第46卷 第2期 西 安 交 通 大 学 学 报 Vol. 46 No. 2
2012年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2012

采用两阶段策略模型(KTSVM)的P2P流量识别方法

丁要军[1,2]，蔡皖东[1]
(1. 西北工业大学计算机学院，710129，西安；2. 咸阳师范学院信息工程学院，712000，陕西咸阳)

摘要：针对识别加密P2P网络流量比较困难的问题，提出一种基于K均值和直推式支持向量机(TSVM)的半监督学习模型——两阶段策略模型(KTSVM，*k*-means based transductive support vector machine)，以提高P2P流量的识别精度. 该模型首先使用K均值半监督聚类算法计算训练集中正例样本的数目，然后根据正例样本的数目来训练TSVM分类模型，提高了TSVM模型的稳定性和准确性. 该模型的优势是可以使用未标注样本和标注样本共同训练分类模型，非常适合于识别标注比较困难的P2P流量. 实验结果表明，在标注样本较少的情况下，该模型的识别精度和稳定性均优于TSVM模型和SVM模型.

关键词：直推式支持向量机；半监督学习；流量识别；对等网络流量；互联网
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2012)02-0045-06

P2P Traffic Identification via *k*-Means Based Transductive Support Vetor Machine

DING Yaojun[1,2]，CAI Wandong[1]
(1. School of Computer Science and Technology，Northwestern Polytechnical University，Xi′an 710129，China；
2. School of Information Engineering，Xianyang Normal University，Xianyang，Shaanxi 712000，China)

Abstract：A new semi-supervised learning model based on *k*-means and transductive support vector machine is proposed to improve the accuracy of P2P traffic identification. The semi-supervised cluster algorithm of *k*-means is used to calculate the number of positive instances in a training set，and then the TSVM model is trained based on the number of positive instances. So the stability and accuracy of TSVM are improved. An important advantage of the model is that the model can be trained by both labeled samples and unlabeled samples，and the model is suitable for the identification of P2P traffic that is difficult to be labeled. Experimental results show that the proposed model is better than TSVM and SVM models in accuracy and stability，and that it is an effective way to improve the accuracy of P2P traffic identification.

Keywords：transductive support vector machine；semi-supervised learning；traffic identification；peer-to-peer traffic；Internet

随着P2P(peer-to-peer)网络技术的广泛应用，P2P已经取代WWW应用成为占用带宽最多的应用协议，P2P流量在中国教育科研网的骨干网上所占的比例已经从过去的0.76%激增到70%左右[1].

收稿日期：2011-06-30. 作者简介：丁要军(1980—)，男，博士生；蔡皖东(通信作者)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2009AA01Z424)；西北工业大学基础研究基金资助项目(JC201149)；咸阳师范学院专项科研基金资助项目(07XSYK268).
网络出版时间：2011-12-01 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20111201.1640.001.html

因此,对 P2P 流量的识别和控制具有重要意义.

传统的 P2P 流量检测主要采用 DPI(deep packet inspection)方法,具有明显的局限性. DPI 方法依据报文应用层中的特征字段来检测,存在两方面的缺陷:一方面,DPI 方法只能识别已知协议特征的 P2P 流量;另一方面,它无法识别加密协议的流量,还牵涉到侵犯用户隐私的问题. 随着 P2P 软件和协议的不断升级,以及加密机制的广泛应用,DPI 方法的检测效率将大大降低. 近几年,基于连接行为和机器学习的 P2P 流量识别方法成为国内外的研究热点. 目前的成果中大多使用监督学习模型对流量进行识别,需要大量的标注数据来训练模型,但是随着加密机制的广泛应用,特别是 P2P 协议的应用,导致很难获取大量的标注数据. 为此,本文提出一种半监督的机器学习模型,可以借助大量未标注数据来训练模型,非常适合于对标注困难的 P2P 流量进行识别.

