基于小波支持向量机的P2P网络流量识别算法

刘 悦，郭拯危
（河南大学 计算机与信息工程学院，河南 开封 475004）

摘 要：对等网络技术引起了广泛关注，其典型的应用有文件共享、即时通信等。为了更好地合理使用、规划P2P网络资源，建立P2P流量识别模型具有十分重要的理论意义和现实价值。提出了一种基于小波支持向量机相结合的P2P流量识别模型，将小波分析中多尺度的学习方法和SVM的优点结合起来，通过小波分析与SVM方法紧致结合，引入满足小波构架和Mercer定理的小波基函数来构造SVM的核函数，建立小波支持向量机的P2P识别算法。实验结果表明该算法能够有效地提高P2P网络流量识别的精度。

关键词：支持向量机；小波；P2P；网络流量
中图分类号：TP393 **文献标识码**：A **文章编号**：1673-629X(2010)10-0107-04

Algorithm for P2P Network Traffic Identification Based on Wavelet SVM

LIU Yue, GUO Zheng-wei
(Computer Science and Technology College, Henan University, Kaifeng 475004, China)

Abstract: Recently, there has been a growing interest in the potential use of peer to peer computing in many applications such as file sharing, instant communication. Therefore, to realize their potential, there is a need of a P2P traffic identification algorithm that facilities the deployment of a network traffic that is optimized in terms of network bandwith. Focuses on developing a novel P2P traffic prediction using wavelet support vector machine. Through the wavelet analysis combined with the SVM method of compact, introduced to meet the wavelet framework and Mercer theorem to construct the wavelet function SVM kernel function, wavelet support vector machines to establish P2P identification algorithm. Experimental results show that the algorithm can effectively improve the accuracy of P2P network traffic identification.

Key words: SVM; wavelet; P2P; network traffic

0 引 言

对等网络计算技术（Peer to Peer Computing，P2P），作为一种新型的网络通信模式，已经被列为影响未来Internet发展的科技技术之一，与网格计算技术（Grid Computing），云计算技术（Cloud Computing）成为分布式计算技术领域的研究热点。P2P的思想改变Internet原来的C/S（Client/Server Computing）不对称的计算模式，每个节点既是资源的提供者（Server），又是资源的使用者（Client），这为大规模的资源共享、互相通信和协同工作提供了灵活和可扩展的计算平台。

P2P的日益广泛应用也带来了许多的负面问题，体现在：（1）消耗大量的网络带宽资源，据统计一些P2P网络流量在Internet上的流量占到50%以上[1,2]，易导致其他应用资源的紧张；（2）安全问题，某些P2P应用软件可以穿过现有的防火墙和安全代理[3]，从而打开个人或企业用户的网络安全防护漏洞，导致个人和企业私密泄露。深入研究P2P网络流量的特征，选取适当的识别模型，进而高效的对P2P网络流量进行识别，及时地采取对策，对P2P网络流量进行有效的控制具有非常重要的理论意义和现实价值。

1 研究现状

文献[4]针对P2P网络流量的识别方法提出了深度报文检测方法，它的识别准确率高，并能识别出具体的P2P协议，缺点是无法识别加密或未知的P2P流量，而且需要对特征码进行逐个匹配，实时性较差；文献[5]提出了基于流特征和基于主机[6]的识别方法，文献[7]采用这种方法对常见的流媒体应用进行了识别，

收稿日期：2010-03-01；**修回日期**：2010-06-08
作者简介：刘 悦（1977-），男，河南开封人，助理工程师，硕士研究生，研究方向为P2P网络、网络流量；郭拯危，教授，研究方向为计算机网络、信息安全。

这种识别方法识别率较高,但是无法识别具体的P2P应用类型。文献[8]使用了基于流特征和应用层字符串匹配的双层特征识别方法,具有较高的识别率,但不能做到实时检测。文献[9]提出了使用朴素Bayesian分类的方法对流量进行识别,提高了识别的精度,但是对其特征以及条件的独立性假设并不很准确,其给出的特征之间不满足独立的条件。文献[10]研究了基于BP神经网络技术的P2P流量识别研究,分析P2P流量特征,构建BP网络,通过该网络的足够训练,得到相关的测试结果。但是神经网络的层数和神经元个数需要人工经验,难以确定,容易陷入局部极小,易出现过学习现象。文献[11]提出了使用SVM的方法进行P2P流量识别,SVM是针对小样本的机器学习方法,它基于结构风险最小化原理,通过解决凸二次优化问题得到全局最优解,其具有较高的推广能力和鲁棒性,但是SVM只是在一个尺度上对样本数据进行分类,对多尺度样本的逼近性能并不能令人满意。

