathrm{Max}(S_i \in A)$、$S_{Min}=\mathrm{Min}(S_i \in A, S_i \neq 0)$、$S_t=\sum_{i=1}^{n} S_i$、均值 S_μ，标准方差 S_σ 作为流特征。为更好地表现流的连续分布特性，选择连续流向相同数据包出现次数 S_c 和 $S_a=S_{Max}/S_{Min}$ 作为流特征(集合 B 选择相同特征)，并选取了流中双向流数据量比例 $f_r=S_t/D_t$、流的连接周期 T 等作为流特征，如表 1 所示。网络流经过

表 1 选取的 17 项网络流特征

网络流特征	特征描述
n	流序列中流原子个数
S_{Max}, D_{Max}	最大数据包容量
S_{Min}, D_{Min}	最小数据包容量
S_t, D_t	总数据包容量
S_μ, D_μ	均值
S_σ, D_σ	标准方差
S_c, D_c	连续同向包出现次数
S_a, D_a	最大最小数据包比
f_r	不同流向流量比
T	连接周期

特征提取,需要进一步归一化处理,才能作为机器学习方法中的特征向量使用。下面介绍使用相关机器学习方法进行流量聚类。

3 基于迭代最优化聚类算法

聚类算法是基于整个数据集内部存在“分组”而产生的一种数据描述方法。基于迭代最优化聚类算法使用误差平方和作为准则函数,通过调整每个样本所属类别,使调整后的准则函数值得以改善。

令 c 表示聚类划分个数,n_i 表示子集 R_i 中的样本数量,x 表示样本特征向量,m_i 表示该子集中样本的均值,则:$m_i=\frac{1}{n_i}\sum_{x\in R_i}x$。定义误差平方和为 $J_e=\sum_{i=1}^{c}\sum_{x\in R_i}\|x-m_i\|^2$,下面考虑利用迭代方法使误差平方和准则达到最小值。

3.1 迭代最优化

令 $J_i=\sum_{x\in R_i}\|x-m_i\|^2$,假设某一样本 x_0 原属于聚类 R_i,现被放入聚类 R_j 中。则 m_j 变成 $m_j^*=m_j+\frac{x_0-m_j}{n_j+1}$,而 J_j 增加为:

$$J_j^*=\sum_{x\in R_j}\|x-m_j^*\|^2+\|x_0-m_j^*\|^2$$

$$=\left(\sum_{x\in R_j}\|x-\frac{x_0-m_j}{n_j+1}-m_j\|^2\right)+\left\|\frac{n_j}{n_j+1}(x_0-m_j)\right\|^2$$

$$=J_j+\frac{n_j}{n_j+1}\|x_0-m_j\|^2$$

假定 $n_i\neq 1$,同样求得 m_i 变成 $m_i^*=m_i-\frac{x_0-m_i}{n_i-1}$,而 J_i 下降为 $J_i^*=J_i-\frac{n_i}{n_i-1}\|x_0-m_i\|^2$。如果 J_i 的减少量比 J_j 的增加量大,则 J_e 变小,这时 x_0 从聚类 R_i 转移到聚类 R_j 中是有利的,即 $\frac{n_i}{n_i-1}\|x_0-m_i\|^2>\frac{n_j}{n_j+1}\|x_0-m_j\|^2$。对有利的转移选取最佳的 j $(j\neq i)$,使得对应的 $\frac{n_j}{n_j+1}\|x_0-m_j\|^2$ 最小。通过对所有样本进行最优化迭代,最终形成 c 个聚类。这样在聚类个数 c 一定的情况下,误差平方和将达到最小值。具体算法如下:

选取 c 个样本点作为初始点,初始化 $m_1,m_2\cdots m_c$
随机选取一个样本 x_0
$k\leftarrow arg\ \min_i\ \|m_i-x_0\|$
分类 x_0 到第 k 个聚类中
如果 $n_k\neq 1$ 则计算
　for $j=1$ to c
　　如果 $j\neq k$ 则
　　$\Delta p_j=\frac{n_j}{n_j+1}\|x_0-m_j\|^2$
　　否则 $\Delta p_j=\frac{n_j}{n_j-1}\|x_0-m_j\|^2$
　$k'\leftarrow \arg\ \min_j\ \Delta p_j$
将样本 x_0 从第 k 个聚类转移到第k'个聚类中
重新计算 J_e, m_k 和 m_k'。
until 所有样本都被选取分类
return $m_1,m_2\cdots m_c$

