第43卷 第2期 西 安 交 通 大 学 学 报 Vol. 43 №2
2009年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2009

利用统计特征的网络应用协议识别方法

徐莉[1]，赵曦[2]，赵群飞[1]，秦涛[3]，
(1. 上海交通大学图像处理与模式识别研究所，200240，上海；2. 上海金融学院信息管理系，201209，上海；
3. 西安交通大学电子与信息工程学院，710049，西安)

摘要：在网络流统计特征的基础上，提出了一种应用协议识别算法. 根据网络流概念，在网络层建立应用协议特征的描述方法，并采用数据量、数据包、时间3种属性全面地描述网络协议的特征. 采用主成分分析方法来确定网络流特征属性的主要成分，以减少环境因素的影响. 结合BP神经网络算法建立的网络协议识别模型，其网络特征具有良好的持久性和稳定性，模型分类结果也不易受网络环境的影响. 真实网络环境下的实验结果显示，所提方法能够准确识别目前网络中常用的应用协议，包括HTTP、FTP、BitTorrent及TELNET，识别准确率达到了97%以上.

关键词：神经网络模型；网络流；应用协议；协议识别
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2009)02-0043-05

Network Application Protocol Identification Based on Statistical Methods

XU Li[1], ZHAO Xi[2], ZHAO Qunfei[1], QIN Tao[3]
(1. Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China; 2. Department of Information Management, Shanghai Finance College, Shanghai 201209, China; 3. School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: An application protocol identification method is proposed based on the statistical characteristics of network flows. The flow features at the network level are extracted according to the concept of network flow, and three attributes: the number of packets, the number of bytes, and time, are used to capture the flow characteristics roundly. Then the principal component analysis algorithm is used to determine the main characteristics of the flow attributes to reduce the effect of environment. Finally, a BP neural network model is given to identify the application protocols. As the features used in the proposed method are more stable, the output results of the model are hence accurate with the change of the network environment. Experimental results in real network environments show that the proposed method can identify several major application protocols accurately, such as HTTP, BitTorrent, FTP and TELNET, and the identification precision is above 97%.

Keywords: neural network model; traffic flow; application protocol; protocol identification

随着信息化大潮的到来，计算机网络应用日益推广，应用协议种类逐渐增加，加之网络代理和端口的动态变化策略的采用，使网络管理愈加困难. 网络应用协议的准确识别可以实现网络运行状况的准确监控和网络资源的有效调配，同时也可提高网络入侵检测的效率和准确性，改善网络服务质量，确保网络安全、高效运行.

最简单的协议识别方法是通过分析网络数据包的端口信息来分析HTTP协议[1]，而现在广泛应用的P2P协议所采用的通信端口都是非知名、动态变

收稿日期：2008-04-21. 作者简介：徐莉(1985—)，女，硕士生；赵群飞(联系人)，男，教授. 基金项目：国家自然科学基金资助项目(60575049)；国家高技术研究发展计划资助项目(2003AA142060).

化的,因此基于端口的网络协议识别技术逐渐失去效力.通过对数据包负载的分析也可实现网络应用协议的划分,这种方法虽然识别准确率较高,但计算复杂[2].为了适应 Internet 流量数据庞大、应用属性动态变化的特点,根据数据包头统计特征进行应用协议识别已成为当前计算机网络领域一个新的研究热点.数据包头的统计特征可以用网络流描述,它是一组具有相同五元组(源 IP、目的 IP、源端口、目的端口、传输层协议)的数据包序列,如图 1 所示.

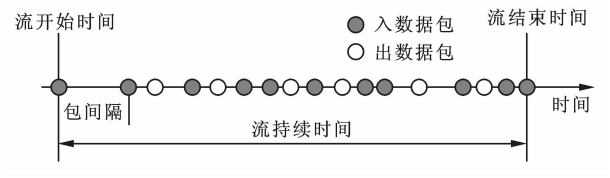


图 1 网络流描述

网络流由一系列流量属性构成,包括数据包大小、数据包方向、数据包间隔和网络流持续时间等.Moore 等人通过直接测量和利用傅立叶变换技术给出了网络流的 249 种属性[3],试图为协议分类研究提供一个完整的属性集合.