文中提出了一种新的基于小波函数和支持向量机相结合的P2P流量识别模型,将小波分析中多尺度的学习方法和SVM的优点结合起来,通过小波分析与SVM方法紧致结合,引入小波基函数来构造SVM的核函数,建立小波支持向量机的P2P识别算法。小波函数具有良好的多尺度学习性能,容易确定网络参数,已经在实际中得到了广泛的应用,此模型能够多尺度的使SVM对识别P2P流量样本逼近,自适应处理P2P流量的非线性变化特征,直至达到要求的精度,而且计算量没有显著增加,具有良好的识别效果。

2 小波支持向量机识别模型

2.1 支持向量机

支持向量机(Support Vector Machine, SVM)是Vapnik等学者基于统计学原理提出的一种机器学习的方法。该方法具有较强的泛化能力,不会出现局部极小和维数灾难等优点,在支持向量机分类算法中,通过核函数将线性不可分数据 x 映射到高维线性可分的特征空间,然后进行有效的处理。所以核函数的选择是影响支持向量机分类效果的关键之一。文献[12]表明,系数变化的核函数有助于提高模型的精度和迭代的收敛速度;另一方面,如果对平滑函数缺乏先验知识,多尺度差值方法是最好的。这两点正是小波所特有的性质,所以以小波函数为核函数的支持向量机具有更高的函数逼近精度和泛化能力。

支持向量机发展与线性可分情况下的最优构造分类超平面,图1表明了它的基本思想。

图1中,两类样本分别使用实心点和空心点来表示,分类超平面用 H 来表示,各类样本距离超平面最近的样本用 H_1、H_2 来表示,并且和超平面的直线平行,H_1 和 H_2 之间的距离称为分类间隔。最优分类超平面就是不仅能将两类样本正确分开,而且使分类间隔最大的超平面。下面的方程就是分类超平面方程:

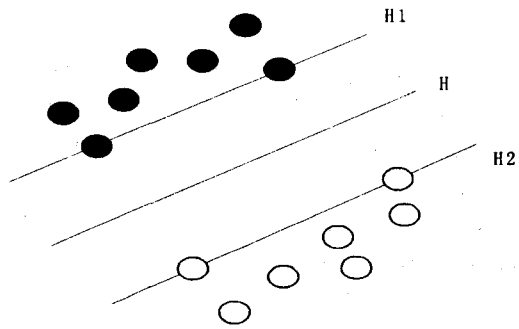

图1 支持向量机最优分类超平面示意图

$$w \cdot x + b = 0$$

通过对它归一化,这样,线性可分的样本集$(x_i, y_i), i = 1,\cdots,n, y \in [-1, +1]$,满足

$$y_i[(w \cdot x_i) + b] - 1 \geqslant 0, i = 1,2,\cdots,n \quad (1)$$

分类间隔值为 $2/\|w\|$,为了使间隔最大,等同于使 $\|w\|$ 最小或 $\|w\|/2$ 最小。使 $\|w\|/2$ 最小的超平面而且达到条件的最小的超平面就为最优分类超平面,支持向量就是处在 H_1、H_2 上的训练样本点。使用Lagrange乘子方法解这个约束最优问题,即在约束条件 $\sum_{i=1}^{n} a_i y_i = 0 \quad a_i \geqslant 0, i = 1,\cdots,n$,为Lagrange乘子,$i = 1,2,\cdots,n$,然后求解下列目标函数 $Q(a)$ 的最大值:

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) \quad (2)$$

这是一个不等式约束下二次函数寻优的问题,存在唯一解。a_i 不为零的解所对应的样本就是支持向量。解上述问题后得到的最优决策函数是:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*) \quad (3)$$

对于非线性问题,只需要将输入向量非线性映射到一个更高维的特征空间,若 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$,则称 $K(x_i, x_j)$ 为内核函数,一个函数是否是核函数的条件是由Mercer定理给出的。相应的最优决策函数变为:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*) \quad (4)$$

