基于机器学习的网络应用识别研究

王变琴 余顺争
(中山大学电子与通信工程系 广州 510275)

摘 要 近年来,网络应用识别在学术和应用领域备受关注和快速发展,已形成一个相对独立的研究领域。基于机器学习的识别方法更是成为研究热点,并且取得一系列初步成果。研究了这方面的相关问题,分别从识别粒度、特征选择及识别算法等方面进行介绍、归纳,并对典型方法进行重点分析,最后指出了存在的问题及未来研究方向。

关键词 应用识别,机器学习,流特征,算法

Research on Machine Learning Based Network Application Identification

WANG Bian-qin YU Shun-zheng
(Dept. of Electronics and Engineering, Sun Yat-sen University, Guangzhou 510275, China)

Abstract In recent years, identifying and categorizing network traffic by application type attracts great interests both in the fields of academic and application, and has already become a relatively independent research realm. Furthermore, the application identification approaches based on Machine learning have been hotspots and have been obtained promising preliminary results. This paper surveyed the current machine learning algorithms about application classification, respectively from fine-grained identification, feature selection, and recognition steps and so on. It focused on the analysis of typical methods and suggested some future research directions.

Keywords Traffic identification, Machine learning, Flow feature, Algorithms

1 引言

目前 Internet 与企业网上运行有多种网络协议和网络应用,各种新应用仍在不断涌现。准确地识别这些应用,对于网络管理员、研究员、服务提供商及用户都具有重要的意义,它是研究差异性服务、QoS 保障、入侵检测、流量监控、计费管理以及用户行为分析的前提和基础。在过去的网络环境中,绝大部分网络流量被 Web, FTP, SMTP, Telnet 等应用所占据。而近年来,应用的数量与种类都有很大的增加,传统应用流量在总流量中的比重越来越少。相反,P2P、流媒体及网络游戏等新应用不断出现,且已经占据了网络流量的 60%以上[1,2]。更重要的是,这些新应用的规范往往不公开并且不遵守默认固定端口的约定。因此,如何能够准确地识别这些复杂的应用,是识别方法必须解决的新问题。

传统的方法是根据各个应用在 IANA[3] 中注册的端口号来标识应用的。例如,若某个 TCP 流使用了端口号 80,8080 或 443,则将其标记为 Web 流量。虽然原则上可以使用在 IANA 中注册的端口号识别各种应用,但是由于许多因素的存在[4],端口识别法正在越来越多地受到限制。

为了解决端口识别方法所出现的问题,许多研究工作[5-9]致力于 IP 分组载荷(payload)的应用识别技术研究。实际中的商用带宽管理及入侵检测工具常采用此类技术,以增强分类精度。此类方法虽然具有很高的识别准确度,但计算复杂度高,且涉及用户隐私,同时也不能识别加密的数据流。

Thomas Karagiannis 等人[10] 提出基于 P2P 网络的连接模式识别 P2P 应用,该方法在识别精度方面优于端口方法,与载荷方法相当,但不依赖载荷。他们在后续的工作中进一步提出[11]利用主机的行为模式来识别流量,它是一种多级别(Social, Functional, Application)的主机行为分析法。通过识别主机所执行的应用通信模式签名(signature)来区分应用。该方法的最大特点是无需解读分组载荷,从而不会涉及用户隐私问题;不需要知道与端口号相关的信息,因此不易被其所误导,也因此被称为 BLINC(Blind Classification)。该方法易于部署,只需要获得网络监控设备捕获的流信息,且对网络拥塞和路径变化不敏感。然而,该方法的扩展性差,分析复杂及应用识别不精细,且在传输层加密时失效。

基于机器学习的智能应用识别方法研究被看成是目前最具有前景的,它引起学者们的广泛关注和快速发展。这类方法仅基于流的统计特性,易于扩展与维护,计算复杂度远低于载荷识别方法。通过近几年的研究,已经获得一系列初步成果[12-29,33-37]。本文的工作在于总结归纳该类研究的相关问题、最新研究动态、重点分析典型方法,最后指出存在的问题及进一步研究方向。

本文第 2 节介绍了应用识别粒度、流的候选特征、机器学

到稿日期:2008-02-22 本文受国家自然科学基金-广东联合基金重点项目(U0735002),国家高技术研究发展计划(863)(2007AA01Z449)资助。

王变琴 女,高工,博士生,主要研究方向为网络应用识别、行为分析与控制,E-mail: wangbq@mail.sysu.edu.cn 或 wangbianqin@tom.com;
余顺争 男,教授,博士生导师,主要研究领域为Internet 流量测量、分析、建模、统计异常检测、无线网络。