1 相关工作

Karagiannis 等人[2]提出了基于传输层行为的 P2P 流量识别方法 BLINC(blind classification),该方法基于{IP,Port}对、传输协议类型等 P2P 协议的连接特征来识别 P2P 协议,由于不依赖于知名 P2P 端口号和应用层特征字段,所以能识别加密网络流和未知 P2P 协议. 但是,因为在不同的网络环境中网络的连接状况差异较大,所以 BLINC 方法的稳定性不好. Moore 等人[3]提出了基于大量传输层特征的朴素贝叶斯模型的流量分类方法,提取了传输层的 248 个统计特征,使用实际流量数据对模型进行训练,对常用协议有很好的分类效果. 但是,由于朴素贝叶斯方法是基于各项属性条件独立的前提,而且需要对大量网络流进行标注以组成训练集,因此代价较高,扩展性不好. 徐鹏等人[4]提出了基于支持向量机(SVM)的流量分类方法,能有效降低冗余属性的干扰,而且不依赖于贝叶斯方法中的先验概率,有很好的分类准确率和稳定性,但是缺点同样在于需要大量标注好的网络流进行训练,而且标注的准确性直接影响最后的分类准确率. 有监督学习方法中一般都采用 l7-fileter[5]来实现训练集的标注. l7-fileter 根据应用层特征字段匹配来识别协议,随着 P2P 协议的不断更新和升级以及加密技术的广泛应用,l7-fileter 的准确性变得无法保障. 如果无法获得一定数量的标注准确的训练集,所有的有监督学习方法的检测准确率都无法保障,因此提出基于半监督学习方法的 P2P 流量识别有一定的意义.

传统的机器学习方法可以分为有监督学习方法和无监督学习方法. 有监督学习首先对已标注的训练集进行学习,训练出分类模型,然后对未标注样本进行类别预测,也就是分类,主要算法有贝叶斯、SVM、决策树分类等. 无监督学习方法完全使用未标注样本进行学习,分组后再人工标记分组结果,主要的代表是聚类算法.

半监督学习介于两者之间,同时使用标注样本和未标注数据进行学习,比较有代表性的有 TSVM(transductive support vector machine)、Co-Training 等算法. 在 Internet 流量分类中,随着加密技术的广泛应用,对网络流的协议标注越来越困难,半监督学习的优势由此得以体现. 半监督学习只需要少量标注网络流,通过引入大量未标注网络流到训练集中来提高检测精度. Erman 等人[6]首次提出将半监督学习方法引入到 Internet 流量分类中,先使用 K 均值算法对标注样本和未标注样本进行聚类,然后根据聚类结果簇中标注样本的标签来标记未标注样本. 可以认为这是聚类方法的一种扩展,但最终都需要进行手工标记.

本文提出一种 KTSVM(K 均值 TSVM)算法,以实现 P2P 流量的识别. 此算法结合了聚类算法和 TSVM 的优势,实验结果表明,在标注样本不充分的情况下,其识别精度明显优于 TSVM、SVM 和半监督聚类等算法.

2 相关概念和原理

P2P 流量识别是一个二分类问题. 判断待检测网络流是否属于 P2P 协议,可以抽象描述为:已知一组网络流$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$($\boldsymbol{X}_i$ 是由网络流的统计特征组成的向量),以及相应的协议类别$\{C_1, C_2\}$,P2P 流量识别的目的是训练分类模型 $f: \boldsymbol{X} \rightarrow C$,判断网络流 $\boldsymbol{X}_i$ 是否属于 P2P 协议.

2.1 相关概念

网络流:在一段时间间隔内,具有相同源 IP、目的 IP、源端口、目的端口、传输层协议的网络报文序列称为网络流,这 5 个属性也称为五元组.

单向流:严格按照五元组取值规则划分的报文序列.

双向流:网络通信中,通信双方建立连接后报文的通信是双向的,但只是目的 IP、目的端口、源 IP、源端口之间的简单互换,这些报文序列原则上也属于同一个流,称为双向流.

流统计特征：以流为单位计算出来的统计特征，包括流的报文大小的期望和方差、报文到达时间间隔的期望和方差等.

