高技术通讯 2010 年 第 20 卷 第 7 期:674 ~ 679

doi:10.3772/j.issn.1002-0470.2010.07.003

一种基于 TCM 主动学习的 P2P 流识别技术①

戴 磊②*** 云晓春** 张永铮*** 吴志刚***

(*中国科学院计算技术研究所 北京 100190)

(**中国科学院研究生院 北京 100039)

摘 要 针对目前基于机器学习的流识别仍然存在着建立分类模型需要大量适用的训练数据,训练数据的标记需要依赖领域专家,因而导致工作量及难度过大和实用性不强的问题,采用主动学习技术提取少量高质量的训练样本进行建模,并结合支持向量机(SVM)分类算法提出了一种基于直推信任机(TCM)的样本筛选方法。实验结果表明,相对于已有的流识别方法,这种方法能够在仅依赖少量高质量训练样本的前提下,保证较高的召回率及较低的误报率,更适用于现实网络环境。

关键词 支持向量机(SVM),主动学习,直推信任机(TCM),机器学习,不确定性采样

0 引 言

P2P 技术能够实现网上的快速高效的直接共享和交互,目前已广泛应用于文件共享、实时通信、流媒体、协同工作等领域。但 P2P 技术的大量使用也带来了相应的问题,已有的研究表明,P2P 流量已占据网络中业务总量的 70%[1],抢占了网络带宽,增大了网络设备负载,降低了网络性能,劣化了网络服务质量,妨碍了正常网络业务的开展和关键应用的普及,因此,必须对 P2P 流量进行识别和控制,这对于网络维护和运营具有重要意义。P2P 流的检测属于流分类领域的研究内容,早期的流分类多依赖应用层端口或检测深层数据包的内容。然而随着 P2P 技术的快速发展,相关的应用不断增加,端口和协议的使用也更加灵活,加密与混淆技术也被应用于 P2P 业务中,基于端口或检查包内容的方法的有效性及性能逐渐降低。由于传统的流分类技术存在弊端,研究者们逐渐转向使用流的统计特性进行流识别的研究,机器学习技术为基于统计特性的流分类提供了一条重要的研究途径,该方法在很大程度上弥补了传统技术的不足,它可以提供较快的分类速度,并且具有对加密流和未知 P2P 流的处理能力。该方法将流的识别问题转化为相应的二分类或多分类问题处理,如基于支持向量机(support vector machine,SVM)算法的流识别方法[3],基于贝叶斯算法的流识别方法[2]等。它们在一定程度上取得了较好的识别效果,但在很大程度上依赖于训练样本集。而在实际应用中,搜集网络中 P2P 数据并对其类别进行正确的标注会具有一定难度并耗费大量人力。针对这种情况,本文提出了一种基于直推信任机(transductive confidence machine,TCM)主动学习的流分类方法,并通过实验证明了其有效性。

1 相关工作

在流分类中使用机器学习技术的思想在文献[4]中的入侵检测中初次提出。文献[5]的作者使用主成分分析(principal component analysis, PCA)和密度估计方法把流分类到不同的应用。文献[6]选取流的 4 个属性作特征,使用最近邻(nearest neighbour, NN)和线性判别式分析(linear discriminate analysis, LDA)方法成功地把网络应用映射到不同的 QoS 类别。文献[7]提出了一种使用无监督的机器学习方法来识别不同网络应用的框架,根据流的统计特征来进行自动分类研究。文献[8]关注流分类的计算性能问题,抽取 26 个流特征,并在特征选择后比较了 5 种分类算法。在与 P2P 流识别相关的研究中,文献[2]将相似的网络应用归为一类业务,并利用核估计的贝叶斯分类器(Naïve Bayes kernel estimation,

① 863 计划(2007AA01Z444,2007AA01Z474,2007AA010501,2007AA01Z467)和国家自然科学基金(60703021,60573134)资助项目。

② 男,1979 年生,博士生;研究方向:计算机网络安全,流分类技术等;联系人,E-mail:dailei@software.ict.ac.cn

(收稿日期:2009-04-03)

NBK)对包括 P2P 流在内的多种业务流进行分类研究。文献[3,9]使用 SVM 分类算法根据不同的流特征对 P2P 流的识别问题进行了研究。文献[10]考虑到实时性问题,根据流的前 4 个包提取统计特征,使用聚类方法区分 P2P 流。

