ISSN 1000-1239/CN 11-1777/TP
48(12): 2253-2260, 2011

一种基于机器学习的 P2P 网络流量识别方法

李致远[1,2]　　王汝传[1]

[1](南京邮电大学计算机学院　南京　210003)

[2](江苏大学计算机科学与通信工程学院　江苏镇江　212013)

(lizhiyuan81@126.com)

A P2P Network Traffic Identification Approach Based on Machine Learning

Li Zhiyuan[1,2] and Wang Ruchuan[1]

[1](*College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210003)

[2](*College of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu* 212013)

Abstract　Peer-to-peer (P2P) overlay networks are typical distributed systems in nature, which have attracted more and more attentions. At present, the P2P technology has been applied in file sharing, streaming media, instant messaging, and other fields. Besides, P2P network traffic accounts for more than 60% of Internet traffic. In order to better manage and control the P2P traffic, it is necessary to study a P2P traffic identification model in depth. Firstly, a machine learning model based on the wavelet support vector machine (ML-WSVM) is proposed to identify known and unknown P2P traffic. In the ML-WSVM model, the combination of the wavelet with the support vector machine is implemented by the wavelet basis function which satisfies the wavelet framework and the Mercer theorem instead of the existing support vector machine kernel functions. The proposed model makes full use of multi-scale features of the wavelet and the advantages of the support vector machine used in the classification. Then, the improved sequential minimization optimization (SMO) algorithm based on a loss function is proposed to solve the optimal hyperplane of the ML-WSVM model. Finally, the theoretical analysis and experimental results show that the ML-WSVM model can greatly improve the identification accuracy and identification efficiency of P2P network traffic, particularly to identify the encrypted packets.

Key words　peer to peer networks; network traffic identification; support vector machine; wavelet function; loss function

摘　要　对等(P2P)覆盖网络作为一种典型的分布式系统日益受到人们的重视.P2P 应用遍及文件共享、流媒体、即时通信等多个领域,P2P 应用所产生的流量占据了互联网流量的 60%以上.为了更好地管理和控制 P2P 流量,有必要对 P2P 流量识别模型进行深入的研究.提出一种基于小波支持向量机的机器学习模型(ML-WSVM)来识别已知和未知的 P2P 流量,ML-WSVM 是通过满足小波框架和 Mercer 定理的小波基函数替换支持向量机核函数的方法,实现小波与支持向量机的结合.该模型充分利用了小波的多尺度特性与支持向量机在分类方面的优势.然后,提出基于损失函数的串行最小化算法来优化求解 ML-WSVM 的最优分类面.最后,理论分析和实验结果表明该方法大大提高了对 P2P 网络流量的识别精度和识别效率,尤其是对加密报文的识别.

关键词　对等网络;网络流量识别;支持向量机;小波函数;损失函数

中图法分类号　TP393.08

收稿日期:2010-07-05;**修回日期**:2010-09-14

基金项目:国家自然科学基金项目(60573141,60773041);国家"八六三"高技术研究发展计划基金项目(2006AA01Z439,2007AA01Z404);南京市高科技资助基金项目(2007 软资 127);江苏省高校科技创新计划资助基金项目(CX10B_198Z)

对等网络(peer to peer networks,P2P)是分布式系统和计算机网络相结合的产物,目前 P2P 受到了学术界和工业界的双重重视.当前 P2P 技术已经渗透到文件共享、文件存储、流媒体、即时通信等多个领域.据 2008 年中国互联网流量研究报告显示,目前国内网络流量的 50%～60%由 P2P 流量占据,并预测在未来几年,P2P 流量将占到网络流量的 70%以上.可见,P2P 流量已经严重地影响到了互联网整体的服务质量,只有通过对 P2P 流量的严格控制才能使互联网成为一个可控、可管、为用户提供高服务质量的网络.而对 P2P 流量进行控制的前提是必须能够准确地识别 P2P 流量.但是,当前的 P2P 流量大都是加密报文,现有的 P2P 流量检测技术往往对这些报文无能为力,使得 P2P 流量识别率大大降低.

本文提出一种基于小波支持向量机的机器学习模型(machine learning model based on wavelet support vector machine, ML-WSVM)以达到提高 P2P 流量识别率的目标.ML-WSVM 充分利用了小波的多尺度特性和支持向量机(support vector machine, SVM)在分类方面的优势,使用小波基函数来构造 SVM 核函数,建立小波支持向量机的 P2P 流量识别模型.然后,通过损失函数的方法优化求解模型的最优分类面.最后,通过实验证明该模型大大地提高了 P2P 网络流量的识别精度,尤其是对加密报文的识别.