2.2　TSVM 算法的原理

Joachims[7]将 TSVM 算法应用于文本分类. TSVM 是基于聚类假设的低密度分割算法，下面简单介绍其原理.

给定一组独立同分布的已标注样本点

$$(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n), \quad \boldsymbol{x}_i \in \mathbf{R}^m, \ y_i \in \{-1, +1\} \tag{1}$$

式中：$\boldsymbol{x}_i$ 是标注样本向量；y_i 是样本所属类别，正例用+1 表示，反例用−1 表示.

另一组来自同一分布的未标注样本点为

$$\boldsymbol{x}_1^*, \ \boldsymbol{x}_2^*, \ \boldsymbol{x}_3^*, \cdots, \ \boldsymbol{x}_k^* \tag{2}$$

因为是未标注样本，故所属类别用 y^* 来表示，它的取值不确定，在算法的执行过程中动态变化.

TSVM 的训练过程可用以下的优化问题描述

Minimize over $(y_1^*, \cdots, y_k^*, \boldsymbol{w}, b, \xi_1, \cdots, \xi_n, \xi_1^*, \cdots, \xi_k^*)$

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_j^* \tag{3}$$

Subject to

$$\forall_{i=1}^{n}: \quad y_i[\boldsymbol{w}\boldsymbol{x}_i + b] \geqslant 1 - \xi_i$$

$$\forall_{j=1}^{k}: \quad y_j[\boldsymbol{w}\boldsymbol{x}_j^* + b] \geqslant 1 - \xi_j^*$$

$$\forall_{i=1}^{n}: \quad \xi_i \geqslant 0$$

$$\forall_{j=1}^{k}: \quad \xi_j^* \geqslant 0$$

式中：$\boldsymbol{w}$ 和 b 分别为支持向量机理论中的线性函数的系数向量和常量；ξ_i 和 ξ_j^* 是松弛变量；C 和 C^* 分别是标注样本和未标注样本的影响因子.

在图 1 中，当仅用标注样本训练模型时，分类面如虚线所示；当使用标注样本和未标注样本共同训练模型时，分类面如实线所示. 显然，实线穿越的是样本分布的低密度区域，而虚线穿越的则是高密度区域.

2.3　TSVM 算法的训练过程

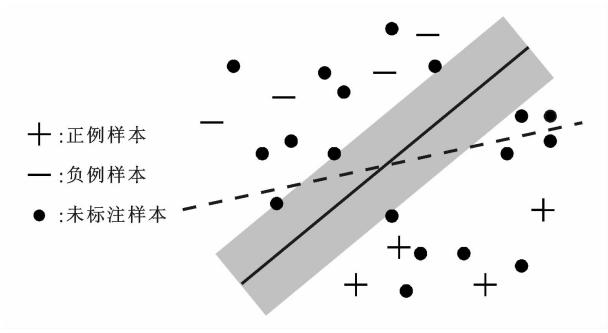
TSVM 模型的训练过程分为以下几个步骤.

(1)指定参数 C 和 C^*，根据训练集中的标注样本进行学习，得到一个初始分类器，并依据特定规则指定测试样本中的正例样本数 N.

(2)用初始分类器对测试样本进行分类，根据判别函数的输出值选取 N 个未标注样本暂时标注为正例，其余标注为反例，并指定一个临时的影响因子 C_{tmp}^*.

(3)根据步骤(2)的标注结果对所有样本重新训练，对新得到的分类器按照特定规则交换一对标注不同的测试样本的标注值，使得式(3)中的目标函数值获得最大下降. 这一步反复执行，直到找不到符合交换条件的样本对为止.

(4)均匀地增加临时影响因子 C_{tmp}^* 的值并返回步骤(3)，当 $C_{\text{tmp}}^* \geqslant C^*$ 时，算法结束，输出结果.