习算法及识别过程等相关问题。第3节对典型方法从算法思想、关键技术与特点等方面进行了分析。第4节总结了现存问题及进一步研究方向。

2 相关问题

2.1 识别粒度

网络应用识别是根据产生流量的应用类型对流量进行识别和分类。由于许多因素限制或不同的需求，应用识别在不同的粒度级别上进行。

(1)粗粒度的识别：按处理方式将应用划分为不同的组，一个组可能包含行为相似的多个应用。例如，文献[4]将流量划分为交互式(例如 Telnet)、块数据传输(例如 FTP-data，Kazaa)、流媒体(例如 Real Media Streaming)及交易型(例如 DNS，HTTPS)，它是对应用进行的粗粒度区分，这种划分可以满足差异性服务机制及 QoS 保障等要求。目前提出的机器识别法大多数只能在这个级别上进行。

(2)应用层协议：识别产生流量的应用层协议。在粗粒度的流分类问题中，每个类别可能包含某些属性类似的多种协议，但协议识别问题必须对流量进行更精细的分类，使得每个类别中的流只使用一种应用层协议，例如 HTTP，FTP，SMTP 等。

(3)应用：识别产生流量的具体应用是最理想的识别结果。例如，P2P 类的具体应用(Morpheus，Limewire，Edonkey，Emule，Bittorren，Gnutella，Fastrac 等)，它是最精细的流量识别。

2.2 流的候选特征

机器学习识别法假设应用发送数据是以某种模式进行的，不同的应用应该具有不同的模式，这些模式被作为识别流的途径，模式可以通过流的统计特性获得。本文为了方便讨论，对流做如下统一描述。

流的描述：设 $X=\{X_1,\cdots,X_n\}$ 是流组成的集合，每个流用 m 个特征刻画，即 $X_i=\{x_{ij}\mid 1\leqslant j\leqslant m\}$，其中 m 为特征数，x_{ij} 为第 i 个流的第 j 个特征值。设 $Y=\{Y_1,\cdots,Y_q\}$ 是应用类别集合，其中 q 是类别数量，Y_i 是类别。机器学习识别法的最终目标是学习将一个 m 维变量 X 映射到 Y。

机器学习识别法的研究对象为流，流可以有许多种候选特征。文献[4]对其进行了较为深入的研究，将其大致分为5类。

(1) 简单分组级特征：直接从分组头部选择或经过简单计算抽取的特征。例如，分组大小均值、各种矩、方差及 RMS (Root Mean Square)大小等。抽样操作对此类特征影响较小。

(2) 流级统计特征：在流级别上的总结统计特征。定义流为分组的单向序列，它们有一些共同的字段值，通常为具有相同的五元组(5-tuple)(源/目的 IP、源/目的端口及 IP 协议类型)字段值。典型的流级特征包括流持续时间均值、每个流的平均数据量、分组数的均值及它们的方差等。有些更复杂的特征也可以从流(或分组)的统计信息中抽取。然而，流数据有一个限制，即在一个流连接中，聚合的多个分组有时可能属于多个应用级连接，这样就扭曲了流级的特征。

(3) 连接级统计特征：可用来跟踪一些与面向传输层连接有关的行为，例如 TCP 连接。一个典型的 TCP 连接从开始到结束，使用预定义好的握手(从客户端到服务器)分组交换。连接处理需要跟踪连接状态以收集连接级的信息。一般情况下，连接级数据较流级数据能提供更丰富的信息，但需要额外的开销，且在测量点受抽样或不对称路由的影响。

(4) 流内或连接内特征：有些特征基于流或 TCP 连接收集，但是需统计每个流内部的分组，例如，流中分组之间到达时间的统计，就需要在分组级收集数据，然后将其分成流(或连接)。流(或连接)内特征包含丢失率、延迟等网络噪声。

(5) 多个流间的统计特征：有些有意义的特征可以通过统计多个流(或连接)之间的信息进行捕获。这些特征与捕获单独的流(或连接)级特征相比，更加复杂，需要较大的计算开销。

以上各种流特征可以从分组头数据(Packet Header Traces)中获得。由于流数据获取是一种被动的流量测量方法，不会给运行的网络造成太大的影响，因此基于流特征的应用识别法是可行的，且易于部署。

2.3 机器学习算法

对于机器学习识别法来说，学习算法是提高识别效果的关键，实用的系统还需考虑计算复杂度及识别过程的吞吐率(速度)等。目前在网络应用识别方面对有监督(supervised)、无监督(unsupervised)及半监督(sem-supervised learning)3种学习形式都有研究。