1 相关工作

目前 P2P 流量检测技术大致分为以下 3 种:基于端口的检测技术[1]、深层数据包检测(deep packet inspection, DPI)技术[2-3]和基于流量特征的检测(transport layer identification, TLI)技术[4-7].但由于端口往往采用随机选择方式以及数据报文被加密导致上述技术对当前的 P2P 协议报文识别率大大降低.针对这种现状,文献[8]提出了基于 BP(back propagation)神经网络的 P2P 流量识别技术.但由于神经网络的拓扑结构难以确定以及在训练过程中,模型容易陷入局部极小或出现过学习使得该技术的识别准确率不是很高.文献[9-10]提出使用 SVM 技术对 P2P 流量进行识别.SVM 是针对小样本的机器学习方法,它是通过求解凸二次优化问题得到全局最优解.但由于现有 SVM 模型中所采用的核函数只能在同一尺度上对样本数据进行分类,对多尺度样本的逼近能力较差.此外,上述方法仅考虑了在最优分类面上的样本成为各类别的概率是相同的,而没有考虑其损失是不同的.即将非 P2P 流量错误地识别为 P2P 流量,从而错误地对该流量进行控制所带来的损失远大于将 P2P 流量识别为非 P2P 流量所造成的损失.

2 小波支持向量机识别模型

2.1 支持向量机

SVM 是 Vapnik 基于统计学原理提出的一种机器学习方法.在该方法中,核函数是影响 SVM 分类效果的关键因素之一.研究发现,系数变化的核函数有助于提高模型的分类精度和收敛速度;另一方面,对于平滑函数缺乏先验知识的情形,使用多尺度差值方法是最好的选择.而小波函数恰好具有上述两种特性,即伸缩性和多尺度[11].因此,以小波函数为核函数的 SVM 具有更强的函数逼近能力和泛化能力.图 1 是 SVM 及其最优分类超平面的示意图.如图 1 所示,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.H_1 和 H_2 分别表示过各类样本中距离分类面最近的样本且平行与分类面的平面,H_1 和 H_2 之间的距离称为分类间隔(margin).

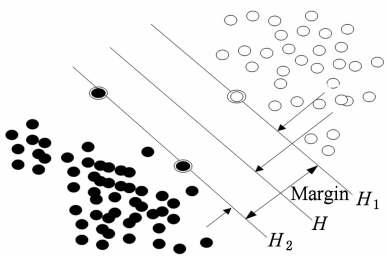

Fig. 1 The optimal hyperplane of support vector machine.
图 1 支持向量机最优分类超平面示意图

最优分类超平面是不仅能将两类样本正确分开使分类间隔最大的超平面.分类超平面方程如式(1)所示:

$$w \cdot X + b = 0. \tag{1}$$

通过对它归一化使得线性可分的样本集(x_i, y_i),$i=1,\cdots,n$,$x \in \mathbb{R}^d$,$y \in \{+1,-1\}$,满足式(2):

$$y_i[(w \times x_i)+b]-1 \geqslant 0, i=1,2,\cdots,n. \tag{2}$$

此时分类间隔为 $2/\|w\|$.要使间隔最大等价于使 $\|w\|^2$ 最小,满足式(2)且使 $\|w\|^2/2$ 最小的分类面即为最优分类平面,H_1 和 H_2 上的样本点即为支持向量.

此时,将问题转化为求解凸二次优化,见式(3):

$$\begin{cases} Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i \times x_j), \\ \sum_{i=1}^{n} a_i y_i = 0, a_i \geqslant 0, i = 1, \cdots, n. \end{cases} \tag{3}$$

在式(3)中,a_i 为 Lagrange 乘子,其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,见式(4)所示:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i (x_i \times x) + b^*). \tag{4}$$

对于非线性分类问题,需通过核函数将非线性输入向量映射到一个线性可分的高维特征空间中.对于任意的核函数都有 $K(x_i, x_j) = \phi(x_i) \times \phi(x_j)$.那么,式(4)的相应核函数表达见式(5):

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*). \tag{5}$$

2.2 特征向量的选取

特征是能够区分 P2P 和非 P2P 流量的一种属性.下面从数据包、数据流和连接层 3 方面对 P2P 流量特征进行分析,从而组成一个三维特征向量.