虚线表示仅用标注样本训练模型的分类面；实线表示用标注样本和未标注样本共同训练模型的分类面

图 1　TSVM 算法的分类原理

3　KTSVM 模型

陈毅松等人[8]提出了 PTSVM 算法来改进 TSVM. 这种算法里不事先设定 N 的值，而是在训练过程中根据一定的原则动态地对无标签样本逐一赋予可能的标签，并对新得到的有标签样本重新训练. 然而，在 PTSVM 算法中每次只确定 1 或 2 个未标注样本的标签，然后进行重新训练并验证标注是否正确，增加了迭代次数和算法的复杂度. 我们在此提出一种基于两阶段策略的分类算法，第一阶段采用半监督聚类算法来确定未标注样本中正例样本的数目 N，第二阶段根据确定的 N 值来训练 TSVM 分类模型. 为便于标记，将这种模型命名为 KTSVM(k-means based transductive support vector machine).

3.1　TSVM 算法的缺陷

使用标注数据和未标注数据共同训练分类模型时，TSVM 的性能要明显优于只能使用标注数据训练的 SVM 模型. 然而，TSVM 算法仍然存在一些缺陷，最主要的缺陷在于算法执行之前必须人为设定未标注数据样本中正例样本的数目 N，由于样本未作标注，所以显然无法准确地确定 N 的值. 虽然算法可以根据已标注样本中正例样本的比例来确定 N 的默认值，但是这样确定的 N 值显然是不准确的. 因为是不同的 2 个样本集，标注样本集中正例样

本的比例与未标注样本集中包含的正例样本数没有必然的联系,所以算法给出的默认值也无法保证算法的稳定性. 如果 N 的实际值与算法的默认值差别较大,算法将产生一个不能正确描述样本分布特征的分类模型.

3.2 半监督聚类算法

一般的聚类算法是无监督的,完全使用未标注数据实现聚类,无法利用样本提供的标注信息. 半监督聚类算法可以对标注样本和未标注样本组成的混合数据集进行聚类,有效利用标注数据的信息,提高聚类的精确度. 在 KTSVM 算法中,使用了半监督的 K 均值聚类算法来确定 N 的值,实现过程如下.

(1)将准备的标注样本集和未标注样本集合并,每一条流表示一个样本,合并的样本集就是一组向量,用($\boldsymbol{x}_1$,$\boldsymbol{x}_2$,…,$\boldsymbol{x}_i$,…,$\boldsymbol{x}_j$,…)表示,第 j 个数据流的协议类别用 y_j 表示. 标注样本中的 y_j 已知,而未标注样本中的 y_j 未知. 每一个样本向量由这条流的 m 个统计特征构成:(x_{i1},x_{i2},…,x_{ik},…,x_{im}).

(2)使用 K 均值算法[9]对样本进行聚类,生成 k 个簇(c_1,c_2,…,c_k). 在 K 均值算法中使用欧氏距离来计算样本间的距离,计算公式如下

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Big[\sum_{k=1}^{m}(x_{ik}-x_{jk})^2\Big]^{1/2} \tag{4}$$

假设第 i 个簇中包含 n_i 个样本,其中标注样本数为 n_{il},未标注样本数为 n_{iu},n_{il} 中属于协议类别 y_j 的样本数记为 n_{ilj}.

(3)假设每一个簇中的所有样本都属于同一协议类别,任意簇 c_i 中的数据流属于协议 y_j 的概率表示为 $P(y_j|c_i)$,使用如下公式计算

$$P(y_j\mid c_i)=\frac{n_{ilj}}{n_i} \tag{5}$$

最终 c_i 中所有数据流的协议类别为

$$y=\arg\max(P(y_j\mid c_i)) \tag{6}$$

(4)计算所有被判定为正例协议的簇中未标注样本总数 N 的值

$$N=\sum n_{iu} \tag{7}$$

式中:i 表示被判定为正例协议簇的序号.

3.3 KTSVM 模型的训练

KTSVM 算法的核心在于使用半监督聚类来确定 TSVM 模型中的重要参数 N,完成了 2 个模型的有机结合,实现了优势互补,比单独的模型分类精度更高,稳定性更好. KTSVM 模型的训练步骤如下.