(1) 有监督学习：给定一个由流组成的集合 $X=\{X_1,\cdots,X_n\}$ 和一个类别集合 $Y=(Y_1,\cdots,Y_q)$，有监督分类问题是定义一个映射 $f:X\to Y$，其中每个流 X_i 被分配到一个类中。一个类精确地包含了被映射到其中的流，即 $Y_j=\{X_i\mid f(X_i)=Y_j\}$，$1\leqslant n$，且 $X_i\in X$。

有监督学习需要大量有类别标签(labeled data)的数据。该类研究有：文献[14]利用最近邻(Nearest Neighbor NN)和线性判别分析(Linear Discriminate Analysis LDA)法将不同的应用类型映射到不同的 QoS 级别，它是有监督学习方法在流量分类方面较早的应用之一；文献[15]深入研究了朴素贝叶斯(Naïve Bayes)法的网络应用识别；文献[16]提出通过简单统计指纹识别应用协议(HTTP，FTP，SMTP)；文献[17]对几种有监督的学习算法(Naïve Bayes，C4.5，Bayesian Network，Naïve Bayes Tree)的性能进行了评价比较。

(2) 无监督学习：给定一个由流组成的集合 $X=\{X_1,\cdots,X_n\}$ 和整数值 k，无监督学习(或聚类)问题就是定义一个映射 $f:X\to\{1,\cdots,k\}$，其中第 i 个流被映射到第 j 个簇 K_j 中去。第 j 个簇 K_j 由所有被映射到该簇中的流组成，即 $K_j=\{X_i\mid f(X_i)=K_j,1\leqslant j\leqslant k,\text{且 } X_i\in X\}$。假设簇的数目 k 是已知的，每个簇 $K_j(1\leqslant j\leqslant k)$ 的实际内容(和解释)作为函数定义的结果被确定，聚类问题的求解视为产生一个簇集 $K=\{K_1,\cdots,K_k\}$ 的过程。

无监督学习在学习时只用无类别标签(unlabeled data)的数据，有类别标签数据只是作为检测其学习性能的测试数据。该类研究有：文献[14]提出基于流的通信模式层次聚类法来构造灵活的流量生成器及文献[18]用期望最大(Expectation Maximization EM)法将流聚成不同的应用类型(Bulk Transfer，Single and Multiple Transactions and Interactive Traffic，amongst others)，它们是无监督学习方法在本领域的初步尝试；文献[19]利用顺序向前特征搜索(Sequential Forward Se-

lection SFS)法和 Autoclass 聚类法进行应用识别(FTP, Telnet, SMTP, DNS, HTTP, AOL Messenger, Napster, Half-Life),其最大贡献在于考虑了特征集的优化问题,为研究者开辟了一个新的探索方向;文献[20]对 K 均值(K-Means)法、高斯混合模型(Gaussian Mixture Mode GMM)及基于隐马尔可夫模型(Hidden Markov Model HMM)的谱聚类法进行了比较评价;文献[21]对聚类算法 K-Means, DBSCA 及 AutoClass 进行评价对比。

(3) 半监督学习:给定一个由流组成的集合 $X=\{X_1,\cdots,X_n\}$,X 可以被分成两部分:$X_l:=\{X_1,\cdots,X_l\}$ 和 $X_u:=\{X_{l+1},\cdots,X_{l+u}\}$,其中 $X_l:=\{X_1,\cdots,X_l\}$ 有对应的类别标识 $Y_l:=\{Y_1,\cdots,Y_l\}$,而 $X_u:=\{X_{l+1},\cdots,X_{l+u}\}$ 则没有类别标识。这是文献[31]对半监督学习法的定义。

介于有监督和非监督学习之间的半监督学习是一种新近发展起来的学习机制,在学习过程中可以同时利用有标签和无标签数据,从而提高分类的准确度。文献[22]首次将半监督识别法用于网络应用识别。

2.4 识别过程

机器学习识别法一般分两步实现:模型建立与分类。先利用训练数据建立模型,然后模型被输入分类器进行分类。图 1 和图 2 显示了基于机器学习的分类体系结构。训练输入数据可以来自采集的流量 traces 或现场捕获。然后,基于五元组信息分解成不同的流来计算流的统计特征。

流特征和算法参数集可以用来建立流模型,见图 1。算法的参数依赖于具体的学习算法从简单到复杂变化,有些算法可能不需要参数。一旦分类器完成训练,就可以对未知的、新的流(基于统计特性)进行识别,见图 2。在实际应用中,一般要求分类器能进行在线实时的应用识别,因此新的流数据要求现场捕获。为了改善学习算法性能,可以对每条流数据进行适当采样。分类的结果可以用来实现 QoS 服务级别映射及流量趋势分析等应用。