1) 数据包大小变化的均方差.P2P 应用的数据包大小源于 P2P 协议自身.P2P 协议报文主要分为信令报文和数据报文.P2P 协议中各类信令报文的长度不同,信令报文的长度与数据报文的长度也不同.P2P 协议内部的实现机制使得 P2P 应用的数据包大小的均方差比非 P2P 应用的数据包要大.

2) 上下行流量的比值.P2P 网络的特点是每个 Peer 节点既是客户端又是服务器.它在从其他节点下载的同时,也上传数据给网络中的其他节点.由于现有的 P2P 应用系统都采用了激励机制,使得每个 Peer 节点的上下行流量大致对称.而对于非 P2P 应用都是客户端发送一个请求,然后由服务器返回客户端所需要的数据.这种 C/S 或 B/S 架构使得上下行流量严格不对称.

3) 资源请求端与资源提供端的 IP 和 Port 数量的比值.以 P2P 文件共享应用为例,一个 P2P 资源请求端要与多个 P2P 资源提供端建立连接,从而形成一对多关系.此时,资源请求端与资源提供端的 IP 和 Port 数量的比值远小于 1.而对于非 P2P 应用,交互双方是一对一关系.此时,资源请求端与资源提供端的 IP 和 Port 数量的比值等于 1.

2.3 核函数的选择

鉴于 P2P 流量属于不确定的非线性流量,而小波又非常适合信号的局部分析和突变信号的检测.考虑将小波基函数替代原有的 SVM 核函数,构建一种小波支持向量机模型,以达到提高 SVM 学习精度的目标.对于小波基函数,选择 Mexican hat 小波.

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 $K(x, y) = \sum_i \lambda_i \psi_i(x) \psi_i(y)$.

对给定平方可积函数 $\psi(x) \in L^2(R)$ 进行傅里叶变换得到 $\psi(\omega)$,$\psi(\omega)$ 满足式(6)的条件:

$$\int_R \frac{|\psi(\omega)|^2}{|\omega|} \mathrm{d}\omega < \infty. \tag{6}$$

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,生成一维小波函数系,见式(7)所示:

$$\psi(x) = |a_i|^{-1/2} \psi\left[\frac{x - b_i}{a_i}\right], a_i, b_i \in R, i \in \mathbb{N}, \tag{7}$$

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

母小波基可生成小波框架,而框架可以用来构造核函数,但需满足 Mercer 条件,见式(8)所示:

$$\iint K(x, x') f(x) f(x') \mathrm{d}x \mathrm{d}x' \geqslant 0. \tag{8}$$

一旦满足 Mercer 条件就可以将 $K(x, x')$ 写成点积形式,即 $K(x, x') = K(\langle x, x' \rangle)$.对于平移不变核函数有 $K(x, x') = K(x - x')$,故可得满足 Mercer 条件的平移不变小波核函数,见式(9)所示:

$$K(x, x') = K(x - x') = \prod_{i=1}^{n} \psi\left(\frac{x_i - x_i'}{a_i}\right). \tag{9}$$

采用 Mexican hat 小波来构造 SVM 核函数,见式(10)所示:

$$K(x, x') = \prod_{i=1}^{n} \left|1 - \left|\frac{x_i - x_i'}{a_i}\right|^2\right| \times \exp\left|-\frac{\|x_i - x_i'\|^2}{2a_i^2}\right|. \tag{10}$$

最后,通过引入小波基函数作为 SVM 核函数,形成小波支持向量机模型,如图 2 所示:

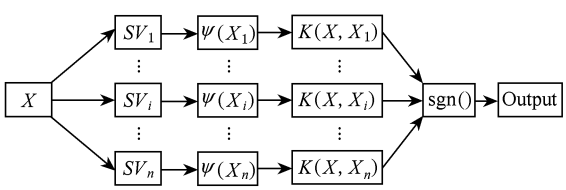

Fig. 2 Wavelet support vector machine model.

图 2 小波支持向量机模型

3 基于损失函数的小波支持向量机优化求解算法

图 3 是 $f(x)$与诊断决策的关系图. 如图 3 所示，对于检测样本 x，如果存在 $f(x)\leqslant -1$ 或 $f(x)\geqslant 1$，则可给出明确的诊断结果. 而在区间 $-1<f(x)<1$ 特别是在 $f(x)=0$ 的邻域内，诊断结果存在一定的不确定性因素，故将该区域称为可疑诊断区.