在网络流特征属性获取的基础上,国内外学者在协议识别方面也做了大量的研究工作,如引入基于概率模型的朴素贝叶斯方法[3],用聚类方法处理协议识别问题[4],将支持向量机方法应用于 P2P 流的识别领域[5],采用高斯滤波方法实现网络协议识别[6]等.上述方法虽然在一定程度上解决了基于统计特征的应用协议识别问题,但是所获取的网络流特征属性随着网络环境的变化而变化,缺乏持久性和稳定性,所以其广泛应用受到了限制.

针对以上问题,本文提出了一种网络协议识别方法.

1 网络流特征提取

根据流量统计特征实现协议识别的关键是网络流量属性集的获取.Moore 等人给出了双向 TCP 流的 249 种属性[3],然而在实际的网络环境中,随着网络带宽的不断增加,待分类的网络流数目通常达到了数十万乃至数百万条,过多网络流导致计算负载过于沉重,而且在所提取的网络流属性中,由于众多网络流属性相互关联,使得分类模型的识别率较低,难以满足现有网络管理的需要.为此,以利于网络管理为基础,以属性易于获取和分析为出发点,本文利用 11 种属性来描述流量数据中双向 TCP 流的主要特征,各属性的具体说明如表 1 所示.

表 1 网络流特征属性及描述

属性	网络流特征	特征描述
$\boldsymbol{x}_1$	存活时间	网络流持续的时间
$\boldsymbol{x}_2$	流方向	流入或者流出监控端
$\boldsymbol{x}_3$	入包个数	监控端接收的数据包数目
$\boldsymbol{x}_4$	出包个数	监控端发出的数据包数目
$\boldsymbol{x}_5$	入字节量	监控端接收到的数据量
$\boldsymbol{x}_6$	出字节量	监控端发送的数据量
$\boldsymbol{x}_7$	最大数据包	包含的最大数据包大小
$\boldsymbol{x}_8$	包大小均值	传送数据包的平均大小
$\boldsymbol{x}_9$	最大包间隔	最大的数据包时间间隔
$\boldsymbol{x}_{10}$	包间隔均值	数据包间隔均值
$\boldsymbol{x}_{11}$	出入字节比	接收和发送的数据量之比

2 应用协议识别算法

2.1 协议识别方法框架描述

本文的网络协议模型共分为网络流特征数据采集、网络流主要属性信息提取、BP 神经网络模型建立以及测试验证等 3 个步骤,其流程框架如图 2 所示,具体描述如下.

步骤 1:根据网络流的定义,将网络数据包汇聚为网络流,提取网络流的特征.

步骤 2:采用主成分分析(PCA)方法提取网络流属性的主要特征.

步骤 3:根据提取出的网络流主要特征,结合 BP 神经网络,构建网络应用协议识别模型.

2.2 网络流主要属性的提取

网络应用的每一次交互行为都可以用一个包含 11 种属性的特征向量来描述,将其分别表示为 $\boldsymbol{x}_1$,$\boldsymbol{x}_2$,…,$\boldsymbol{x}_{11}$,对应关系详见表 1.这样,网络用户的行为集合就可以用一个 $n\times 11$ 的矩阵 $\boldsymbol{X}$ 来表示

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{x}_{1,1} & \boldsymbol{x}_{1,2} & \cdots & \boldsymbol{x}_{1,11}\\ \boldsymbol{x}_{2,1} & \boldsymbol{x}_{2,2} & \cdots & \boldsymbol{x}_{2,11}\\ \vdots & \vdots & & \vdots\\ \boldsymbol{x}_{n,1} & \boldsymbol{x}_{n,2} & \cdots & \boldsymbol{x}_{n,11}\end{bmatrix}=[\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{11}] \quad (1)$$

式中:$\boldsymbol{x}_{i,j}$ 表示第 j 个流量属性在第 i 条网络流样本上的观察值;n 为网络流的条数.然后,采用 PCA 方法实现网络流主要特征的提取[7].λ_1,λ_2,…,λ_{11} 是协方差矩阵的特征值序列,$\boldsymbol{u}_1$,$\boldsymbol{u}_2$,…,$\boldsymbol{u}_{11}$ 是每个特征值分别对应的特征向量,其中特征值满足 $\lambda_1>\lambda_2>\cdots$