2.2 特征向量的选取

特征作为向量的一维数据,特征选取就是能够充分体现出P2P和非P2P流量的区别,常规相关以六元组标示的流相关的统计特征、平均流持续时间、每条流

的平均字节数、每条流的平均包数目、包到达时间间隔等参数,也可以对 P2P 流量进行识别,但是不能使用过多的特征信息,过多的特征会使支持向量机分类器的计算时间大大增加,还会降低分类器的性能。因此需要在保持支持向量机性能可以接受的情况下,尽可能地降低特征维数,通过对 P2P 流量的特征进行分析,可以从数据包、数据流和连接层面进行显效的 P2P 流量特征分析,从而形成三维相对特征向量:

1)数据包大小变化的均方差值:当流的数据包的大小有变化并且变化的跨度比较大时,每个流的包大小的均方差值就很大,而且流的持续时间比较长。P2P 流独特的包大小格式来源于 P2P 文件共享协议。P2P 文件共享协议分为两个步骤:信令和数据传输。目前的 P2P 系统提供并行文件的下载机制。它允许一个文件被分割成许多小块,进行传输的数据包从小到大。因此,在 P2P 流与其他应用流量相比在数据包的大小变化上有明显的不同,P2P 流的数据包大小变化的均方差值比常见的 web 流量要大。

2)上行流量和下行流量的比值:P2P 网络的一个重要特性就是每个节点在从其它节点下载的同时,也在提供上传服务,每个节点的上行流量与下行流量都是大体对称的。对于常态网络流量,如 HTTP,FTP,一般都是客户发送一个请求,然后服务器返回客户机所需要的数据。在这种网络结构中,上行流量与下行流量是不平衡的。通过比较某个连接的上行流量和下行流量,当这个比值在一定范围内时,就认为这个连接是 P2P 连接,那么所对应的流量就为 P2P 流量。

3)对端 IP 地址数量与 port 数量的比值:TCP 的连接特征表现为 P2P 流量具有多个对端 IP 和对端 port。对于 P2P 文件传输,一个 P2P 客户端和一个或多个 P2P 客户端建立连接。相对于源端,其对端的 IP 地址数量较多,对端的 IP 和对端的 port 都是随机分布的,而且对端 IP 和对端 port 个数比值接近 1。非 P2P 流量二者的比值小于 1,采用对端 IP 的数量以及对端 port 个数的比值可以作为 P2P 流量的一个识别特征。

2.3　核函数的选择

核函数 $K(x_i,x_j)$ 的要求是满足 Mercer 定理,在这个要求内怎样选择它,选择不同的函数可以构造出不同的支持向量机。目前,仍然没有一种对确定的针对 P2P 网络流量识别问题构造出适当核函数的有效方法。目前针对 P2P 网络流量的识别普遍采用了径向基(RBF)核函数,因为 RBF 核相比其它核函数具有较少的超参数,计算难度较小,能够使用于所有分布的样本。

选择适当的核函数有助于提高模型的精度和迭代的收敛速度。文中从核函数的构架上引入小波基函数 Mexican hat 小波函数,对支持向量机的结构进行优化。P2P 网络流量呈现出突发性,不确定的非线性流量特征,小波分析适合于信号的局部分析和突变信号的检测,结合小波分析引入满足框架理论的小波基函数来构造支持向量机的核函数,建立小波支持向量机的识别算法,能充分提高支持向量机的学习精度。

在平方可积空间 $L^2(R)$,若 $F=\{\psi_i\}$ 是一个框架,且有增序正数列$\{\lambda_i\}$,使得函数 $K(x,y)$ 可表示成下列形式:$K(x,y)=\sum_i \lambda_i\psi_i(x)\psi_i(y)$,只要满足 Mercer 条件就可以作为在 SVM 中使用的核函数。

给定平方可积函数 $\psi(x)\in L^2(R)$ 的傅里叶变换为 $\psi(\omega)$,且满足条件

$$\int_R \frac{|\psi(\omega)|^2}{|\omega|}d\omega<\infty$$

则以函数 $\psi(x)\in L^2(R)$ 为母小波,采用不同的平移和伸缩因子,可生成一维小波函数系,即

$$\psi(x)=|a_i|^{-1/2}\psi\left[\frac{x-b_i}{a_i}\right],a_i,b_i\in R,i\in Z \quad (5)$$