3.2 迭代最优化聚类算法分析

假设有 n 个待分类的样本,它们都在 d 维空间中。首先要计算每个样本点与各个划分中心点的距离,并选取最小距离进行初步分类。每个距离需要 $O(d^2)$次运算,故初步分类过程需要 $O(cd^2)$次运算。样本点与各个聚类中心点的距离存储在一个数组中,相应的空间复杂度为 $O(c)$。在迭代最优化过程中,样本点可以利用已存储的距离数组计算转移的增减量 $\triangle p$,并求出最小量,这一过程需要 $O(c)$次计算。通过对 n 个样本点进行选取划分,其计算次数为 $O(cn(d^2+1))$,故总的计算复杂度为 $O(cnd^2)$。通常 $n>>c$,由此可见算法的性能主要取决于样本数量 n 和特征空间维数 d。

4 实验与分析

NLANR 为因特网社团进行网络分析研究提供网络数据包跟踪记录。本文使用 NLANR 提供的 Auckland-VI[7]数据包跟踪记录作为实验数据集,通过将数据包转换成网络流,并进行序列化处理,最终形成上万条流序列。随机选取了 2 000 条流序列进行特征提取及归一化处理,以此作为基于迭代最优化聚类实验的数据样本。

考虑到 Auckland-VI 提供的跟踪记录只包含 IP 层和传输层信息,同时为了检验聚类方法的合理性,使用了 Auckland-VI 2001 年的跟踪记录。早期的网络流一般使用基于端口的方法进行识别,从而对聚类的结果进行验证。这可能有一定的误差,但对大部分的网络流(特别是早期的)来说都是适用的。

4.1 聚类数目

聚类分析中一个重要的环节就是找到数据集中客观存在的类别数目。在使用迭代最优化算法进行聚类时,通过重复对 $c=1$、$c=2$、$c=3$ 等情况进行聚类尝试,并观察准则函数如何随 c 变化。如图 4 所示,误差平方和 J_e 是 c 的单调递减函数。当 $c\leqslant 6$ 时,J_e 会随着 c 的增加

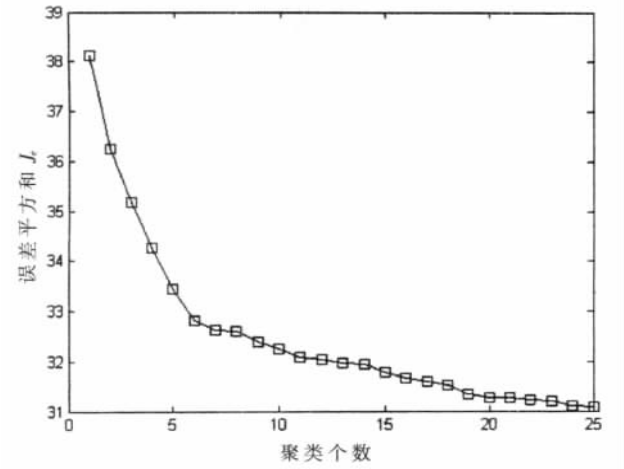

图 4 聚类数目与误差平方和 J_e 关系图

迅速减少，之后下降速度明显减缓，直至 J_e=0 时为止。这说明在随机选取的 2 000 个数据样本中的确客观存在 6 个稠密且分得很开的聚类。在 c=6 的基础上进行分析，因为它更好地表达了数据集的内部结构。

4.2 聚类分布

本文使用端口识别方法对采样的 2 000 条网络流进行识别，以验证聚类结果的合理性。经过端口识别发现，在 2 000 条网络流中主要包括 HTTP (80,8080)、HTTPS (443)、SMTP (25)、POP3 (110)、BT (6881－6889)、EDONKEY (4662)、NTP (123)、DNS (53) 等应用，而未知流量约占 5% 左右。