(1)合并训练集和测试集,运行 K 均值聚类算法,使用欧氏距离计算样本间的距离,并根据需要设置簇的数量 k,若是二分类则 k 为 2,生成 2 个簇.

(2)使用半监督聚类算法计算未标注样本中正例样本的数量 N.

(3)指定参数 C 和 C^*,根据训练集中的标注样本进行学习,得到一个初始分类器.

(4)用初始分类器对测试样本进行分类,根据判别函数的输出值以及步骤(2)中计算的 N 的值,选取 N 个未标注样本暂时标注为正例,其余标注为反例,并指定一个临时的影响因子 C^*_{tmp}.

(5)根据步骤(4)的标注结果对所有样本重新训练,对新得到的分类器按照特定规则交换一对标注不同的测试样本的标注值,使得式(5)中的目标函数值获得最大下降. 这一步反复执行,直至找不到符合交换条件的样本对为止.

(6)均匀地增加临时影响因子 C^*_{tmp} 的值并返回步骤(5),当 $C^*_{\text{tmp}}\geqslant C^*$ 时,算法结束,输出结果.

4 实验结果分析

4.1 准备数据集

实验数据是在校园网出口采集的 Internet 流量,随机挑选 50 GB 的 PCAP 数据作为实验数据,使用 OpenDPI[10]对 PCAP 文件进行解析和标注,最后挑选 3 种 P2P 协议的数据流和 HTTP 协议的数据流,如表 1 所示.

表 1 实验数据集

类别	协议	数据流数
正例	BitTorrent	14 811
	eDonkey	14 182
	Thunder/WebThunder	12 253
反例	HTTP	41 346

将表 1 中的数据集分成训练集和测试集,训练集作为标注数据,而测试集作为未标注数据,如表 2 所示.

表 2 训练集和测试集的构成

数据集	正例数	反例数	合计
训练集	4 125	4 135	8 260
测试集	37 121	37 211	74 332

从表 2 可以看出,训练集和测试集的数据流比例接近 1∶9,训练集和测试集中正、反例的比例均接近 1∶1.

考虑到要分别进行 SVM 的实验和 TSVM、

KTSVM 的实验，故需要构造不同的训练集. 在 SVM 实验中将数据集的 BitTorrent、eDonkey 和 Thunder 协议样本标记为＋1，将 HTTP 协议样本标记为－1. TSVM 和 KTSVM 实验中用到的训练集包含标注样本和未标注样本，标注样本中属于 BitTorrent、eDonkey 和 Thunder 协议的样本标记为＋1，属于 HTTP 协议的样本标记为－1，未标注样本标记为 0.

4.2　特征提取

基于机器学习方法的协议识别一般采用网络流传输层上的统计特征. Moore 等人[11]列举出 248 个流统计特征，得到了广泛应用. 本实验中使用一个 248 个特征的约减集[12]，为降低分类模型对端口信息的依赖性，删去了目的端口和源端口 2 个特征，具体特征如表 3 所示.

表 3　网络流统计特征

序号	流特征	描述
1	Push_pkts_serv	TCP 头中具有 push 位的包数(服务器端到客户端)
2	Init_win_bytes_clnt	初始窗口中发送的所有字节数(客户端到服务器端)
3	Init_win_bytes_serv	初始窗口中发送的所有字节数(服务器端到客户端)
4	Avg_seg_size_serv	段的平均大小(服务器端到客户端)
5	Ip_bytes_med_clnt	IP 包中所有字节的中位数(客户端到服务器端)
6	Act_data_pkt_clnt	TCP 数据负载大于 1 字节的包数(客户端到服务器端)
7	Data_bytes_var_serv	数据包中包含的所有字节数的方差(服务器端到客户端)
8	Min_seg_size_clnt	观测到的最小段的大小(客户端到服务器端)
9	RTT_samples_clnt	发现的 RTT 样例的数量(客户端到服务器端)
10	Push_pkts_clnt	TCP 头中具有 push 位的包数(客户端到服务器端)
11	Class	应用协议类别

4.3　实验结果分析

实验平台使用一台个人计算机，CPU 为 Intel Pentium-4，主频为 2.80 GHz，内存为 1 GB，Windows XP 操作系统. 采用 SVM-Light[13]作为实验工具，核函数选用径向基函数(RBF).