在过去几年中,流分类领域的研究者们通过使用经典的数据挖掘和机器学习方法,推动了流识别技术的发展。

2　主动学习模型

为了克服学习算法对训练样本集的过分依赖,文献[11,12]提出了主动学习的思想,研究表明,主动学习方法已在多个领域得到成功应用。主动学习算法可以由以下 5 个组件进行建模:

$$A = (C, L, S, Q, U) \tag{1}$$

式中,C 为一个或一组分类器;L 为一组已标注的的训练样本集;Q 为查询函数,用于在未标注的样本中查询信息量大的样本;U 为整个未标注样本集;S 为督导者,可以对未标注样本进行标注。主动学习算法分为两个阶段。第一阶段为初始化阶段,随机从未标注样本中选取一小部分,由督导者标注,作为训练集建立初始分类器模型;第二阶段为循环查询阶段,S 从未标注样本集 U 中,按照某种查询标准 Q,选取一定的未标注样本进行标注,并加到训练样本集 L 中,重新训练分类器,直至达到训练停止标准为止。可以看出,主动学习方法的研究目标是通过筛选优质样本,降低督导者对样本类别标注的工作量并减少无用或噪音数据对分类器的负面影响,而在主动学习算法模型中,查询函数是进行样本筛选的主要组件,因此,查询函数的设计是主动学习方法的研究重点。

3　基于主动学习的流分类

精简训练样本有助于减少样本标注工作量,降低算法的计算开销。为了有效筛选高质量样本,本文提出了一种基于主动学习算法的流分类算法,算法中查询函数 Q 的设计基于不确定性采样(uncertainty based sampling, UBS)策略,并以 TCM 置信机制对样本的不确定性进行估测,分类器 C 则使用了以能够处理非线性、小样本、高维度等实际问题而著称的支持向量机[3, 9]。

查询函数设计是主动学习研究的重点,目前主要的查询策略包括不确定性采样(UBS)与投票选择方法(query by committee, QBC)。UBS 策略假设分类器总是选择分类中最不确定的样本交由督导者标注,QBC 策略则首先根据概念在搜索空间的分布获得概率假设,利用这个假设预测样本的标注。QBC 选择投票产生的标签与假设预测的标签不一致的样本交由督导者标注。由于到目前为止还未出现高效实用的 QBC 实现方法[13,14],基于 UBS 策略的查询函数设计仍是主要的研究方向。

3.1　TCM 置信

UBS 策略选取分类器最不确定的样本交由督导者标注,因此对未标注样本的不确定性进行评估是 UBS 策略查询函数设计的基础。评估未标注样本分类不确定性的最好方法是对其分类结果进行置信,直推信任机(TCM)是一种适用范围较广的置信机制[15,16]。TCM 基于随机性检测,只需要满足 iid 假设即可(即待归类的样本以及用于训练的数据集都是独立且同分布的),检测函数的值称为 P 值,P 值定义为待分类样本属于已存在的几类样本空间的概率估测值。其相对于某类样本空间的值越大,则表明它属于该类样本空间的可能性越大。

为了计算 P 值,还需要定义一种称为奇异值(strangeness)的指标。对于 SVM 分类器,其对偶优化模型 A' 为

$$\begin{cases} \max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i^{\mathrm{T}} x_j) \\ s.t. \sum_{i=1}^{l} y_i \alpha_i = 0 \\ 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,l \end{cases} \tag{2}$$

其中 $\alpha_i, i = 1,2,\cdots,l$ 为 lagrange 乘子。优化问题 A' 是一个凸二次规划问题,其局部最优解即为全局最优解。若 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$ 为模型 A' 的最优解,则

$$\omega^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i \tag{3}$$

根据 KKT 互补条件,最优解必满足

$$\alpha_i (y_i (\omega^{\mathrm{T}} x_i + b) - 1 + \xi_i) = 0 \quad i = 1,2,\cdots,l \tag{4}$$

由式(2)–(4)可知,对应于 lagrange 乘子 $\alpha_i = 0$ 的样本对分类问题不起什么作用,而只有对应于 lagrange 乘子 $\alpha_i > 0$ 的样本(支持向量)对计算 ω^* 起作用,从而决定分类结果,并且 α_i 值越小,其样本对分类器影响越小。因此,定义样本作为类别 y 对应于 lagrange 乘子的 α_i 为样本归属类别 y 的奇异值[17]。结

合奇异值,可以给出 P 值的计算方法。

定义1 待检测样本 i 相对于类别 y 的 P 值的计算方法为

$$P_y(\alpha_{iy}) = \frac{\#\{j:\alpha_{jy} \geqslant \alpha_{iy}\}}{n} \tag{5}$$

其中,# 表示训练样本集中满足条件的样本个数;α_{iy} 为待检测样本归属类别 y 的奇异值;α_{jy} 为训练样本集中标注为 y 类的样本奇异值;n 为训练样本集中归属类别 y 的样本个数。不难看出,P 值取值区间为[0,1],并且其值越大,待检测样本的奇异值相对于类别 y 的训练样本越小,样本归属于类别 y 的可能性越大。

