基于优化 SVM 的 P2P 协议识别 *

毛 灵, 陈兴蜀, 吴仲光, 谭 骏, 杜 敏
(四川大学 计算机学院, 成都 610065)

摘 要: 针对 P2P 应用提出了一种采用 DFI 深度流分析的方法,通过还原会话流,提取 P2P 数据流的各种属性特征,采用 Grid Search、遗传算法、粒子群算法三种不同算法优化的支持向量机对网络数据流进行分类。通过实验测试,在 P2P 与非 P2P 的多种应用中,使用支持向量机进行设计的分类器分类准确率较高,均在 90% 以上,最高能达到 97%。

关键词: 粒子群算法; 遗传算法; 支持向量机; P2P; 协议识别
中图分类号: TP393.0 **文献标志码**: A **文章编号**: 1001-3695(2011)07-2750-04
doi:10.3969/j.issn.1001-3695.2011.07.099

Novel P2P protocol identification algorithm based on optimized SVM

MAO Ling, CHEN Xing-shu, WU Zhong-guang, TAN Jun, DU Min
(*College of Computer Science, Sichuan University, Chengdu* 610065, *China*)

Abstract: To overcome the shortcomings of P2P application, this paper proposed a new P2P identification algorithm which used DFI deep flow analysis by extracting the various properties of P2P data stream characteristics, using support vector machine which was optimized by Grid Search, generation algorithm and particle swarm optimization to assort the network data streams. Experimental results of the P2P and non-P2P classification show that the classifier by SVM has a high accuracy over 90%, and the highest can be 97%.

Key words: PSO; GA; SVM; P2P; protocol identification

随着计算机通信网络的不断发展,通过网络共享资源的 P2P 应用越来越广泛。虽然 P2P 应用从很大程度上缓解了 C/S 架构模型中服务器的压力,但是 P2P 应用也存在很大的安全隐患:首先 P2P 应用占用了大量的网络带宽,影响其他应用的正常网络需求[1];其次 P2P 应用易泄露用户信息,威胁个人信息安全[2]。因此,对 P2P 应用的识别和监控显得十分必要。

P2P 应用的识别及控制经历了几个阶段:最初的 P2P 应用采用固定的端口进行通信,如 BT 通信端口为 6881 到 6889,因此,Sen 等人[3]提出了基于端口的 P2P 识别技术,只需要获取或封堵这些端口信息即可;之后的 P2P 应用不再采用固定端口,大量采用随机端口,有的甚至采用传统 80 端口进行数据传输,因此基于端口的 P2P 监控方法不再适用。之后研究者采用了深度数据包检测(deep packet inspection,DPI)技术,该方法主要基于数据包应用层特征码进行协议识别[4],它具有精度高以及实时性等优点。随着 P2P 技术的不断发展,P2P 应用中普遍运用加密技术,使得该方法也不再适用。在最近几年,越来越多的研究者使用基于 P2P 协议流量统计的特征进行分类识别[5,6],使用模式识别[7]、机器学习等领域的算法根据流量的统计信息对协议进行分类。在理论方面取得了较大的突破。但是基于流量统计特征的识别方法计算量大,一般用于离线分析,不能够满足实时性要求,因此不能够在线实时地对网络流量进行分析,识别出其中的 P2P 流。

针对这些问题,本文提出了一种新的 DFI 识别技术,它基于 P2P 数据流的各种属性进行统计分析,以实现不同 P2P 应用以及非 P2P 应用的分类,进而实现对 P2P 应用的监管。不同的协议具有不同的特征,本文通过获取数据包会话流的各种属性,对其进行统计分析,通过特征提取,选择出最能识别出各种应用的多种属性,并利用粒子群算法[8]优化的支持向量机[9]进行分类。

1 基于行为统计特征的 P2P 识别

传统的协议识别是根据网络中数据包的特征码进行识别,而随着加密技术的应用,这些特征已经很难提取。而基于行为特征统计的识别方法则放弃了数据包内容,仅提取 IP 包头信息,得到各种统计数据,并利用模式识别方法进行非 P2P 应用以及 P2P 各种应用的分类。

1.1 基本概念

本文采用数据包的传输特性进行研究,但单个数据包携带的信息太少,不足以对网络协议的分类产生较大作用,需要将单个数据包与和它有联系的其他数据包的信息相融合才能够形成某个协议的属性特征,因此此处引入了会话流的概念。