式中 $|a_i|^{-1/2}$ 为归一化系数,a_i,b_i 分别为伸缩和平移因子,$\psi(x)$ 就为依赖于参数 a_i,b_i 的小波基函数。

一些母小波基可生成小波框架,而框架就可以用来构造核函数,但必须满足 Mercer 条件:

$$\iint k(x,x')f(x)f(x')dxdx'\geqslant 0 \quad (6)$$

就可以将 $k(x,x')$ 写成特征空间中的点积形式,即 $K(x,x')=K(<x,x'>)$,对于平移不变核函数同样可以得到 $K(x,x')=K(x-x')$,故可得满足 Mercer 条件的平移不变小波核函数:

$$K(x,x')=K(x-x')=\prod_{i=1}^{n}\psi\left(\frac{x_i-x'_i}{a_i}\right)$$

文中采用 Mexican hat 小波来构造支持向量机的核函数,Mexican hat 小波函数为:

$$\psi(x)=(1-x^2)\exp\left|-\frac{x^2}{2}\right|$$

得到的新的核函数为:

$$K(x,x')=\prod_{i=1}^{n}\left|1-\left|\frac{x-x'}{a_i}\right|^2\right|\cdot \exp\left|-\frac{\|x-x'\|^2}{2a_i^2}\right| \quad (7)$$

通过引入小波基函数作为支持向量机的核函数,形成新的小波支持向量机(WSVM)(见图 2)。

3　模型算法实现步骤

(1)训练步骤:

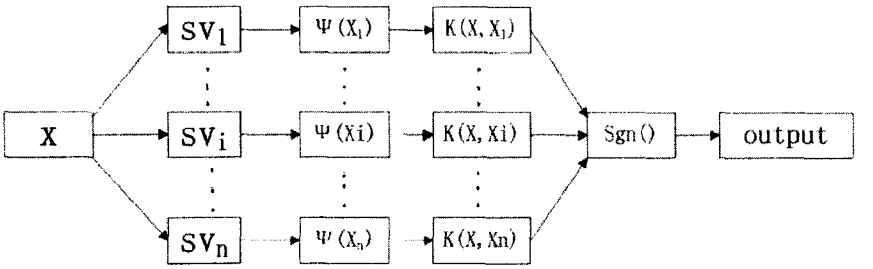

图 2 小波支持向量机网络结构图

第一步:输入两类训练样本向量 $(x_i, y_i)(i = 1, 2, \cdots, N, x \in R^n, y \in \{-1, 1\})$,类号分别为 ω_1, ω_2。如果 $x_i \in \omega_1$,则 $y_i = -1$;$x_i \in \omega_2$,则 $y_i = 1$。

第二步:选取 Mexican hat 小波来构造支持向量机的核函数 $\psi(x) = (1 - x^2)\exp\left|-\frac{x^2}{2}\right|$。

第三步:求解公式(2)的最优解,得到最优 Lagrange 乘子 a_i。

第四步:利用样本库中的一个支持向量 X,代入公式(4),左值 $f(x)$ 为其类别值(-1 或 1),可得到偏差值 b^*。

(2) 分类步骤:

第一步:输入待测样本 X。

第二步:利用训练好的 Lagrange 乘子 a_i、b^* 和核函数,求解公式(3)。

第三步:根据 sgn($f(x)$) 的值,输出类别。如果 sgn($f(x)$) 为 -1,则样品属于 ω_1;如果 sgn($f(x)$) 为 1,则样品属于 ω_2。

4 实验及分析

实验的主要目的是比较 SVM 采用径向基核函数与小波支持向量机 P2P 流量识别模型(WSVMP2P)来识别真实 P2P 网络流量的精确度,流量从河南大学网络中心的路由器出口进行采集。

实验选取台湾大学林智仁副教授等开发设计的一个快速有效的 LIBSVM 模式识别软件包。

实验步骤是:

1)按照 LIBSVM 软件包所要求的格式准备网络流量数据集;

2)对流量数据进行归一化处理,将所有的训练数据和测试数据都映射到[-1,1]区间;

3)选用不同的核函数;

4)采用交叉验证选择最佳误差惩罚因子 C 与核函数参数 γ,取值分别为 $C = 32$, $\gamma = 0.03$;

5)采用最佳参数 C 与 γ 对整个训练集进行训练获取支持向量机模型;