3.2 查询函数的设计

对于二分类问题,我们可以利用 TCM 方法为每一个待判定的样本计算两个 P 值,每个 P 值都对应了该样本属于每类样本的置信度(confidence)估测值。将这两个 P 值表示为 P_x 和 P_y,P_x 即样本属于 x 类的置信估计值,而 P_y 即样本属于 y 类的置信估计值,当 $P_x > P_y$ 时,表示该样本更接近于类别 x,而且,P_x 越大且 P_y 越小则表明将样本归作 x 类的结果更为准确和可信,反之,则意味着样本归属类别 y 的可能性更大。不难看出,当 $P_x \approx P_y$ 时,样本归属的不确定性较高。那么,对于基于 UBS 策略的主动学习方法,我们给出如下查询函数(query function):

$$Q(i) = |P_x - P_y| \tag{6}$$

式(4)中,$Q(i)$ 表示 P 值的绝对偏差程度,$Q(i)$ 越小,表示分类器对该样本的预测结果越不确定,从而该样本即为该选择函数选择出来的需要标记且学习的下一个样本。

3.3 主动学习方法的终止策略

主动学习算法进入循环阶段后,需要有相应的策略对其进行终止,以减少样本筛选的执行时间。为了使算法终止后所产生的样本集可以接近最佳训练效果,本文使用均方预测误差率作为循环的终止条件指标。针对训练集 D 在测试样本 x 下的预测函数 f^*,均方预测误差率的计算方法如下:

$$f_{error} = \frac{MSE(f^*(x,D))}{f(x,D)} \tag{7}$$

其中,MSE 用于计算分类错误率的均方预测误差(mean squared prediction error,MSE),f 为使用 SVM 检测算法获取的预测分类错误率。

根据统计学习理论,均方预测误差 MSE 包含偏差(bias)与变化值(variance)两部分,当新的选择样本不断加入,偏差应该快速下降,而变化值也逐渐趋于稳定,此时检测效果往往接近最佳时刻。因此,使用相对的均方预测误差作为终止策略,能够有效限定偏差与变化值范围,使样本选择结果接近最优。通过计算测试集 x 预测分类错误率与实际分类错误率的方差,再取与实际分类错误率比值,可获得终止条件指标,设定经验阈值 ε,当该值小于 ε 时,终止样本选择循环。

3.4 基于 TCM 的主动学习算法

根据上面构造的选择函数以及终止策略,我们给出了针对该算法的主动学习方法伪代码(如图1所示)。在图1中,假定有少量样本已标注好训练集 L 和一个未标注的样本池 U。当算法进入循环阶段时,在每一轮中,按顺序选取样本池 U 中的每一个样本,使用选择函数进行计算,选取其中 $Q(i)$ 为最小值对应的样本 i,请求标注后并将其加入以标记好的训练集 L 中再进行训练,直到满足终止条件均方预测误差率 f_{error} 小于阈值 ε 为止。最后得到的训练集 l 即为选择与精简后的训练集。

```
将训练集初始化为 L 并将 Q_min 置 1;
repeat
{
    for each sample j in U do
    {
        从 U 中按顺序选取样本 j 并计算其 P 值;
        if (Q(j)<Q_min)
        {
            Q_min ← Q(j);
            i ← j;
        }
    }
    将样本 i 加入训练集 L 并从样本池 U 中删除该样本;
    根据测试样本集 x 计算 f_error;
}
until (f_error < ε)
```

算法参数说明:L 为少量以标注样本,U 为大量未标注样本池,x 为测试样本集,ε 为算法终止阈值,Q_{min} 为最小选择函数值

图1 基于 TCM 与 SVM 的主动学习方法

4 实验及分析

为了评估所提出的流分类方法的有效性,进行了大量的实验。在实验中,首先使用特征选择技术降低数据维度,并对该领域著名的核估计贝叶斯分类(NBK)方法、SVM 方法以及经典的 K 近邻(K-nearest neighbors,KNN)方法的检测性能进行了比较,然

后,为了评估特征选择的影响,比较了特征选择优化前后 SVM 算法的检测性能与计算性能的变化,最后,为了验证主动学习方法的有效性,在特征选择优化后,比较了选择相同数量样本的情况下采用主动学习方法与随机采样方法选择样本后分类效果的差异。为了评价 P2P 流的识别效果,实验中使用了两个标准指标:检测率(true positive rate,TP)和误报率(false positive rate,FP)。TP 指被正确检测为待检测类别的样本的数量与测试集这一类样本的总数量的比值,FP 指错误判为待检测类别的样本的数量与测试集中非待检测样本的数量的比值。