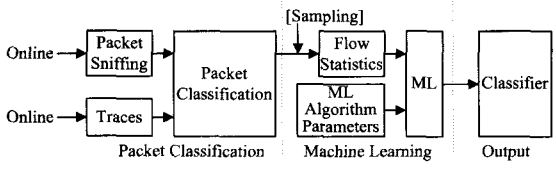

图 1 机器学习法训练过程

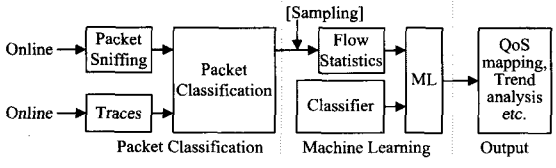

图 2 机器学习法分类过程

3 典型方法分析

最近几年,基于机器学习的应用识别方法研究已经成为热点,研究成果不断出现。比较有代表性的研究有文献[15]基于有监督贝叶斯的网络流量分类;文献[20,23]基于聚类的早期应用识别及文献[22]半监督网络流量分类。它们的研究相对系统、完善,在某种程度上代表了目前的研究方向、动态及水平。下面对其进行详细分析,希望对研究者有一定的启发作用。

3.1 贝叶斯法的流量分类

Andrew W. Moore 等人[15]系统地研究了基于贝叶斯法的应用识别(Bulk, Database, Interactive, Mail, Services, WWW, P2P, Attack, Games, Multimedia)。研究对象为完整的 TCP 连接,即具有相同五元组信息的两个主机之间所有的分组(从第一个 SYS 分组开始到最后 FIN ACK 之间的所有分组)。利用 248 不同特征(discriminator)[27]来描述流,这些特征包括流持续时间的统计、TCP 端口信息、载荷大小统计及分组到达时间的傅立叶变换等特征。

基本的分类算法为朴素贝叶斯法[24],它是一种最简单的方法。有监督贝叶斯分类包括两个步骤:(1)通过对训练集(流的类别归属已经标识)计算建立统计模型;(2)将建立的统计模型应用于网络流量中未知的、新的流分类。每个新流 X_i 获得属于特定应用的概率:

$$p(Y_j \mid X_i)=\frac{p(Y_j)f(X_i \mid Y_j)}{\sum_{Y_j} p(Y_j)f(X_i \mid Y_j)} \quad (1)$$

其中 $p(Y_j)$ 代表类别 Y_j 的先验概率,$f(X_i \mid Y_j)$ 为给定 Y_j 时 X_i 的条件概率。该方法假设流的特征之间是相互独立的,且同服从高斯分布。

实验结果显示每个流的分类精度约 65%。通过利用核密度估计(Kernel Density Estimation)对算法本身进行优化及特征选择与降维压缩[28]方法的改善,使分类精度超过 95%,与传统的方法(精度 50~70%)相比,精度有很大的提高。该作者在文献[29]中进一步探索了这种方法基于 NetFlow 数据的轻量级应用分类的可行性和算法性能。结果表明该方法的识别准确度>91%,稍低于流级(>91%)的识别准确度。

方法特点:最大特点是利用手工分类(基于载荷)的网络流量数据作为训练和测试数据,识别精度高,且学习算法简单、易于使用,但模型受限于特征量之间是独立同分布的,这种假设是不现实的。同时,这种方法特征量太多,难以实现在线实时识别。有监督学习法另一个缺点是不能识别以前未出现过的应用。

3.2 聚类法的早期应用识别

Laurent Bernaille 等人[21,23]利用聚类法进行早期应用识别。识别的对象为 TCP 连接,不包含 TCP 控制分组(SYN, Keep-Alive,不含载荷的 ACK)。选取的特征为 TCP 连接的开始 p 个分组的大小和方向。选用了 3 种聚类算法来识别应用(POP3, SMTP, SSH, HTTPS, POP3, HTTP, FTP, Edonkey and Kazaa)。

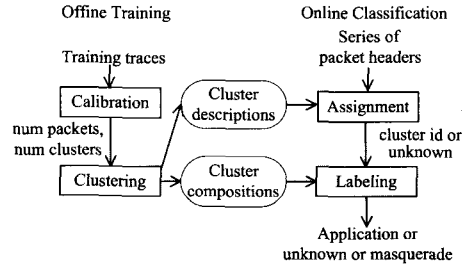

图 3 方法总流程

基本思想:直观上,开始几个分组的载荷大小能够捕获应用协商阶段的行为,通常这是一个预定义的消息序列,不同的应用之间是有区别的。该方法将一个 TCP 连接的前 p 个分组的大小和方向作为相应应用的行为,p 的值为整数,通过实