定义 1 TCP 会话流。如果主机 A 与 B 之间通过 TCP 协议进行通信,其通信端口分别为 portA、portB。设 p 为一条 TCP 会话流,则 p 包含主机 A 与 B 的端口 portA、portB 之间在一个 SYN 包发起之后,第一个 FIN 包发起之前这两个端口之间所传输的所有数据包,若有多个 SYN 包,则以第一个为准,其他的均丢弃。

定义 2 UDP 会话流。如果主机 C 与 D 之间通过 UDP 协

收稿日期: 2010-11-23; **修回日期**: 2011-01-16 **基金项目**: 国家"973"计划重点基础研究发展资助项目(2007CB311106);国防重点实验室基金资助项目(NEUL20090101)

作者简介:毛灵(1988-),男,四川岳池人,硕士研究生,主要研究方向为信息安全、P2P(maoling1988@163.com);陈兴蜀(1968-),女,四川成都人,教授,博导,博士,主要研究方向为信息安全、计算机网络;吴仲光(1955-),男,四川遂宁人,副教授,主要研究方向为自动控制;谭骏(1985-),男,重庆人,博士研究生,主要研究方向为信息安全、P2P、机器智能;杜敏(1986-),男,陕西宝鸡人,硕士研究生,主要研究方向为信息安全.

议进行通信,其通信端口分别为 portC、portD。设 p 为一条 UDP 会话流,则 p 包含主机 C 与 D 的端口 portC、portD 之间在 t 时刻内在这两个端口之间所传输的所有数据包(在本文的研究中选择 t 为 60 s)。

TCP 会话流与 UDP 会话流统称为会话流。

定义 3　会话流属性特征向量。设 $s(p)$ 为一条会话流 p 所表现的一个属性特征,则 $S^m(p)=[s_1(p),s_2(p),\cdots,s_m(p)]^{\mathrm{T}}$ 为一条会话流 p 的 m 个属性的特征矢量。部分属性特征如表 1 所示。

表 1　部分属性特征

序号	各种属性	序号	各种属性
1	流持续时间	5	数据包发送时间间隔
2	流发送包个数	6	数据包接收时间间隔
3	流接收包个数	7	ACK 包的个数
4	流接收的字节数	8	乱序包的个数

1.2　系统框架

系统整体设计情况如下:

a)首先采集数据样本。对 P2P 与非 P2P 应用的数据包进行采集并按流属性进行统计。

b)数据预处理。去噪等处理。

c)特征选择。选取出 P2P 和非 P2P 应用中对分类影响较大的一些特征属性。

d)分类器设计及优化。利用粒子群算法优化支持向量机参数,根据学习样本数据,测试不同参数下分类准确率,迭代出最优参数。分类器主要包括训练与测试阶段。训练阶段通过支持向量机找出学习样本数据中的分类面,以及支持向量,得到协议特征库。测试阶段则是利用协议特征库中的分类面,将测试数据进行分类,得到最终分类结果。

流程如图 1 所示。

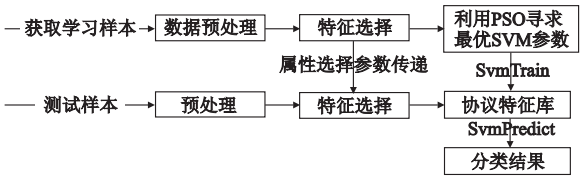

图1　协议识别流程

2　基于 SVM 的分类器算法设计

2.1　支持向量机

支持向量机(support vector machine)是 Cortes 和 Vapnik 于 1995 年首先提出的,它在解决小样本、非线性及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中[8]。

支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度,accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中,以期获得最好的泛化能力。

支持向量机通过寻找不同类之间的支持向量,以较少的样本数据,获得较好的分类能力。

假设训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(R_n,y)^l$,其中 $x_i\in R_n,y_i\in y=\{-1,1\}$(针对二分类问题),$i=1,\cdots,l$。支持向量机的目的就是寻找最优的分类面 H(图 2):