6)利用获取的模型在支持向量机模型(SVM)和小波支持向量机模型(WSVMP2P)中分别进行测试与预测。

从河南大学网络中心的路由器出口,利用端口映射,采集到了 5 组流量数据,前两组数据集是 P2P 流量和非 P2P 流量,第三组数据集是混合流量,按 hour 为单位收集,第四组数据集是按 day 为时间单位收集,第五组以 week 为时间单位收集,规模由小到大,采集的流量类型包含了文件共享、流媒体、网络电话、即时通讯等常见的 P2P 流量的应用类型,便于后期实验对比。通过整理,数据集描述见表 1。

表 1 实验采集的数据集描述

SET	during	interval	p2p	p2p type	flows	bytes
set 1	1 hour	15min	100%	Bittorrent	265.2k	183.2M
set 2	2 hour	15min	0	------	311.2k	36.4M
set 3	3 hour	15min	53.40%	Bittorrent	328.6k	45.6M
set 4	1 day	15min	85.60%	eDonkey pplive	3.2M	2.6G
set 5	1 week	15min	11.20%	ppstrem pplive eDonkey Bittorent	28.4M	22.9G

通过图 3 可以看出,基于小波支持向量机的 P2P 流量识别方法(WSVMP2P),随着训练数据集的增加,训练时间的累积,平均分类准确率都能保持一定的稳定性,没有出现明显的抖动,并且识别的精确度达到

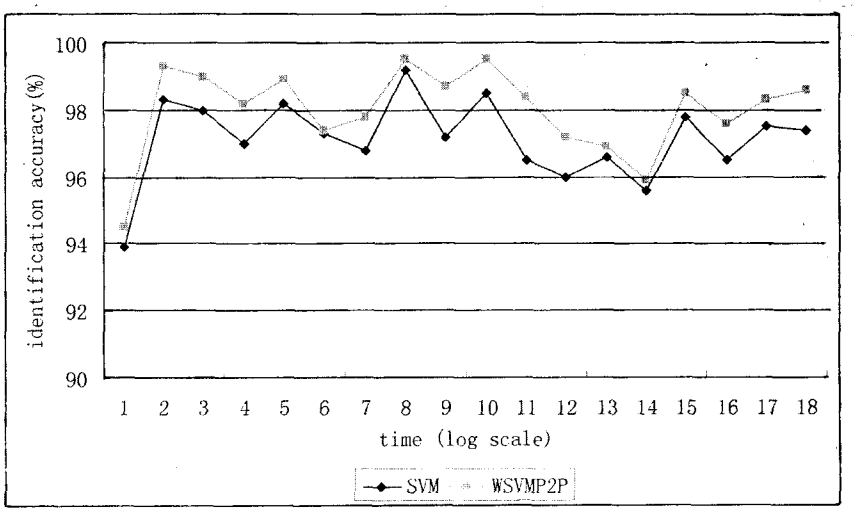

图 3 两种模型的识别精度对比图

98%,由此可看出,基于小波支持向量机的识别方法在实际问题中获得了比原有流量识别方法更高的精度。

5 结束语

提出了基于小波支持向量机的P2P网络流量的

(下转第 114 页)

上,对其做了改进,增加了滑动窗口机制,使之适应P2P视频点播。在此基础上给出了基于BitTorrent的P2P VOD系统的架构,改进的BitTorrent协议是基于BitTorrent的P2P VOD系统的一部分,通过对完整系统的仿真实验验证了改进后的协议的有效性。

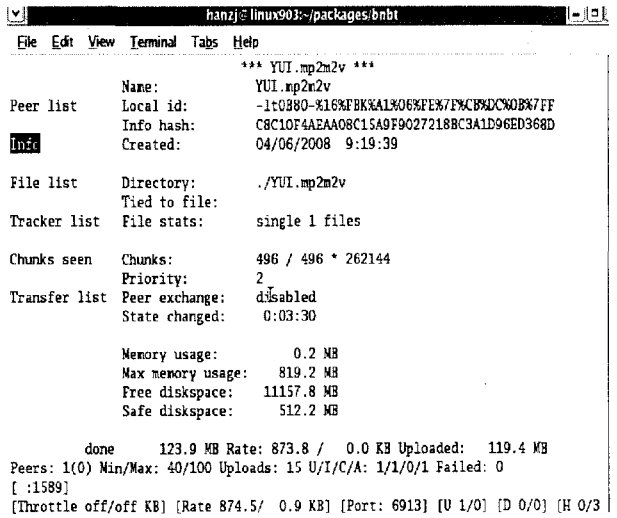

图 6 服务端显示的文件下载过程

参考文献:

[1] 吴广智. VOD视频点播核心技术研究[J]. 中山大学学报论丛,2006,26(2):126-128.