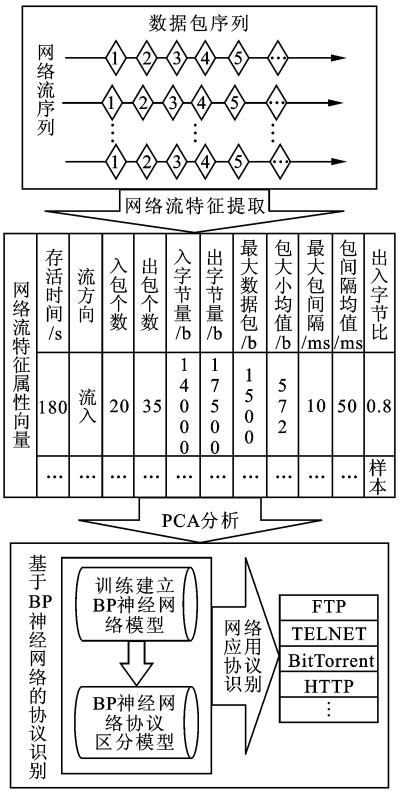


图 2　协议识别流程框架图

$>\lambda_{11}$. 通过方差贡献率 α 可衡量不同特征值和特征向量在重构原始空间时的重要性，通常 $\alpha\geqslant 80\%$. 满足式(2)的 m 值对应的特征向量构成一个特征子空间，它继承了原始数据集的主要特征，而余下的 $11-m$ 维特征属性主要是包含着噪声.

$$\alpha \leqslant \sum_{i=1}^{m}\lambda_i / \sum_{i=1}^{11}\lambda_i \tag{2}$$

2.3　网络应用协议识别模型

BP 神经网络是一种多层前馈型神经网络，它包含输入层、隐含层和输出层，可以逼近任意连续函数，广泛应用于模式分类并取得了很好的结果[8]. 根据协议识别需要，可建立(I,J,K)的协议识别网络模型，其中 I 为输入层神经元数，J 为隐含层神经元数，K 为输出层神经元数. 隐含层神经元数 J 采用经验式确定[9]，即

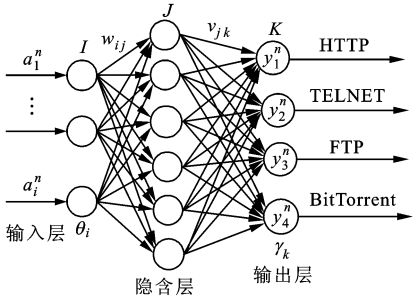
$$J=(K+I)^{1/2}+H \tag{3}$$

式中：H 为介于$[1,10]$之间的整数. 经过实验对比，选择隐含层节点数为 6，可以取得理想的实验结果. 将通过 PCA 算法提取后的特征属性 $P_n=(a_1^n,\cdots,a_i^n)$作为神经网络的输入，其中的 n 表示训练集中的第 n 条网络流记录，$a_i^n(i=1,\cdots,I)$为提取后的网络流特征属性. 本文根据实验结果选取 $I=3$，输出层神经元节点数为 4，期待输出值 $Y_n=(y_1^n,\cdots,y_4^n)$，期待输出值与应用协议间的对应关系如表 2 所示.

表 2　神经网络期待输出与应用协议对应关系

应用协议	Y_n			
	y_1^n	y_2^n	y_3^n	y_4^n
HTTP	1	0	0	0
FTP	0	1	0	0
BitTorrent	0	0	1	0
TELNET	0	0	0	1

据此，网络协议识别的 BP 网络结构如图 3 所示.



w_{ij}：输入层与隐含层的连接权重；v_{jk}：隐含层与输出层的连接权重；θ_i、γ_k：输入层、输出层神经元阈值

图 3　基于神经网络的协议识别模型

3　实验结果

3.1　网络数据收集

为了验证本文方法，在实验室环境下采用被动监听的方式采集网络流量. 图 4 所示是网络流量数据采集平台的拓扑结构图，其中电脑终端是监控对象，它们经过路由器连接至 Internet，在路由器处设置镜像服务器，用于采集流经路由器的网络流量.