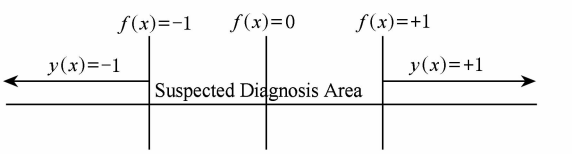

Fig. 3 The relationship between $f(x)$ and diagnosis decision.

图 3 $f(x)$与诊断决策的关系图

为了定量描述可疑诊断区诊断结果的不确定性程度，现引入诊断可信度函数.

定义 1. 诊断可信度函数. 诊断可信度函数 $T(X,y)$是衡量将当前运行状态 X 诊断为 y 的可信程度. 显然对于两类诊断问题{正常为 1，故障为 -1}，对任一诊断样本都有

$$T(x,1)+T(x,-1)=1. \tag{11}$$

诊断可信度函数定义见式(12)(13)：

$$T(x,1)=\frac{1}{2}[1+f(x)], \tag{12}$$

$$T(x,-1)=\frac{1}{2}[1-f(x)]. \tag{13}$$

显然在分类面上 $T(x,1)=T(x,-1)$，则有 $f(x)=0$.

传统的 SVM 理论只考虑了在最优分类面上的样本成为各类别的概率是相同的情况，而没有考虑其损失是不同的. 在流量识别问题中，需要考虑一个比误判更为广泛的概念——风险. 对于 P2P 流量检测，可能将非 P2P 流量错误地识别为 P2P 流量，从而错误地对流量实施控制，其带来的损失远大于将 P2P 流量识别为非 P2P 流量所带来的损失. 因此，在进行判决时，必须考虑不同的分类错误所引起的损失. 下面用决策表的形式描述该问题，表 1 即为 P2P 流量识别问题的决策表.

在表 1 中，当 $L_{11}=L_{12}=0$ 时，即正常分类，不存在选择. 而对于流量识别问题，一般假定 $L_{21}\geqslant L_{12}$，即把非 P2P 流量错误地识别为 P2P 流量所带来的损失大于将 P2P 流量识别为非 P2P 流量产生的损失. (L_{21},L_{12})的选择不是随机的，而是由系统运行过程中对数据进行统计得到的.

Table1 P2P Traffic Decision Table

表 1 P2P 流量决策表

Decision	Actual State	
	P2P Traffic	Non-P2P Traffic
P2P Traffic	L_{11}	L_{12}
Non-P2P Traffic	L_{21}	L_{22}

定义 2. 基于损失函数的诊断可信度函数. 引入损失函数后，记 $T'(x,1)$为考虑损失样本 X 属于第 1 类的可信度函数，$T'(x,-1)$表示考虑损失样本 X 属于第 2 类的可信度函数. 此时，基于损失函数的诊断可信度函数定义见式(14)(15)：

$$T'(x,1)=\frac{L_{21}T(x,1)}{L_{12}T(x,-1)+L_{21}T(x,1)}, \tag{14}$$

$$T'(x,-1)=\frac{L_{12}T(x,-1)}{L_{12}T(x,-1)+L_{21}T(x,1)}. \tag{15}$$

显然在分类面上，$T'(x,1)=T'(x,-1)$. 由此可得修正后的最优分类面见式(16)所示：

$$f(x)=\frac{L_{12}-L_{21}}{L_{12}+L_{21}}. \tag{16}$$

下面分别给出模型的训练分类步骤和 P2P 流量识别步骤.

1) 训练分类步骤

① 选择具有明确属性的标准样本(x_i,y_i)，$i=1,\cdots,n$，$x\in\mathbb{R}^d$，使得 $y\in\{+1,-1\}$；

② 根据特征预处理(将采集到原始数据包转换为 Weka 软件可以识别的文件格式)后的训练向量训练小波支持向量机，调整小波支持向量机参数到最佳状态(将两类训练样本准确无误的分开且分类间隔最大)，并得到相应的 Vapnik 意义(小样本、非线性及高维模式下的分类识别)下的超平面 $f(x)=0$；

③ 选择式(16)替代步骤②得到的决策函数作为实际决策函数，并修正最优分类面.