[2] 李 杰,李 毅. P2P实时点播系统[J]. 电脑知识与技术,2008,15(6):1055-1057.

[3] 程先宏. 多媒体网络QoS技术简述[J]. 科技信息(科学教研),2007,24(35):235-236.

[4] 吴国庆. 对等网络技术研究[J]. 计算机技术与发展,2008,18(7):100-103.

[5] 段翰聪. P2P流媒体分发技术研究[D]. 成都:成都电子科技大学,2007.

[6] Deshpande H, Bawa M, Garcia-Molina H. Streaming live media over a peer-to-peer network[R]. USA: Stanford University, 2001.

[7] Tran D A, Hua K A, Do T T. A peer-to-peer architecture for media streaming[J]. IEEE Journal on Selected Areas in Communications, 2004, 22(1): 121-133.

[8] COHEN B. Incentives to build robustness in BitTorrent[EB/OL]. 2009-03. http://www.bittorrent.org/bittorrentecon.pdf.

[9] 欧阳荣,雷振明. BitTorrent类型P2P系统模型研究与性能分析[J]. 北京邮电大学学报,2006,29(2):113-117.

[10] 王 珏. BitTorrent下载技术研究[J]. 科技广场,2005,5(2):26-27.

[11] 聂 哲. BitTorrent技术探讨与性能改进[J]. 现代计算机,2007,8(6):107-109.

[12] 杨 刚. 基于Linux的嵌入式媒体播放器研究[D]. 重庆:重庆大学,2007.

(上接第 110 页)

识别算法,利用了小波基函数的稀疏变化和多尺度差值可以有效地提高SVM模型的精度。WSVM除了具有支持向量机本身的有点外,其小波核函数是近似相交的,能够以很高的精度逼近任意函数,而传统使用的SVM的径向基核函数是相关的,甚至是冗余的,因而小波支持向量机的计算量相对更小,更加适合解决P2P突变性网络流量的识别问题。

参考文献:

[1] Cache logic 中国互联网流量分析报告[EB/OL]. 2005. http://www.cachelogic.com/home/pages/research/p2p2005.php.

[2] Saroiu S, Gummadi P K, Gribble S D. A measurement study of peer-to-peer file sharing systems[C]//Proceeding of the Multimedia Computing and Networking 2002. San Jose, California: ACM Press, 2002:156-170.

[3] 蒋海明, 张剑英, 王青青,等. P2P流量检测与分析[J]. 计算机技术与发展,2008,18(7):75-79.

[4] WANG Rui, LIU Yang, YANG Yuexiang, et al. Solving the app-level classification problem of P2P traffic via optimized support vector machines[C]//Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications. Jinan, China: IEEE Computer Society Press, 2006: 534-539.

[5] 黄烟波, 周磊戈. 基于流特征的P2P流量识别方法研究[J]. 计算机技术与发展,2009,19(9):46-48.

[6] 吴 敏,王汝传. 基于主机的P2P流量检测与控制方案[J]. 计算机技术与发展,2009,19(10):26-29.

[7] 宫 婧, 孙知信, 陈二运. 一种基于流量行为分析的P2P流媒体识别方法[J]. 计算机技术与发展,2009,19(9):128-131.

[8] 王春枝, 李 涛. 基于双层特征的P2P流量检测[J]. 计算机技术与发展,2009,19(7):238-241.

[9] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques [C]//ACM SIGMETRICS 2005. Banff, Alberta, Canada: ACM Press, 2005:50-60.

[10] 沈富可,常 潘. 基于BP神经网络的P2P流量识别研究[J]. 计算机应用,2007,27(12):44-45.

[11] Yang Ai-min. A P2P Network Traffic Classification Method Using SVM[C]// The 9th International Conference for Young Computer Scientists. [s.l.]: IEEE, 2008.

[12] 崔锦泰. 小波分析导论[M]. 程正兴译. 西安:西安交通大学出版社,1997.