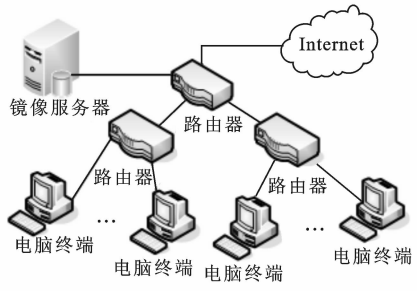


图 4　网络流量收集系统拓扑图

对收集到的网络流量，根据网络流定义将数据包汇聚为网络流并提取 11 种网络流特征. 采用特征

字段和网络端口相结合的方式标定应用层所采用的应用协议. 在本文实验中,共收集了 HTTP、FTP、BitTorrent 及 TELNET 4 种网络应用数据,数据情况如表 3 所示. 鉴于建立网络应用节拍模式的需要,实验中提取的网络流为完整的 TCP 交互流,即数据集不包含数据包个数少于 3 的网络流. 随机抽取每种协议网络流的 50%作为训练网络流集合,并建立网络应用协议识别模型,其余数据集将作为测试网络流集合,最后测试所建模型的准确性.

表 3 采集到的网络应用协议数据

应用协议	HTTP	FTP	BitTorrent	TELNET
流数	2 724	388	768	1 062
样本比	0.551	0.079	0.155	0.215
训练数	1 362	194	384	531

3.2 PCA 分析结果

实验中收集到的每一条网络流数据都是一个 11 维的向量,n 条流的特征数据就形成了 $n\times11$ 的特征矩阵. 采用 PCA 分析方法寻找网络流属性的主要特征,由式(3)计算所得的不同应用协议的方差贡献率,结果如图 5 所示. 由图 5 可知,FTP 应用协议和 HTTP 应用协议的方差贡献率在属性为 3 维时就已经超过了 80%,BitTorrent 和 TELNET 应用协议的方差贡献率在属性为 3 维时分别接近和超过 70%,可见网络应用协议特征可以用 3 维数据有效描述.

经过 PCA 分析认为,可以用 3 维特征有效地描述原始网络流特征的主要属性. 图 6 给出了提取出的主要属性在 3 维空间中的分布特征. 由图 6 可以看出,经过特征提取处理后,网络应用协议特征数据在 3 维空间中呈现不同的特性,相同的应用协议在 3 维特征空间中聚集为相同的簇,表现出了很强的相似性,具有良好的可区分性. 这为识别应用协议奠定了有效的基础.

3.3 BP 神经网络识别结果

经过 PCA 分析认为,可以将原有的 11 维网络流特征映射到 3 维空间. 随机选取网络数据集的 50%作为训练样本集,并建立 BP 神经网络模型,再利用其余数据集作为测试样本来验证本文算法的准确性. 在本文中,选取神经网络系统误差为 0.01,协议识别模型各步的误差曲线如图 7 所示. 由图 7 可以得出,经过近 4 次迭代,协议识别模型达到规定的系统误差,经过 10 次迭代后,识别模型趋于收敛.

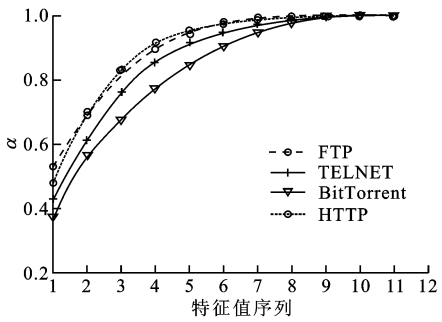


图 5 应用协议的低维方差贡献率

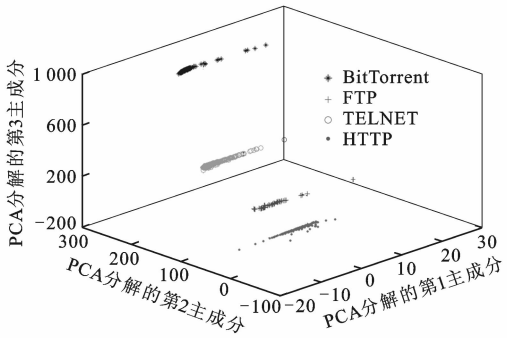


图 6 降维处理后不同应用协议特征的分布

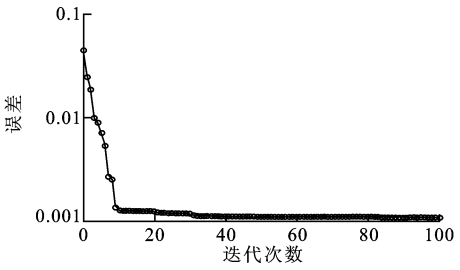


图 7 协议识别模型各步的系统误差

利用测试数据集得到的测试结果如表 4 所示. 由表 4 结果可以看出,本文模型可以准确实现网络流协议类型的划分. 根据原始网络流特征,仅采用 BP 神经网络识别算法的协议识别结果如表 5 所示. 结合表 4、表 5 可以看出,单独使用 BP 神经网络算法的识别准确率要低于 PCA 与 BP 相结合的协议识别算法. 这是因为,所提网络流的 11 类特征数据之间存在着一定的关联性,由此影响了识别准确率.