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

使得分类间隔 $2/\|\omega\|$ 达到最大值,即求 $\|\omega\|$ 的最小值。

由于在实际中几乎不能实现线性分类,在分类样本中存在一些个别错误样本(图 3),因此需要将原来的样本进行升维处理,并且引入松弛变量。

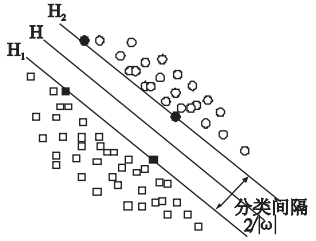

图2　支持向量机模型

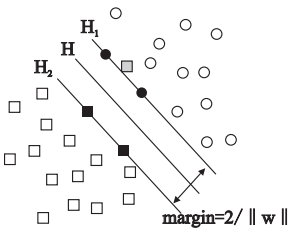

图3　错误样本处理

最终得到的 SVM 分类计算公式如下:

$$\min\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.}\quad y_i[(wx_i)+b]\geqslant 1-\xi_i(i=1,2,\cdots,n)\quad \xi_i\geqslant 0$$

其中:$w=\sum_{i=1}^{n}(a_iy_ix_i)$,$\langle w,x\rangle+b=\langle\sum_{i=1}^{n}(a_iy_ix_i),x\rangle+b$。

采用的 RBF 核函数为

$$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2),\gamma>0$$

2.2　分类器设计算法

根据协议识别流程,在还原数据流并统计其特征后,得到每一条流对应的数据向量,经过特征选择后,选择出对分类影响较大的特征属性。系统的核心就是分类器,根据输入的流特征向量计算并判断出其所属的应用。

鉴于选取的属性特征较多,而 SVM 对高维运算以及分类有着无可比拟的优势,因此选择支持向量机作为分类器的设计思想。

算法 1　分类器算法:样本训练与分类预测

输入:特征选择后的训练样本及测试样本,参数 C 和 γ 值。

输出:测试样本预测分类结果。

a)根据得到的 C 和 γ,得到训练函数中 options 表达式

cmd = ['-c',num2str(bestc),'-g',num2str(bestg)]

b)学习样本训练

model = svmtrain(train_label,train,cmd)

c)测试样本分类预测

[predict_label,accuracy] = svmpredict(test_label,test,model)

其中:train_label、test_label 分别表示训练样本和测试样本的权值,代表属于不同的类;train、test 分别表示训练样本和测试样本;cmd 表示训练的一些参数,包括 C、γ;model 表示训练模型特征表示;predict_label 为测试样本实验分类结果;accuracy 表示测试样本分类准确率。

3　分类器优化

在进行 SVM 分类计算中,需要事先确定惩罚因子 C 以及核函数 RBF 的参数 γ。但是不同的 C 和 γ 值对分类的结果影响巨大,在以往主要是按经验方式选取参数,结果具有不确定性。因此需要通过一定的规则来寻找实现最优分类的 C 和 γ 值。下面将介绍三种不同的 SVM 参数优化算法:Grid Search、GA(遗传算法)、PSO(粒子群算法)。

3.1　基于 Grid Search 的优化

Grid Search 的主要思想是:将 C 和 γ 的函数表达式分别作为直角坐标系中的 X、Y 轴,在一定的范围内,选择一定的步长,依次计算不同 C 和 γ 对应的结果,然后根据结果来选择最优的参数。

首先根据经验选取参数的取值范围,如 $2^{-4}\sim2^4$;然后再选择步长,即参数的每一次变化的差距值,如 0.5;最后将所有可能的选择依次遍历,利用 SVM 特有的交叉验证 cross validation 计算结果,通过最后的比较,得出最佳参数。根据以往实验,一般在结果相同的情况下,选择 C 值较小的参数对。

算法 2　利用 Grid Search 优化 SVM 参数算法

输入:经过特征选择的训练样本 train,训练样本对应的权值 train_label。

输出:SVM 最优参数 bestc(C)、bestg(γ)。

a)参数初始化

设定 C、g 的变化范围:$c_{\min} \leqslant C \leqslant c_{\max}$,$g_{\min} \leqslant g \leqslant g_{\max}$;以及 C、g 的变化步长 c_{step},g_{step};

b)遍历计算适应度

for($C = c_{\min}$;$C \leqslant c_{\max}$;$C += c_{\text{step}}$)

{

c)for($g = g_{\min}$;$g \leqslant g_{\max}$;$g += g_{\text{step}}$)

d){fitness(i) = svmtrain(train_label,train,cnd)}

计算不同 C、g 对应的适应度,其中 cmd 是由 C、g 组成的函数;