2) 流量识别步骤

① 对抓取到的 P2P 流量进行特征预处理；

② 判断该流量是否已经存在于已经识别的 P2P 记录表中，如已存在，则转 P2P 流量控制模块，如果不存在，则转到步骤③；

③ 利用训练步骤中得到的决策函数对流量性质进行判定. 如果是 P2P 流量，则更新 P2P 记录表.

4 性能评估及模型复杂性分析

4.1 实验环境设置

为验证 ML_WSVM 对 P2P 流量的识别准确率，

按照上述原理实现了 ML_WSVM 的原型系统,并将之部署在校园网核心路由器的旁路上,图 4 即为网络的拓扑结构:

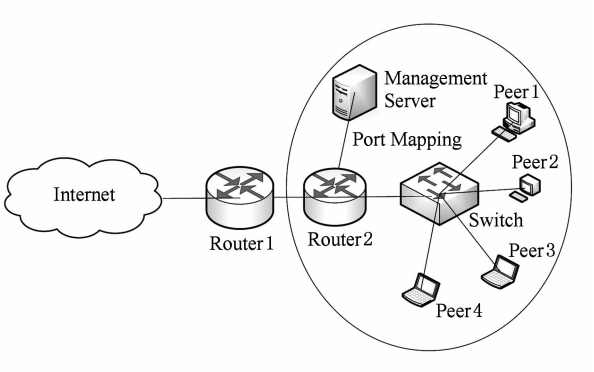

Fig. 4 Experimental topology.

图 4 实验拓扑图

在图 4 中,管理服务器型号为 Dell PowerEdge 2850,操作系统为 SUSE 10.0 (64b).管理服务器上要安装 Weka 软件,主要用于离线模型训练和在线流量识别.4 个 Peer 分别为笔记本电脑和台式机.实验中,这 4 个 Peer 同时运行 P2P 应用程序(BitTorrent,eMule,ARES 和 PPlive)进行资源发现和下载.之所以选择上述 4 种应用软件不仅因为它们更常用,还因为上述 4 种软件对关键报文和数据分片都采用了加密技术.其中关键报文包括预登录返回报文(首次与网络交互所产生的报文,称为预登录报文)和分片下载请求报文.这两类报文对 P2P 流量检测与控制起到了关键作用,尤其是分片下载请求报文对 P2P 流量识别率影响较大.

4.2 性能评估

1) ML_WSVM 对 P2P 流量的识别准确率

对于 P2P 流量的识别准确率从两方面验证,一是对 P2P 流量和非 P2P 流量的识别准确率;二是对混合 P2P 流量中每一类 P2P 协议的识别准确率.

具体实验步骤如下.

① 在局域网中,利用 NetMate 工具采集网络数据包并计算网络流量特征,按流进行分组并标注是否属于 P2P 应用.流量数据是每隔 1 min 采集一次,采集到的流量分成 3 个数据集,分别命名为 enry1.final.arff,enry2.final.arff 和 enry3.final.arff.

② 从数据源文件中提取流相关信息,将三维特征向量进行数据的向量化,形成训练和分类数据源.

③ 利用 Weka-3.6.0 (Waikato environment for knowledge analysis, Weka)软件进行 ML_WSVM 模型训练,这是一个离线分析的过程.实验中采用 10-fold 交叉验证方法.表 2 是(L_{21},L_{12})=(2,1)时训练集的分类精度、真阳性率、漏检率和误检率.

Table 2 Classification Accuracy, True Positive Rate, False Negative Rate and False Positive Rate of Rraining Set

表 2 训练集的分类精度、真阳性率、漏检率和误检率 %

Data Sets	Classification Accuracy	True Positive Rate	False Negative Rate	False Positive Rate
Entry1	99.90	96.4	1.2	4.3
Entry2	97.9	99	1.0	4.6
Entry3	99.23	100	0.4	3.9
Mean	99.01	98.47	0.86	4.23

④ 表 3 是从校园网中心的核心路由器出口采集并整理后的流量数据.

Table 3 Description of Data Sets

表 3 数据集描述

Sets	During Time/h	Interval /min	The ratio of P2P traffic to total traffic/%	Types	Flows /MB	Output Traffic /MB
Set1	1	10	100	ARES	0.2652	183.2
Set2	2	10	0		0.3112	36.4
Set3	3	10	53.40	BT ARES	0.3286	45.6
Set4	24	10	85.60	Emule ARES	3.2	2600
Set5	168	10	11.20	PPlive	28.4	10000

如表 3 所示,前两组数据集是 P2P 流量和非 P2P 流量;第 3,4,5 组数据集是混合流量,以小时(h)为单位收集.