表 4 PCA 结合 BP 网络协议识别算法的测试结果

应用协议	HTTP	FTP	BitTorrent	TELNET
流数	1 362	194	384	531
识别数	1 341	190	375	523
识别准确率/%	98.5	97.9	97.6	98.5

表 5　BP 网络协议识别算法的测试结果

应用协议	HTTP	FTP	BitTorrent	TELNET
流数	1 362	194	384	531
识别数	1 251	180	365	503
识别准确率/%	91.8	92.7	95	94.7

经过 PCA 降维处理后，去除了原始数据之间的关联性，减少了网络流量属性的记录数量，这不但提高了 BP 识别模型的识别准确率，也降低了计算复杂度.

4　结束语

网络应用协议的有效识别是近几年网络安全领域的研究热点，也是提高网络管理效率的基础. 本文提出了一种网络流特征刻画方法，即利用网络层参数描述网络应用协议的特征模式，采取 PCA 方法提取网络流特征的主要属性，最后结合 BP 神经网络建立不同网络应用协议识别模型. 真实网络环境下的实验结果表明，本文方法能够有效识别网络应用协议的类型. 我们的下一步工作是研究基于流量分类方法的实时应用系统.

参考文献：

[1] MOORE D, KEYS K, KOGA R, et al. The CoralReef software suite as a tool for system and network administrators [C]// Proceedings of the 15th USENIX Conference on Systems Administration. Berkeley, CA, USA: USENIX Association, 2001: 133-144.

[2] ROESCH M. SNORT: lightweight intrusion detection for networks[C]// Proceedings of the 13th USENIX Conference on Systems Administration. Berkeley, CA, USA: USENIX Association, 1999: 229-238.

[3] ZUEV D, MOORE A W. Traffic classification using a statistical approach[C]// 6th International Workshop on Passive and Active Network Measurement. Berlin, Germany: Springer-Verlag, 2005: 321-324.

[4] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms [C]// Proceedings of ACM SIGCOMM Workshop on Mining Network Data. New York, USA: ACM, 2006: 281-286.

[5] 王锐，王逸欣，樊爱华，等. 一种跨层 P2P 流量检测方法 [J]. 计算机应用，2006，26(12)：30-32.
WANG Rui, WANG Yixin, FAN Aihua, et al. A P2P traffic detection method between levels [J]. Computer Applications, 2006, 26(12): 30-32.

[6] CROTTI M, DUSI M, GRINGOLI F, et al. Traffic classification through simple statistical fingerprinting [C]// Proceedings of ACM SIGCOMM Computer Communication Review. New York, USA: ACM, 2007: 5-16.

[7] DUDA R O, HART P E, STORK D G. Pattern classification [M]. 2nd ed. Beijing, China: China Machine Press, 2004: 576-580.

[8] 程洪，郑南宁，高振海，等. 基于主元神经网络和 *K*-均值的道路识别算法[J]. 西安交通大学学报，2003，37(8)：812-815.
CHENG Hong, ZHENG Nanning, GAO Zhenhai, et al. Road recognition algorithm using principal component neural networks and *K*-means [J]. Journal of Xi′an Jiaotong University, 2003, 37(8): 812-815.

[9] HECHT-NIELSON R. Neuron computing [M]. Reading, MA, USA: Addsion Wesley, 1990: 124-125.

（编辑　苗凌）

西安交通大学两团队获 2008 年度国家级教学团队荣誉称号

日前，教育部公布了 2008 年度国家级教学团队名单，西安交通大学 2 个教学团队榜上有名，获得国家级教学团队荣誉称号. 这 2 个团队分别是电气学院王兆安教授作为带头人的电力电子与新能源技术研究中心教学团队和理学院马知恩教授作为带头人的大学数学系列课程教学团队.

本次评比经过学校推荐、省级评审、专家评审等程序，全国共评选出 300 个国家级教学团队.

（来源：交大新闻网 http://xjtunews.xjtu.edu.cn）