}

e)max(fitness(i));根据计算得到的适应度找出其中的最大值,并记录其对应的 C、g 的值,即为 bestc 和 bestg

3.2　基于 GA 的优化

遗传算法(genetic algorithm)是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法。

遗传算法的基本运算过程如下:

a)初始化。设置进化代数计数器 $t = 0$,设置最大进化代数 T,随机生成 M 个个体作为初始群体 $P(0)$。

b)个体评价。计算群体 $P(t)$ 中各个个体的适应度。

c)选择运算。将选择算子作用于群体。选择的目的是把优化的个体直接遗传到下一代或通过配对交叉产生新的个体再遗传到下一代。选择操作是建立在群体中个体的适应度评估基础上的。

d)交叉运算。将交叉算子作用于群体。所谓交叉是指把两个父代个体的部分结构加以替换重组而生成新个体的操作。遗传算法中起核心作用的就是交叉算子。

e)变异运算。将变异算子作用于群体。即是对群体中的个体串的某些基因位上的基因值作变动。

f)终止条件判断。若 $t > T$,则以进化过程中所得到的具有最大适应度个体作为最优解输出,终止计算。

3.3　基于 PSO 的优化

粒子群算法(PSO)最早是在 1995 年由 Eberhart 和 Kennedy 共同提出的,其基本思想是受他们早期对许多鸟类的群体行为进行建模与仿真研究结果的启发,他们的模型及仿真算法主要利用了生物学家 Hepper 的模型[9]。

粒子群算法是非常依赖于随机的过程,这是与优化规划的相似之处,此算法中朝全局最优和局部最优靠近的调整非常类似于遗传算法的交叉算子。此算法也使用了适应度的概念,这是所有进化计算方法所共有的特征。

设 $z_i = (z_{i1}, z_{i2}, \cdots, z_{iD})$ 为第 i 个粒子($i = 1,2,\cdots,m$)的 D 维位置矢量,根据事先设定的适应值函数计算 z_i 当前的适应值,即可衡量粒子位置的优劣;$v_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$ 为粒子 i 的飞行速度,即粒子移动的距离;$p_i = (p_{i1}, p_{i2}, \cdots, p_{id}, \cdots, p_{iD})$ 为粒子迄今为止搜索到的最优位置;$p_g = (p_{g1}, p_{g2}, \cdots, p_{gd}, \cdots, p_{gD})$ 为整个粒子群迄今为止搜索到的最优位置。

在每次迭代中,粒子根据以下公式更新速度和位置:

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1 r_1 (p_{id} - z_{id}^{k}) + c_2 r_2 (p_{gd} - z_{id}^{k})$$

$$z_{id}^{k+1} = z_{id}^{k} + v_{id}^{k+1}$$

其中:$i = 1,2,\cdots,m$;$d = 1,2,\cdots D$;ω 为惯性权重;k 为迭代次数;r_1 和 r_2 为[0,1]之间的随机数,两个参数是用来保持群体的多样性;c_1 和 c_2 为学习因子,也称加速因子,其作用是使粒子具有自我总结和向群体中优秀个体学习的能力,从而向自己的历史最优点以及群体内历史最优点靠近。

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{k_{\max}} \times k$$

其中:$\omega_{\max}$ 为初始权重;$\omega_{\min}$ 为最终权重;$k_{\max}$ 为最大迭代次数;k 为当前迭代次数。这个函数使粒子群算法在刚开始的时候倾向于全局搜索的开掘,然后逐渐转向于开拓,从而在局部区域调整解,这种改进使得粒子群算法的性能得到很大的提高。

算法 3　利用 PSO 优化 SVM 参数算法

输入:经过特征选择的训练样本 train,训练样本对应的权值 train_label。

输出:SVM 参数最优解 bestc(C)、bestg(γ)。

a)参数初始化

确定粒子群算法的加速因子 c_1、c_2,种群规模 sizepop,进化次数 maxgen,以及 SVM 的参数。

b)生成初始粒子和速度

for i = 1:sizepop do

生成初始随机种群;

生成初始化速度;

计算初始适应度

jitness(i) = svmtrain(train_label,train,cmd)得到初始 fitness 矩阵;

end;

c)找极值及极值点

找出全局极值 global_fitness = min(fitness)

及对应极值点 global_x;

个体极值初始化 local_fitness = fitness

及极值点 local_x 初始化;

d)迭代寻优

for i = 1:maxgen

　for j = 1:sizepop

　do

　　速度更新;

　　种群更新;

　　计算更新后适应度;

　end

　个体最优更新;