(1)对比 KTSVM 算法与 SVM 算法的分类精度. 在测试集不变的情况下，从训练集中按照正例数与反例数为 1∶1 的比例分别挑选 20、40、80、120、160、500、1 000、2 000、4 000、8 000 个样本作为训练集，进行 SVM 分类实验. 将测试集中样本的类别标记设置为 0，并将其作为未标注样本分别添加到 SVM 的训练集中，组成 KTSVM 实验的训练集. 实验结果如图 2 所示.

从图 2 中的结果可以看出，当训练集样本数量不充分时，半监督学习算法 KTSVM 的优势体现得很明显，而当训练集样本数量达到一定程度时，2 种算法的精度相当. 然而在现实中，获取大量标注准确的训练集是一件困难的事情.

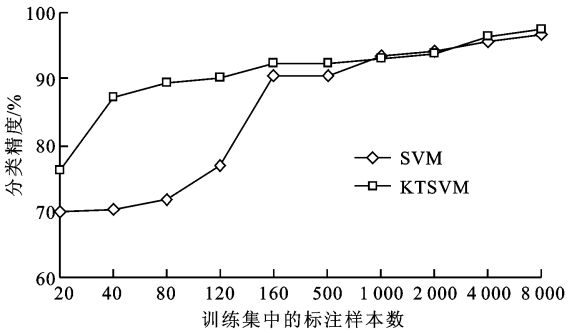


图 2　2 种算法的分类精度与训练集样本数量的关系

(2)比较 KTSVM 算法与 TSVM 算法的分类精度. 在测试集不变的情况下，从训练集中按照正例数与反例数为 9∶1 的比例分别挑选 20、40、80、120、160、500、1 000、2 000、4 000、8 000 个样本，然后将测试集中样本的类别标记设置为 0，并将其与挑选的训练集样本合并，组成半监督实验的训练集. 此时，标注样本中正、反例数的比例为 9∶1，而未标注样本中正、反例数的比例为 1∶1. 实验结果如图 3 所示.

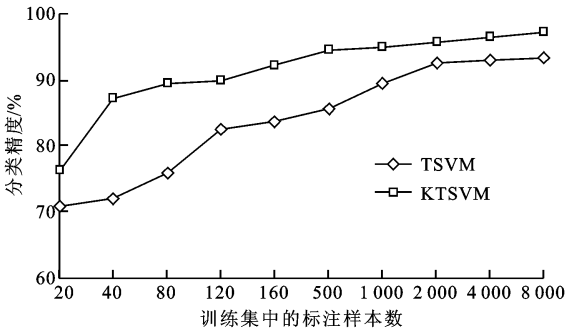


图 3　标注样本中正、反例数的比例为 9∶1 时 2 种算法的分类精度与标注样本数量的关系

从实验结果可以看出，当训练集标注样本中的正、反例数的比例与未标注样本中正、反例数的比例差异较大时，TSVM 算法所确定的 N 值与实际值差距较大，而 KTSVM 算法根据聚类结果确定的 N 值与实际值比较接近. 在一般情况下，未标注样本中正、反例数的比例与标注样本中正、反例数的比例并无直接联系，所以 KTSVM 模型比 TSVM 更加稳定.

为考察算法的运行效率，对比了 KTSVM 和 SVM 两种算法的时间效率，结果见表 4、表 5.

表 4 2 种算法的训练时间

样本数	t_{KTSVM}/s	t_{SVM}/s
20	0.63	0.02
40	1.18	0.05
80	1.34	0.08
120	2.60	0.12
160	2.82	0.13
2 000	9.43	2.33

表 5 2 种算法的测试时间

样本数	t_{KTSVM}/s	t_{SVM}/s
20	0.03	<0.01
40	0.06	<0.01
80	0.08	<0.01
120	0.09	<0.01
160	0.12	<0.01
2 000	0.43	0.03

因为 KTSVM 算法中需要首先使用 K 均值半监督算法确定参数，并且还需要循环增加未标注样本的影响因子和动态更新未标注样本的标签，所以在训练时间上消耗较多，时间效率相对 SVM 稍低. 然而，因为最终的识别环节主要是参照测试时间，所以在识别过程中 KTSVM 算法的效率是可以接受的.