按照本节的步骤②③,分别对表 3 中的 5 个数据集进行训练,得到 5 种 ML_WSVM 模型.之后,任意选取一种识别模型,同时将其他 4 个数据集作为测试集,实验重复 10 次.至此,便得到 5 种 P2P 流量识别模型的平均真阳性率、平均漏检率和平均误检率.最终,选择真阳性率最高、漏检率及误检率相对较低的数据集 4 训练出的模型作为 P2P 流量识别模型,离线分析过程结束.下面用上述离线分析得到的 P2P 流量识别模型对 P2P 流量进行实时检测,时间为 1000 min,每次只打开一种 P2P 应用软件,实验重复 10 次后取均值.

图 5 是各种 P2P 协议的真阳性率随时间变化的曲线.如图 5 所示,ML_WSVM 对各种 P2P 协议的真阳性率始终保持在较高的水平,没有出现明显

的波动,对多数 P2P 流量的识别精度能达到 96% 左右.

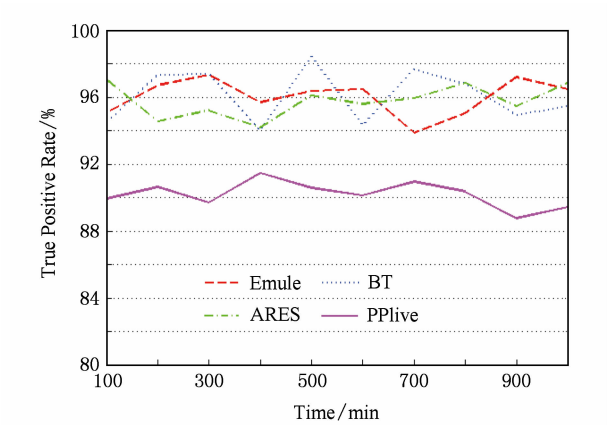

Fig. 5 Comparison of true positive rate of ML-WSVM model.

图 5 ML-WSVM 模型的真阳性率比较

图 6 是各种 P2P 协议的误检率随时间变化的曲线. 如图 6 所示,3 种 P2P 应用的误检率都低于 10%,但 PPLive 的误检率较其他 3 种应用略高. 这是因为实验采用的识别模型是基于 set4 数据集,而该数据集中没有包含 PPLive 流量. 尽管各类 P2P 应用产生的流量特征相似,但 P2P 文件共享和 P2P 流媒体所产生的流量特征还有所差别,这导致了误检率的上升. 从上述实验的整体效果看,ML-WSVM 流量识别模型对各类 P2P 流量的实时检测效果明显.

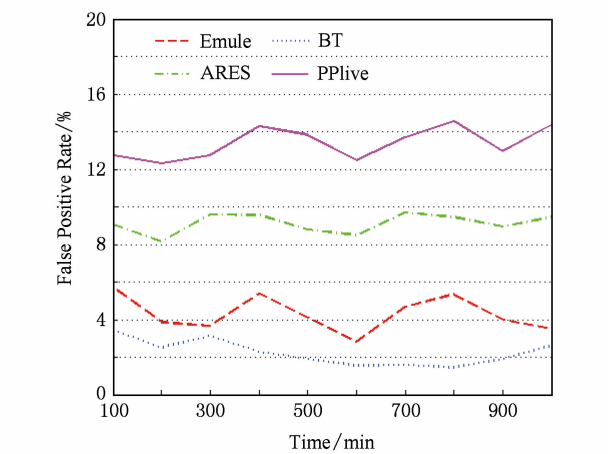

Fig. 6 Comparison of false positive rate of ML-WSVM model.

图 6 ML-WSVM 模型的误检率比较

2) ML_WSVM 对加密报文的识别准确率

P2P 协议中的报文分为加密和非加密两种,加密报文集中在预登录返回报文、分片下载请求及其相应的返回报文. 此外,有效数据在传输过程也是加密的,其目的就是让协议分析者找不到特征码,无法对 P2P 流量进行识别和控制. 下面就来验证 ML_WSVM 模型对加密报文的识别准确率.