　群体最优更新;

end

e)根据 d)中迭代寻优,获取最优值 bestc、bestg

4　实验结果及分析

4.1　实验环境

为了进一步验证系统的可行性,笔者选择从四川大学教育科研网上获取一系列数据进行实验。实验环境搭建如图 4 所示。

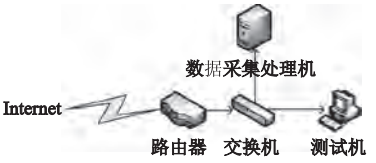

图4　数据采集环境

如图 4 所示,测试机通过交换机、路由器连入四川大学教育科研网。同时在交换机上设置端口镜像,将测试机上流入流出的数据完全复制并传递到数据采集处理机上,然后在数据采集机上使用抓包软件进行数据的抓取,以及数据的后续处理。

为了获得较纯净的数据,实验中选择在测试机上一次只运行一个程序,分别为 Web、BT、Emule。由于抓取到的数据以数据包的形式组织在一起,因此需要根据数据包的五元组信息将数据包重组为数据流,并对数据流中的属性特征进行相应的统计。

4.2　样本训练

在获取的数据流中,每一种协议选择 1 200 条数据流,其中 800 条数据流作为训练样本,另外 400 条作为测试样本。根据已经得知的数据及其所属分类,利用算法 1 进行 SVM 最优参数计算,然后利用算法 2 中 svmtrain 进行支持向量机训练,得到数据行为特征库。

4.3　样本预测

根据得到的数据行为特征库,利用 svmpredict 对测试样本进行预测,得到测试分类结果,并与其真实所属分类进行对比,计算出预测分类准确率。

针对某一协议,其数据流分为 TCP 和 UDP 数据流,若在获取数据中,TCP 所占比例为 p_1,则 UDP 比例为 $1-p_1$;其中 TCP 的分类识别率为 q_1,UDP 识别率为 q_2,则该协议最终的分类准确率为 $p_1q_1+(1-p_1)q_2$。

4.4　计算结果与分析

通过上述分析,对常见的非 P2P 与 P2P 进行了相关测试,选择了 Web、BT、Emule 三种协议。

通过测试,首先根据不同的特征利用优化算法得到各自的最优参数。采用 Grid Search 的参数优化结果如图 5 所示。

根据图 5 显示,区域内的数字表示在对应 c 和 g 的参数作用下能达到的最高分类识别率。在 $\log_2 c=4(c=16)$,$\log_2 g=0(g=1)$的情况下,最高识别率能达到 96.11%。

采用 GA 遗传算法的参数优化结果如图 6 所示。

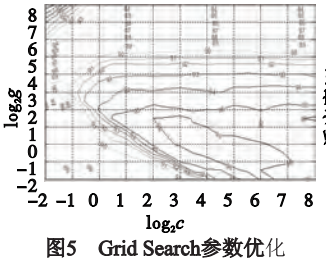

图5　Grid Search参数优化

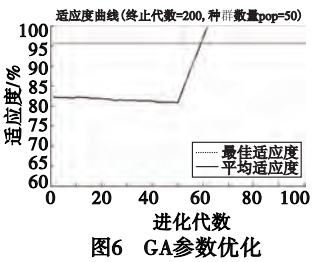

图6　GA参数优化

通过结果分析,在进化到 60 代以后,平均适应度开始趋于稳定。采用 PSO 的参数优化结果如图 7 所示。通过结果分析,在进化到 150 代以后,适应度开始趋于稳定,最大值达到 95.8%。

在同样的输入条件,使用三种不同的优化方法,得到各自最优的参数后,所得分类识别率结果如图 8 所示。

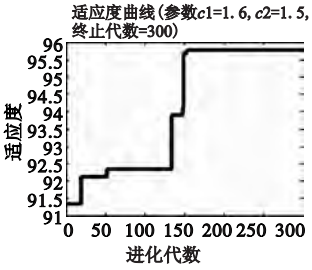

图7　PSO参数优化

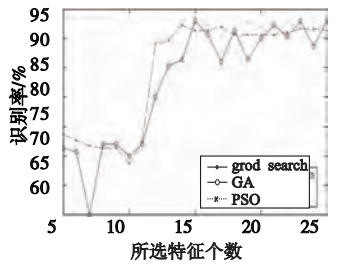

图8　三种优化算法识别率对比

通过对三种不同优化 SVM 参数的方法进行分类研究,根据结果可以得出:

a)分类识别率。使用 GA 遗传算法的分类识别率最低,Grid Search 次之,PSO 最高。

b)识别率的稳定性。使用 Grid Search 在稳定区域相对平滑,PSO 次之,GA 最差。

c)识别收敛。GA 算法在选择属性达到 15 个或以上才开始到达一个相对稳定的区域,而 PSO 和 Grid Search 则在选择 12 个属性后开始趋于稳定。