验确定。分类器包含一套 TCP 连接到应用的映射规则。识别过程分两个阶段：离线训练阶段和在线分析阶段，见图 3。

(1) 训练阶段：通过聚类将训练集(所有目标应用的 TCP 连接的代表样本)分割成具有相似行为的 TCP 连接组，以获得应用行为模型。首先，解析数据并完成基于开始 p 个分组的大小的连接表示，然后校准聚类算法。这一步主要是搜索得到最好聚类的簇数和分组数。最后运行聚类算法，得到具有相似行为的 TCP 连接分组。该方法评价了 K-Means，GMM[24]及基于 HMM 谱聚类[25]3 种著名算法的性能。研究结果表明，虽然 HMM 谱聚类法代表的信息丰富，但聚类效果与基于简单欧氏距离空间的 GMM 法相当。训练阶段输出两个集合：一个包含每个簇的描述，另一个为每个簇中出现的应用。

(2) 分类阶段，分类器将两个方向的分组头序列作为输入。解析和转换模型抽取五元组信息和分组大小，分析器过滤掉控制流(TCP 握手阶段的 3 个分组)并存储每个流方向上的分组的大小，当得到某个连接的前 p 个分组时，将信息发送给分配模型，该模块基于簇描述分配连接所属的簇，由它确定一个新的 TCP 连接是否属于预定义的簇，不属于任何簇的连接标记为未知。这种方法允许分类器检测新的应用或已知应用的新的操作模型。最后，由标记模块基于该簇成分选择与连接有关的最可能的应用。

实验结果表明，一个 TCP 连接的前 4 个分组足以用来识别应用。对于已知应用识别准确度超过 90%，识别新的作为未知应用的概率为 60%。该作者在文献[26]中用类似的方法识别 SSL 加密连接中的应用，利用握手阶段前几个分组实现早期识别应用，准确率在 85%。

方法特点：最大的特色是选用的特征简单，并考虑到在线早期应用识别，为应用的跟踪、分析控制提供了支持。然而，此方法也有许多的缺陷[23]：需要考虑分组的顺序，而实际观察到的分组可能不是原来的顺序；另外，不同的应用可能具有类似的行为；不支持对数据抽样操作；通过对分组内容充填很容易骗过分类器；不能对 UDP 流量进一步分类。

3.3 基于 K-Means 半监督的流量分类

设计网络流量识别模型面临两个主要问题：第一，有标记的样本是很少的，且难以获得。用少量有标记的样本，传统的有监督学习方法不能识别以前未出现过的流。第二，不是所有产生流的应用类型都是预先知道的。传统的有监督学习方法强制性地将每个流映射到 q 个已知类别，不能检测新的流类型。半监督学习是处理大量的无标记数据和少量有标记数据的一种学习分类方法，适用于大量数据不断产生，同时对这些数据不容易标记或者花费代价太大。

Jeffrey Erman[22]首次提出了基于流的半监督学习的流量识别方法。作者开始选用了 25 个流候选特征，然后使用向后贪婪特征选择方法[30]对特征进行了优选，得到 11 个特征。选择 K-Means 算法识别应用(HTTP，P2P，EMAIL，FTP，P2P，Encryoted，STREAMING，DATABASE，CHAT 等)。

主要思想：将有监督和无监督方法结合，设计了一种新的分类方法——半监督学习方法[31]。此方法由两个阶段构成：利用聚类算法[24]划分训练集，此训练集由少量有标识的流及大量未标识的流组成；利用有效的已标识流来获得从簇到不同已知 q 类别(Y)的映射，此阶段可能存在没有被映射的流，说明可能存在没有标记的流。学习的结果是一个划分集。具体步骤如下：

(1)聚类。首先将有效的训练集进行聚类。利用 $d(X_i, X_j)$度量向量 X_i 与 X_j 之间的相似性，得到划分(相似的样本出现在相同的簇中，不相似的样本出现在不同的簇中)。有多种相似性度量，该作者使用欧氏距离作为度量：

$$d(X_i, X_j) = [\sum_{k=1}^{m}(x_{ik} - x_{jk})^2]^{1/2} \tag{2}$$

可以选择不同的聚类算法，基于之前对算法的研究与比较[21]，此处作者选用了 K-Means 算法进行聚类。

(2) 簇到应用的映射。K-Means 算法的输出是一个簇集，r_k 代表簇的中心。对于给定的流特征向量 X_i，利用式(3)将其分配给某个簇(通过寻找离 X_i 最近的中心)。

$$C_k = \arg\min_k d(X_i, r_k) \tag{3}$$

其中 $d(.,.)$是在聚类阶段中选择的欧氏距离度量。式(3)相当于最大似然法聚类分配方案。

利用概率分配找到聚类到标记的映射：给定一个数值 C_k，要确定 C_k 属于类 y_j 的概率，记为 $P(Y=y_j \mid C_k)$，称作后验概率，其中 $j=1,\cdots,q$(q 是应用类型数量)，$k=1,\cdots,K$(K 是簇数量)。通过最大似然法估计 $P(Y=y_j \mid C_k)$。没有任何标识样本分配的聚类被定义为“未知”应用类型，代表以前没有识别的应用类型。