针对开源的 P2P 文件共享软件 BitTorrent, eMule 和 ARES,通过代码了解到预登录返回报文、分片下载请求及其相应返回报文的构包过程. 用 VC 开发一个客户端软件,软件功能是构造加密报文并发送到外网. 实验中共发送 100 个不同协议和不同类型的加密报文,并选用 DPI 技术、TLI 技术、基于 BP 神经网络的识别技术、基于 SVM 的识别技术以及 ML_WSVM 五种方法对该加密报文进行识别,表 4 即为实验进行 10 次实验后取均值的结果:

Table 4 Encrypted Packet Identification Accuracy

表 4 加密报文的识别精度

Method	Identification Accuracy /%	Variance
DPI	2.75	1.25
TLI	84.35	2.98
BP	70.64	3.21
SVM	81.70	1.37
ML_WSVM	90.30	0.88

如表 4 所示,ML_WSVM 对加密的 P2P 协议报文的识别准确率明显高于其他几种 P2P 流量识别技术. 而目前在互联网上使用最广泛的 DPI 技术对加密协议报文的识别准确率仅达 2.75%左右.

3) ML_WSVM 与现有检测技术的识别率对比

表 5 是不同的算法在相同数据集上的识别准确率. 对于 DPI,TLI 和 BP,采用表 3 中的 Set3,Set4 和 Set5 数据集为测试集,实验重复 10 次后取均值,如表 5 前 3 项所示. 对于 SVM 和 ML_WSVM,也采用 Set3,Set4 和 Set5 数据集,所不同的是这 3 个数据集既是训练集又是测试集. 实验重复 10 次取均值,如表 5 后两项所示:

Table 5 Identification Accuracy of Different Algorithms

表 5 不同算法的识别准确率

Method	Identification Accuracy /%	Variance
DPI	79.72	2.21
TLI	93.35	1.87
BP	81.17	3.5
SVM	90.50	1.7
ML_WSVM	95.94	1.1

从表 5 中发现,在同样的数据集为测试集以及

同样的数据集为测试集和训练集的情况下,ML_WSVM 对 P2P 流量的平均识别准确率最高且最稳定;TLI 和 SVM 算法的识别准确率及稳定性次之;DPI 和 BP 算法的识别精度和稳定性最差.

下面针对测试集和训练集采用不同数据集的情况,对各类 P2P 识别算法的识别精度进行比较测试,表 6 即为各类识别算法在不同数据集上的识别准确率.首先从 Set1 到 Set5 五个数据集中,任选一个作为训练集,之后再选一个作为测试集.对于不需要训练的识别方法,直接使用测试集进行测试.如表 6 所示,采用不同数据集作为测试集和训练集时,各类 P2P 流量识别方法的识别精度与表 5 的实验结果一致.此外,实验还证明了 ML_WSVM 比其他 P2P 流量识别方法具有更强的泛化能力.

Table 6 Identification Accuracy on Different Data Sets

表 6 不同数据集上的识别准确率 %

Training Sets	Testing Sets	DPI	TLI	BP	SVM	ML_WSVM
Set3	Set4	74.13	96.82	79.07	88.37	91.60
Set3	Set5	83.06	90.65	82.93	86.32	93.43
Set3	Set1	90.03	89.49	85.45	89.05	98.87
Set4	Set5	70.72	93.80	80.01	85.53	98.54
Set4	Set1	69.29	94.98	87.51	90.49	97.32
Set1	Set4	76.48	90.81	78.85	93.85	95.50
Identification Accuracy		77.29	92.76	82.30	88.94	95.88

4.3 ML-WSVM 模型的复杂性分析

定理 1. ML-WSVM 模型的计算复杂度仅由超平面边缘上的少数支持向量决定,而与样本集的规模无关.ML-WSVM 模型的空间复杂度与样本集的规模 n 呈线性关系.

证明. 根据 ML-WSVM 模型的机制,首先通过非线性变换 Φ 将样本空间映射到一个高维空间,然后在该空间中求取最优线性分类面.这种非线性变换是通过定义小波核函数 $K(\boldsymbol{x}_i,\boldsymbol{x})$ 实现的.此时的优化函数见式(3),相应的 ML-WSVM 见式(5),其中 $\boldsymbol{x}_i$ 为支持向量,$\boldsymbol{x}$ 为未知向量.ML-WSVM 在分类函数形式上类似于一个神经网络,其输出是若干中间层节点的线性组合,而每一个中间层节点对应于输入样本与一个支持向量的内积.由于最终的 ML-WSVM 模型中只包含未知向量与支持向量的内积的线性组合,因此识别时的计算复杂度仅取决于支持向量个数,而与样本集的规模无关.对于模型最耗时间和空间的寻优过程,采用串行最小化算法(sequential minimization optimization, SMO).由于 SMO 不涉及二次规划数值解法,因而不必将核函数矩阵整个存放于内存中,则 SMO 使用的内存与样本规模 n 呈线性关系[12].于是,ML-WSVM 模型的空间复杂度与样本集的规模 n 也呈线性关系.