按照迭代最优化算法将 2 000 条网络流聚成 6 类 ($c1$,$c2$…$c6$) 后，结果如表 2 所示。分析发现：$c1$ 约占总流数的 49 %，其特点是流原子个数较少 (≤7)，持续周期短 (1s)，流原子 D 端数据量不大 (约 3 000 B)。该类主要是网页流量，一般是通过网页获取 HTML 页面、图标和小图片等较小的对象时产生流量。$c2$ 主要包含的是 SMTP 和 POP 流量，该类中流原子 D 端总数据量均值达到了 17 000 B 且方差较大。还有部分通过 HTTP 获取较大对象的流量也包含在 $c2$ 中。c3、c4 主要包含 P2P 应用流量，如 BT、EDONKEY 之类。之所以分为两类是因为 $c3$ 与 $c4$ 相比，其流原子 D 端数据包较小 (约 300 B)，而 c4 流原子 D 端数据包多为 1 500 B，同时这两类在不同流向比和持续周期方面相差较大 ($c4$ 的持续周期均值达 70 s)，因此判断 $c3$ 主要聚集的是 P2P 通信中会话建立，连接信息交互及资源搜索等产生的流量，而 $c4$ 则是真正使用 P2P 进行文件下载的聚类。$c5$ 的特点是流原子 S 端数据较小，但流原子个数多，不同流向比接近 1，它主要包含 telnet、https 等应用。$c6$ 中几乎 95 % 都是 DNS 应用，其流原子个数很少，一般只有 2~3 个。

表 2 聚类分布情况

聚类子集	流数	包含应用
$c1$	982	HTTP，HTTPS…
$c2$	270	SMTP，POP3，HTTP…
$c3$	145	BT，EDONKEY…
$c4$	446	BT，EDONKEY，FTP…
$c5$	108	TELNET，HTTP…
$c6$	49	DNS…

4.3 有效性验证

聚类算法的有效性可以通过聚类精确度进行计算。聚类精确度衡量了聚类算法将单一应用归为一类的能力。定义聚类精确度 H_i 为聚类 i 中占主要成分的应用流在该类中所占百分比，即 $H_i=\max(\frac{N_{ij}}{N_i})$ $(1\leqslant i\leqslant 6)$，其中 N_i 表示聚类 i 中总流数，N_{ij} 表示应用 j 在聚类 i 中的流数。尽管每个聚类中 H_i 值各不相同，但最大化的 H 均值更能体现聚类性能。使用 k－均值聚类、层次聚类和迭代最优化聚类算法对以上的 2 000 个样本进行分类并分析比较，如图 5 所示。以迭代最优化算法为例，$c1$ 和 $c6$ 中主要应用分别是 HTTP 和 DNS，其流特征明显，聚类精确度较高 (≥90 %)；$c2$ 中的应用主要以 SMTP 为主，约占该类 60 % 流数；$c3$ 和 $c4$ 中是聚类精确度只有 50 %~60 %，这是因为其中各种应用较多，所占比例较为分散；$c5$ 的应用以 telnet 为主，约占该类 80% 流数。

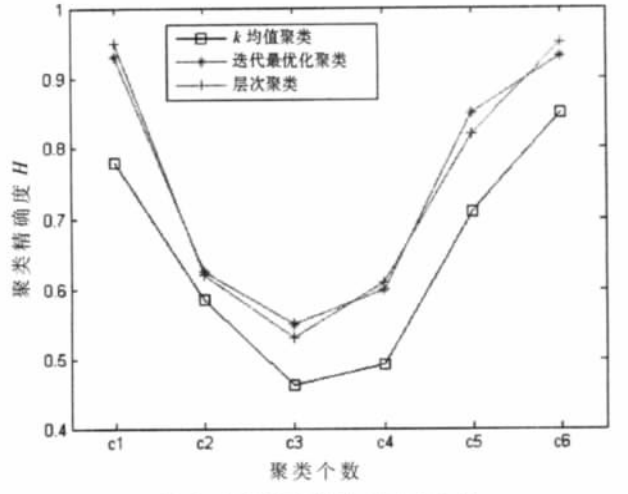

图 5 各类算法聚类精确度比较

虽然 k－均值聚类与迭代最优化聚类均属于迭代优化算法范畴，但 k－均值聚类对初始类中心比较敏感，易于陷入局部最优，这使其 H 均值只能达到 65%。而层次聚类的效果与迭代最优化算法相当，其 H 均值都达到了 74 %，但它的时间复杂度太大 ($O(cn^2d^2)$)，同时空间复杂度达到了 $O(n^2)$，不宜于大规模聚类。因此，基于迭代最优化的聚类算法因其计算复杂度低、算法简单，更适用于特征维数高、样本数量多的流量分类中。

如何合理地分类未知流量、有效地提高网络流量分类效率，并降低分类算法的计算复杂度，是目前众多学者对流量分类研究的重点。本文根据网络流性质，提出

(下转第 127 页)