最后，利用最大后验概率函数做类别判决函数，即确定每一类的后验概率，然后将 C_k 分配到具有最高概率的类中：

$$y = \arg\max_{y_1,\cdots,y_q}(P(y_j \mid C_k)) \tag{4}$$

其中 C_k 是从式(3)获得的离 X_i 最近的簇。聚类企图形成不相交的组，在每一个组内部对象之间有强的相似性[24]，半监督方法正是基于这样的事实训练分类器。可以验证这样的假设：如果在每一个簇中，有少数的流已经被识别，则有合理的根据建立该簇到应用类型的映射。

基于以上方法，该作者分别实现了离线和实时的分类原型系统，并且实现了 Bro[32]入侵检测系统的实时分类，同时还研究了分类器的寿命及其重新训练技术问题。研究结果表明分类器在长时间内是适用的，当网络的使用模式有巨大变化(包括引入新的应用)时需要重新训练分类器。这种架构可以部署在网络边界和核心处[33]。

方法特点：通过少量有标记的流和大量无标记的流训练能够提高分类器的识别速度和精度；同时，算法具有较强的鲁棒性，能识别未知应用和行为有变化的已知应用。该作者为网络流量分类提出了一种新的方法思路，但有许多问题有待深入研究。例如，作者只用了一种聚类算法，缺乏与其它算法进行比较等。

3.4 机器学习识别法评价

有监督、无监督和半监督方法都有各自的局限。有监督方法需要大量高质量的有标记数据来构造分类器，而无监督方法通常需要消耗很多的计算资源，并且产生的类别并不一定与人的判断一致。半监督方法则是它们的折衷，通过利用一些有标记的数据来影响聚类过程，试图得到更好的聚类效果。

机器学习识别法由于选择的特征与识别算法不同，以及训练数据集的偏斜，使得性能上有差异，但总体上它们具有一些共同特性。在此分别从准确度、计算复杂度、扩展性及鲁棒

性等方面给予评价。

(1) 准确度:一个好的方法应该有较高的准确性、较低的误报率和漏报率。当前所研究的方法大多数只能对流进行粗粒度分类,且识别的种类有限。准确性的评价指标很不统一,研究者选用或自定义了多种度量对自己的方法进行评价。文献[34]建议用流准确度(Flow Accuracy)(正确分类的流数量占测试集中流总量的百分比)和字节准确度(Byte Accuracy)(正确分类的字节数量占测试集中字节总量的百分比)对方法性能进行评价。现有的基于流特征的识别法的流准确性都在70%左右,高于端口方法,低于载荷方法。

(2) 复杂度:为了能够满足高速网络,一个好的算法必须有较低的时间、空间复杂度。由于要进行多个特征分析,该类方法的复杂度高于端口方法,但是远低于载荷方法。

(3) 扩展性:方法的好坏还应该考虑它的扩展性。好的方法应该易于扩展,以适应网络快速发展变化特性。由于这类方法不需要知道协议或应用的具体细节,因此当协议或应用规范的细节发生变化时不需要更新相应的方法,只需要用新的数据集重新训练识别器。

(4) 鲁棒性:健壮的方法应该在一定的参数摄动下,能够维持其某些性能不变。由于某些流特征(例如分组到达时间、流持续时间等)对网络动态变化(例如网络拥塞)极其敏感,使基于这些特征的方法稳定性、准确性受到影响。

(5) 对加密流的识别:加密的载荷流数据使基于载荷的识别法失效。然而,根据流特征有可能识别加密流。文献[26,35]进行了这方面的探索研究,已经取得初步成效。

4 存在问题及未来研究方向

机器学习方法识别应用是目前的一个新的热点研究方向,虽然已经取得一些初步研究成果,但这方面的研究还很不成熟,许多问题有待进一步研究解决。

(1) 首先,缺乏评价方法的统一标准数据集(尤其是含载荷的数据集)及评价指标[36]。目前研究者多数采用收集的部分公开数据集(不含载荷)或自生成的数据与自定义的各种性能评价指标验证算法,给方法的测试评价与比较带来一定的困难。

(2) 缺少对流特征集选择、优化的系统研究。对流特征选择虽有研究[15,19],但多数算法只是简单地选择一种或多种流特征组合。选择或抽取的流特征是影响方法性能的一个重要因素。