此外,由于 ML-WSVM 模型采用了离线分析和在线检测的策略,且模型旁路部署,因此,即便在大规模流量检测时也不会影响识别准确率和效率,对主干网不会增加负担.

5 结 论

本文提出了一种 ML-WSVM 的 P2P 流量识别方法,该方法通过使用小波基函数构造 SVM 核函数的方式来建立基于小波支持向量机的 P2P 流量识别模型.然后,采用基于损失函数的方法优化求解 ML-WSVM 模型使其获得最优的分类平面.ML-WSVM 模型充分利用了小波的伸缩性和多尺度特性以及 SVM 在分类方面的优势,使得模型具有更强的分类能力和实时检测能力.通过多项实验及其结果表明 ML-WSVM 较之其他 P2P 流量识别方法能够大大提高 P2P 流量的识别精度,尤其是对加密报文的识别.下一步工作是考虑将该模型移植到 FPGA 设备中,作为一个独立的硬件模块使用.

参 考 文 献

[1] Sen S, Wang J. Analyzing peer to peer traffic across large networks [J]. IEEE Trans on Networking, 2004, 12(2): 137-150

[2] Sen S, Spatscheck O, Wang D. Accurate, scalable in-network identification of P2P traffic using application signatures [C] //Proc of the 13th Int Conf on World Wide Web. New York: ACM, 2004: 512-521

[3] Wang R, Liu Y, Yang Y, et al. Solving the app-level classification problem of P2P traffic via optimized support vector machines [C] //Proc of the 6th Int Conf on Intelligent Systems Design and Applications. Piscataway, NJ: IEEE, 2006: 534-539

[4] Karagiannis T, Broido A, Faloutsos M, et al. Transport layer identification of P2P traffic [C] //Proc of the 4th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2004: 121-134

[5] Auld T, Moore A W, Gull S F. Bayesian neural networks for Internet traffic classification [J]. IEEE Trans on Neural Networks, 2007, 18(1): 223-239

[6] Zuev D, Moore A. Traffic classification using a statistical approach [G] //LNCS 3431: Proc of the 6th Int Workshop on Passive and Active Network Measurement. Berlin: Springer, 2005: 321-324

[7] Constantinou F, Mavrommatis P. Identifying known and unknown peer-to-peer traffic [C] //Proc of the 5th IEEE Int Symp on Network Computing and Appli-cations. Piscataway, NJ: IEEE, 2006: 93-102

[8] Chen H, Hu Z, Ye Z, et al. Research of P2P traffic identification based on BP neural network [C] //Proc of the 1st Int Symp on Computer Network and Multimedia Technology. Piscataway, NJ: IEEE, 2009: 1-4

[9] Yang A, Jiang S, Deng H. A P2P network traffic classification method using SVM [C] //Proc of the 9th Int Conf on Young Computer Scientists. Piscataway, NJ: IEEE, 2008: 398-403

[10] Liu F, Li Z, Nie Q. A new method of P2P traffic identification based on support vector machine at the host level [C] //Proc of the Int Conf on Information Technology and Computer Science. Piscataway, NJ: IEEE, 2009: 579-582

[11] Chui C. An Introduction to Wavelets [M]. New York: Academic Press, 1992

(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)

[12] Bai Peng, Zhang Xibin, Zhang Bin, et al. Supptor Vector Machine and Its Application in Mixed Gas Infrared Spectrum Analysis [M]. Xi'an: Xidian University Press, 2008 (in Chinese)

(白鹏, 张喜斌, 张斌, 等. 支持向量机理论及工程应用实例[M]. 西安: 西安电子科技大学出版社, 2008)

Li Zhiyuan, born in 1981. PhD from the College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China. Student member of China Computer Federation. His main research interests include P2P networks security and wireless sensor networks.

Wang Ruchuan, born in 1943. Professor and PhD supervisor of Nanjing University of Posts and Telecommunications. Senior member of China Computer Federation. His main research interests include P2P computing, the security of network, sensor networks, etc.