式中 $p_{ij}=s_{ij}/|s_j|$,$|s_j|$是 s_j 中样本属于类 C_i 的概率。

在属性 A 上该划分获得的信息增益为:

$$\sigma=H(C)-E(A)$$

根据上面的计算得到每个属性的权重系数为:

$$w_i=\sigma_i/\sum_{i=1}^{m}\sigma_i$$

从分析中知道,该权重系数反应了样本中各个属性的重要程度,权重系数值越大则该属性越重要,对分类的贡献越大。

在确定了样本属性重要度后,就可以构造基于样本属性重要度的代价敏感支持向量机。

4 实验结果

本文利用 MATLAB 软件进行模拟实验,对+1 类和-1 类的分类性能进行比较,在三维空间中引入两类不同的样本:正类和负类,并引入了一定数量的噪声和野值数据。为了验证所提算法的有效性,利用所提算法进行了一系列比较实验。在实验中,模拟用的训练样本和测试样本均随机产生,样本数据情况如表 1 所示。

表 1 训练集和测试集样本

	训练集	测试集
样本数	847	638
正类样本数	427	315
负类样本数	420	323

在实验中考虑正类的错分代价大于负类的错分代价,分别用 C-SVM、Cost-sensitive SVM 和属性加权的 Cost-sensitive SVM 进行性能测试,表 2 所示为分类准确率的比较。由表 2 可见代价敏感支持向量机分类算法提高了错分代价高的类别的分类精度,在进行属性加权后,总体的分类精度也得到了提高。

表 2 三种算法的测试结果比较

	C-SVM	Cost-sensitive SVM	属性加权 Cost-sensitive SVM
正类分类正确率/%	97.73	99.55	100
负类分类正确率/%	98.18	96.82	98.37
总体正确率/%	97.82	98.39	99.17

本文在对支持向量机分析的基础上,提出了对样本属性加权型的代价敏感加权支持向量机。数值实验的结果表明,该方法能够提高错分代价敏感的类别的分类精度,同时整体的分类性能也得到了提高。但是如何确定代价系数仍然是一个需要解决的问题,也是笔者下一步要研究的方向。

参考文献

[1] 范昕炜,杜树新,吴铁军.可补偿类别差异的加权支持向量机算法[J].中国图像图形学报,2003,8(7):1037-1042.

[2] 贾银山,贾传荧. 一种加权支持向量机分类算法[J]. 计算机工程,2005,10(5):35-39.

[3] LIN C F, WANG S D. Fuzzy support vector machine [J]. IEEE Trans. On Neural Networks, 2002, 13(2):464-471.

[4] 陈小娟, 刘三阳. 一种新的模糊支持向量机算法[J]. 西安文理学院学报:自然科学版,2008,11(1):1-4.

[5] 汪延华,田盛丰. 样本属性重要度的支持向量机方法[J]. 北京交通大学学报,2007,10(5):43-46.

[6] 赵靖.基于 SVM 算法的垃圾邮件过滤研究与实现[D]. 北京:北京交通大学,2005.

(收稿日期:2008-12-11)

(上接第 124 页)

流序列化的思想,将网络流分解成流原子,并在此基础上进行流特征选取,以获得更能表征其内在性质的特征。接着将基于迭代最优化聚类算法应用到流量分类中,通过实验发现网络流具有不同行为模式,如下载模式、单向模式和交互模式等。这种迭代最优化的方法对进行流量分类研究具有重要意义。下一步的工作是研究如何提高聚类的精确度,以及如何把该方法应用到在线学习的网络流量分类中。

参考文献

[1] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network application[C]. in Passive & Active Measurement Workshop 2005.

[2] SU Hui Kai, WU Cheng Shong, CHEN Kim Joan. Session classification for traffic aggregation[C]. IEEE International Conference 2004.

[3] MOORE A W, ZUEV D. Internet traffic classification using bayesian analysis techniques[J]. Joint International Conference on Measurement and Modeling of Computer Systems. 2005:50-60

[4] MCGREGOR A, HALL M, LORIER P, et al. Flow clustering using machine learning techniques[C]. Lecture Notes in Computer Science, 2004

[5] ZANDER S, NGUYEN T, ARMITAGE G. Self-learning IP traffic classification based on statistical flow characteristics[J]. Passive and Active Network Measurement 2005.

[6] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms[C]. Proceedings of the 2006 SIGCOMM workshop 2006.

[7] NLANR Measurement and Network Analysis Group. Trace 20010613-060000-1[EB/OL].http://pma.nlanr.net/Traces/Traces/long/auck/6/

[8] DUDA R O, HART P E, STORK D G. 模式分类[M]. 李宏东,姚天翔,译. 第 2 版, 北京:机械工业出版社, 2006

(收稿日期:2008-11-26)