(3) 学习算法只是几种经典算法的选用,缺乏算法之间的对比分析、新算法的设计。如何选择或设计新的识别算法,使其在特征集一定的情况下达到最好的识别效果,是影响方法性能的另一个重要因素。

(4) 很少考虑抽样问题。目前的研究多是基于所采集到的全部数据,但随着网络速度的不断提高,抽样是不可避免的。

(5) 已有方法识别的应用类型数量极其有限,且不精细。此外,许多方法是离线或在流结束时方能识别应用,它们不能用于实时的流跟踪、分析控制。

理想的识别方法应该具有高准确度、高效率、低开销,以实现准确实时在线的早期应用识别,同时具有良好的扩展性,且对网络噪声不敏感。综合上述分析,未来的研究方向可以有以下几个方面:

(1) 特征选择及优化问题研究。特征选择与优化的研究目标是能够找到区分度好、特征量少、计算简单及对抽样操作的支持,且能在流结束前尽早识别流的特征集。

(2) 最优识别算法的研究。研究目的是寻找或设计适合流特征的高效率的算法。根据网络流量特性,基于有监督和无监督的半监督学习算法将是一个非常有前景的研究探索方向。

(3) 组合方法的研究。如何更好地结合各类方法,使其优势互补,以达到精细地识别更多应用类型,是一个重要的研究方向。虽然已经有一些简单的组合思想[7,36],但研究还不够深入,可能需要更多的方法进行更复杂的集成。

对此,我们提出一种组合设计方案,将端口方法、流特征法和载荷的方法结合起来:首先利用基于端口的方法识别有固定端口的应用;对于未能识别的流利用基于流特征的方法对流进行粗分;然后,在每一类中使用基于载荷的方法标识具体应用。这样,既可以避免在大范围中使用基于载荷方法的计算开销,又可以增加基于流特征方法的精细度和准确性,实现了方法优势互补。目前这种组合方案作为网络应用个体行为分析与控制系统的组成部分正在研究当中。

参 考 文 献

[1] Cache Logic. Peer-to-Peerin2005, http://www.cachelogic.com/home/pages/research/p2p2005.php, 2005

[2] Sen S, Wang J. Analyzing Peer-to-Peer Traffic across Large Networks. IEEE/ACM Transaction Networking, 2004, 12(2): 219-232

[3] IANA. Internet Assigned Numbers Authority (IANA). http://www.iana.org/assignments/port-numbers

[4] Roughan M, Sen S, Spatscheck O, et al. Class-of-Service Mapping for QoS: A Statistical Signature-based Approach to IP Traffic Classification // IMC'04. Taormina, Italy, October 2004

[5] Sen S, Spatscheck O, Wang D. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures // WWW'04. New York, USA, May 2004

[6] Haffner P, Sen S, Spatscheck O, et al. ACAS: Automated Construction of Application Signatures // SIGCOMM'05. Philadelphia, USA, August 2005

[7] Moore A, Papagiannaki K. Toward the Accurate Identification of Network Applications // PAM'05. Boston, USA, March/April 2005

[8] Kumar S, Dharmapurikar S, Yu F, et al. Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection // SIGCOMM '06. Pisa, Italy, September 2006

[9] Ma J, Levchenko K, Krebich C, et al. Unexpected Means of Protocol Inference // IMC'06. Rio de Janeiro, Brasil, October 2006

[10] Karagiannis T, Broido A, Faloutsos M, et al. Transport Layer Identification of P2P Traffic // IMC'04. Taormina, Italy, October 2004

[11] Karagiannis T, Papagiannaki K, Faloutsos M. BLINK: Multilevel Traffic Classification in the Dark // SIGCOMM'05. Philadelphia, USA, August 2005

[12] Mena A, Heidemann J. An Empirical Study of Real Audio Traffic // Infocom'03. Tel-Aviv, Israel, March 2000

(下转第38页)

Conf. on Management of Data. June 2003

[4] Cheng R, Xia Y, Prabhakar S, et al. Efficient indexing methods for probabilistic threshold queries over uncertain data// Proc. of the 30th Intl. Conf. on Very Large Data Bases. 2004

[5] Cheng R, Xia Y, Prabhakar S, et al. Efficient join processing over uncertain-valued attributes// CIKM'06. November 2006

[6] Chau M, Cheng R, Kao B. Uncertain Data Mining: A New Research Direction // Invited Paper, the Workshop on the Sciences of The Artificial (WSA) 2005. National Dong Hwa University, Taiwan, Dec. 2005

[7] Chau M, Cheng Reynold, Kao B, et al. Uncertain Data Mining: An Example in Clustering Location Data // the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference(PAKDD 2006). Singapore, April 2006