5 结论与展望

使用有监督的机器学习方法进行 Internet 流量分类的成果较多，但随着加密技术的广泛应用以及协议更新速度的不断加快，准确标注 P2P 协议流变得更加困难. 半监督学习方法可以在标注样本数量不充分的情况下，借助未标注数据的信息来提高分类精度，在 P2P 流量识别中具有明显优势. 我们将半监督学习算法 TSVM 引入 P2P 流量分类，提出了 KTSVM 算法，考虑到 TSVM 算法无法准确确定参数 N 的情况，借助半监督聚类来确定更接近实际的 N 值. 从实验结果可以看出，KTSVM 算法的精确度和稳定性更好.

由于条件所限，目前只使用了校园网流量来进行实验，这是因为校园网中 P2P 流量相对较多，更容易发现 P2P 网络流的一些特征. 在下一步的工作中，我们将争取使用骨干网的流量数据进行实验，以发现更多的流量特征，进一步改进算法模型.

参考文献：

[1] 徐鹏，刘琼，林森. 改进的对等网络流量传输层识别方法[J]. 计算机研究与发展，2008，45(5)：794-802.
XU Peng，LIU Qiong，LIN Sen. An improved transport layer identification of peer-to-peer traffic [J]. Journal of Computer Research and Development，2008，45(5)：794-802.

[2] KARAGIANNIS T，PAPAGIANNAKI K，FALOUTSOS M. BLINC：multilevel traffic classification in the dark [C]// Proceedings of the 2005 ACM SIGCOMM Conference. New York，USA：ACM，2005：229-240.

[3] MOORE A W，ZUEV D. Internet traffic classification using Bayesian analysis techniques [C]// Proceedings of the 2005 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems. New York，USA：ACM，2005：50-60.

[4] 徐鹏，刘琼，林森. 基于支持向量机的 Internet 流量分类研究[J]. 计算机研究与发展，2009，46(3)：407-414.
XU Peng，LIU Qiong，LIN Sen. Internet traffic classification using support vector machine [J]. Journal of Computer Research and Development，2009，46(3)：407-414，

[5] Geeknet Inc. L7-filter [EB/OL]. [2010-06-02]. http://l7-filter.sourceforge.net/.

[6] ERMAN J，MAHANTI A，ARLITT M. Semi-supervised network traffic classification [C]// Proceedings of the 2007 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems. San Diego，California，USA：ACM，2007：369-370.

[7] JOACHIMS T. Transductive inference for text classification using support vector machines [C]// Proceedings of the 1999 International Conference on Machine Learning (ICML). New York，USA：ACM，1999：200-209.

[8] 陈毅松，汪国平，董士海. 基于支持向量机的渐进直推式分类学习[J]. 软件学报，2003，14(3)：451-460.
CHEN Yisong，WANG Guoping，DONG Shihai. A progressive transductive inference algorithm based on support vector machine [J]. Journal of Software，2003，14(3)：451-460.

[9] 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.

[10] IPOQUE Company. OpenDPI 1.2.0 [EB/OL]. [2010-06-10]. http://www.opendpi.org/.

[11] MOORE A W，ZUEV D，CROGAN M. Discriminators for use in flow-based classification，RR-05-13 [R]. London，England：University of London，2005：1-14.

[12] LI Wei，CANINI M，MOORE A W. Efficient application identification and the temporal and spatial stability of classification schema [J]. Computer Networks，2009，53(6)：790-809.

[13] JOACHIMS T. SVM-Light [EB/OL]. [2010-06-20]. http://svmlight.joachims.org/.

（编辑　葛赵青　赵大良）