其中 GA 算法稳定后主要在 90% 左右波动,而且对选择的属性要求较高,当相差一种统计属性时,其结果波动较大。

使用 Grid Search 与 PSO 效果基本一致,但是使用 Grid Search 算法有一个前提条件,就是其参数设定范围是根据以往实践经验来获取的,而且参数值都是按照一定步长增加的,属于离散数据,优化结果一般都接近最优值,而不是等于最优值,这也是影响分类的一个因素。而使用 PSO 优化算法,其初始条件则没有这些限制。

之后,使用不优化的 SVM 进行分类,与使用优化算法的 SVM 分类结果进行对比,如图 9 所示。

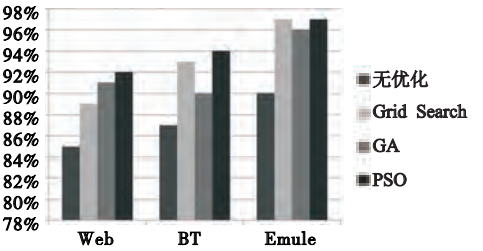

图9　不同优化算法对于不同应用的识别率对比

通过图 9 可以看到,无论 P2P 还是非 P2P,使用任意一种优化算法后的分类识别率均高于未优化的分类测试。

经过一系列的测试,使用 PSO 优化的 SVM 进行分类器设计,可以避开数据加密的难题,很好地识别出网络中的各种 P2P 与非 P2P 应用。

5　结束语

P2P 应用越来越广泛,对 P2P 研究也是近几年的热点。随着技术的发展,越来越多的研究开始倾向于基于统计、模式识别和神经网络等方向发展。但是理论始终走在前面,大部分的 P2P 监控实际应用还处于基于端口号、基于应用层特征码进行 P2P 协议识别的阶段。

本文通过统计知识,基于网络层,统计出以会话流为单位的网络数据包各种属性。利用模式识别原理,在数据分类中采用支持向量机,能很好地支持高维数据的快速分类,实现各种 P2P 应用的协议识别。这在协议识别领域,特别是 P2P 方面的协议识别,具有很大的创新性。在未来的研究中,笔者准备利用加权等方式对分类算法进行改进,同时研究找出更多的反映不同应用的数据包属性,达到利用更少的属性,获取更高的分类准确率,同时考虑在实际网络环境中数据包监管的实时性。

参考文献:

[1] BEN A N, GUILLEMIN F. Impact of peer-to-peer applications on wide area network traffic: an experimental approach[C]//Proc of IEEE Global Telecommunications Conference. Texas: IEEE Press, 2004: 1544-1548.

[2] KIM J T, PARK H K, PAIK E H. Security issues in peer-to-peer systems[C]//Proc of the 7th International Conference on Advanced Communications Technology. Phoenix: IEEE Press, 2005:1059-1063.

[3] SEN S, WANG Jia. Analyzing peer-to-peer traffic across large networks [J]. IEEE/ACM Trans on Networking, 2004,2(2):219-232.

[4] SEN S, SPATSCHECK O, WANG D. Accurate, scalable in-network identification of P2P traffic using application signatures[C]// Proc of ACM WWW'04. New York:ACM Press, 2004:512-521.

[5] KARAGIANNIS T, BRPODP A. Transport layer identification of P2P traffic[C]//Proc of the 4th ACM SIGCOMM Conference on Internet Measurement. Sicily: ACM Press, 2004:121-134.

[6] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]//Proc of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. Alberta:ACM Press, 2005:50-60.

[7] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,2004.

[8] 邓乃扬,田英杰. 支持向量机——理论、算法与拓展[M]. 北京:科学出版社,2009.

[9] 纪震,廖惠连,吴青华. 粒子群算法及应用[M]. 北京:科学出版社, 2009.

[10] XU Ke, ZHANG Ming, YE Ming-jiang, *et al*. Identify P2P traffic by inspecting data transfer behavior[J]. Computer Communications, 2010,33(10):1141-1150.

[11] 王素丽,高洁,孙新德. MATLAB 混合编程与工程应用[M]. 北京:清华大学出版社,2008.