[8] Wang Kay Ngai, Ben Kao, Chun Kit Chui, et al. Efficient Clustering of Uncertain Data // the Proc. of the Sixth International Conference on Data Mining (ICDM'06). 2006

[9] Sau Dan Lee, Ben Kao, Reynold Cheng. Reducing UK-means to K-means// the Proc. of the 1st Workshop on Data Mining of Uncertain Data (DUNE2007). 2007

[10] Acharya S, Alonso R, Franklin M, et al. Broadcast disks: Data management for asymmetric communications environments // Proc. of the ACM SIGMOD Intl. Conf. on Management of Data. May 1995

[11] Acharya S, Muthukrishnan S. Scheduling on-demand broadcasts: New metrics and algorithms// Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'98). October 1998

[12] Vaidya N H, Hameed S. Data broadcast in asymmetric wireless environments// Workshop on Satellite Based Information Services(WOSBIS). Nov. 1996

[13] Acharya S, Franklin M, Zdonik S. Balancing push and pull for data broadcast// Proc. of the ACM SIGMOD Intl. Conf. on Management of Data. May 1997

[14] Govindan R, Hellerstein J, Hong W, et al. The sensor network as a database. Technical Report, 02-771. Computer Science Department, University of Southern California, 2002

(上接第 23 页)

[13] Dewes C, Wichmann A, Feldmann A. An analysis of Internet chat systems// IMC '03. Miami Beach, FL, USA, October 2003

[14] ez-Campos F H, Smith F D, Jeffay K, et al. Statistical Clustering of Internet Communications Patterns. Computing Science and Statistics, July 2003, 35

[15] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques// SIGMETRICS'05. Banff, Alberta, Canada, June 2005

[16] Crotti M, Dusi M, Gringoli F, et al. Traffic Classification through Simple Statistical Fingerprinting. ACM SIGCOMM Computer Communication Review, January 2007

[17] Williams N, Zander S, Armitrage G. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, October 2006

[18] McGregor A, Hall M, Lorier P, et al. Flow ClusteringUsing Machine Learning Techniques// PAM '04. Antibes Juan-les-Pins, France, April 2004

[19] Zander S, Nguyen T, Armitage G. Automated Traffic Classification and Application Identification using Machine Learning // LCN'05. Sydney, Australia, November 2005

[20] Bernaille L, Teixeira R, Salamatian K. Early Application Identification// CoNEXT'06. Lisboa, Portugal, December 2006

[21] Erman J, Arlitt M, Mahanti A. Traffic Classification Clustering Algorithms// SIGMETRICS'06 (MineNet). Pisa, Italy, September 2006

[22] Erman J, Mahanti A, Arlitt M, et al. Offline / Online Traffic Classification Using Semi-supervised Learning. Technical report. University of Calgary, 2007

[23] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26

[24] Duda R O, Hart P E, Stork D G. Pattern Classification. Second edition. Wiley, 2001

[25] Bernaille L, Soule A, Jeannin M-I, et al. Blind application flow recognition through behavioral classification. Technical report. Laboratoire d'Informatique de Paris 6, Universit'e Pierre et Marie Curie, 2005. http://www-rp. lip6. fr/ bernaill/techreport. pdf

[26] Bernaille L, Teixeira R. Early Recognition of Encrypted Applications// PAM'07. Louvain-la-neuve, Belgium, April 2007

[27] Moore A W, Zuev D. Discriminators for use in flow-based classification. Technical report. Intel Research, Cambridge, 2005

[28] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution// ICML'03. Washington DC, USA, August 2003

[29] Hongbo J, Moore A W. Lightweight Application Classification for Network Management // INM' 07. Kyoto, Japan, August 2007

[30] Guyon I, Elisseef A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003, 3: 1157-1182

[31] Chapelle O, Scholkopf B, Zien A. Semi-supervised Learning. Cambridge, MA: MIT Press, 2006

[32] Paxson V. Bro: A System for Detecting Network Intruders in Real-time. Computer Networks, 1999, 31(23/24): 2435-2463

[33] Erman J, Mahanti A, Arlitt M, et al. Identifying and Discriminating Between Web and Peer-to-Peer traffic in the Network Core// WWW'07. Banff, Canada, May 2007

[34] Erman J, Mahanti A, Arlitt M. Byte Me: A Case for Byte Accuracy In Traffic Classification // SIGMETRICS'07 (MineNet). San Diego, CA, June 2007

[35] Crotti M, Dusi M, Gringoli F, et al. Detecting HTT PTunnels with Statistical Mechanisms // CC'07. Glasgow, Scotland, June 2007

[36] Salgarelli L, Gringoli F, Karagiannis T. Comparing Traffic Classifiers. ACM SIGCOMM Computer Communication Review, 2007, 37(3): 65-68