4.1　数据集

实验中使用了 Moore 等人专为流分类实验而制作的数据集[18]。在这个数据集中每一个流作为一个样本,样本中包括许多特征,如连接的持续时间、包间隔时间、传输的数据包数量等。每一个流有 248 个特征,表 1 列出了部分特征示例,文献[18]中有对这些特征更详细的描述。

表 1　流特征示例

特征
Flow Duration
TCP Port
Packet inter – arrival time(mean, variance, …)
Payload size(mean, variance, …)
Initial window bytes

除了 248 个特征之外,每个样本还有一个流的应用类型标注,如 WWW, P2P, MAIL, BULK 等。表 2[2]列出了数据集中所有的类,如 BULK 类是由 ftp 流组成的,并且包括了 ftp 控制流与 ftp 数据流。关于流中类详细的描述见文献[18]。

表 2　网络流的类型

(每一个类型下含有多个应用实例)

Classification	Example Application
BULK	ftp
DATABASE	postgres, sqlnet, oracle, ingres
INTERACTIVE	ssh, klogin, rlogin, telnet
MAIL	imap, pop2/3, smtp
SERVICES	X11, dns, ident, ldap, ntp
WWW	www
P2P	KaZaA, BitTorrent, GnuTella
ATTACK	Internet worm and virus attacks
GAMEA	Half – Life
MULTIMEDIA	Windows Media Player, Real

数据集共有 10 组数据,包含大量样本。为了方便处理,我们从中抽取 1676 条 P2P 类别样本以及 6006 条其它类别样本构成实验数据。同时,将其它类别样本归作一类,标注为类别 OTHER,转换为二分类问题进行实验。处理后的获取数据集中各类流的数目如表 3 所示。

表 3　数据集的流数目统计结果

Flow classes	Flow numbers
OTHER	6006
P2P	1676

4.2　特征选择与对比实验

实验数据中的样本包含大量特征,其中存在很多特征对于分类器而言是冗余或无关的,为了提高分类器性能,减少冗余无关特征的影响,实验中首先使用快速相关性特征选择(fast correlation-based filter,FCBF)算法[19]对独立于实验数据的训练数据进行特征选择。表 4 列出了特征选择优化后选取的分类特征。

表 4　FCBF 特征选择选择选取的特征

The selected features in extended feature set
Server port
Maximum of bytes in packet
The maximum number of sack blocks seen in any packet (client(server)
The total number of pure ack packets(client(server)
The total bytes of data found in the retransmitted data(client(server)
The total number of ACK packets seen carry sack information(server(client)
The total number of packets with the URG bit turned on in the TCP header (server(client)
The average full – size RTT sample (server(client)
The standard deviation of full – size RTT samples (client(server)
Third quartile of bytes in packet
Minimum of packet inter – arrival time
FFT of packet IAT (client(server,frequent #3)

在特征选择后,对实验数据使用 5 折交叉验证对各种分类方法进行比较,实验结果如表 5 所示。为了保证评测与比较公平性,实验分别对 SVM,KNN 两种算法挑选不同的参数(NBK 算法无可选参数),取最优结果作比较。其中,SVM 的参数的类型为 CSVM,核函数为径向基函数(radial basic function, RBF),gamma 取值 0.005,C 取值 2048;KNN 算法的

近邻个数为3。不难看出,相对于其它算法,SVM算法提供了较高的检测率及较低的误报率。

表5 对比实验结果

算法	TP(%)	FP(%)
NBK	72.3	0.7
SVM	96.4	1.3
KNN	95.6	1.3

表6是特征选择前后SVM算法的检测结果,可以看出,经过FCBF特征选择后,SVM算法在性能大幅提升,而其检测能力基本未受到负面影响,这表明FCBF特征选择对分类结果的负面影响极小。

表6 特征选择前后SVM分类器实验结果

SVM算法	TP(%)	FP(%)	建模时间(s)	检测时间(s)
原始数据集特征	96.5	1.6	73.16	22.78
选择后数据集	96.4	1.3	4.11	4.83

4.3 采用主动学习方法实验结果

对主动学习方法的评价一般要考虑两方面的效果:(1)某个定量的训练集采用主动学习方法相对于采用被动学习方法所提升的性能的程度;(2)为了达到某个性能指标主动学习方法相对于被动学习方法所需要的训练样本集的精简程度。为了验证基于TCM的主动学习方法的有效性,我们在相同数量样本的情况下,关注了采用主动学习方法与随机采样方法选择样本后分类效果的差异。在实验中,为了确保实验结果的可信性,将实验数据均分成5组,分别利用其中4组作训练样本,剩下1组作测试,这样就得到5组实验结果。在每组实验中都从训练样本中抽出5个P2P样本与5个OTHER样本构成初始训练集 l,将剩余的训练样本作为样本池 U,然后筛选100个样本进行对比。

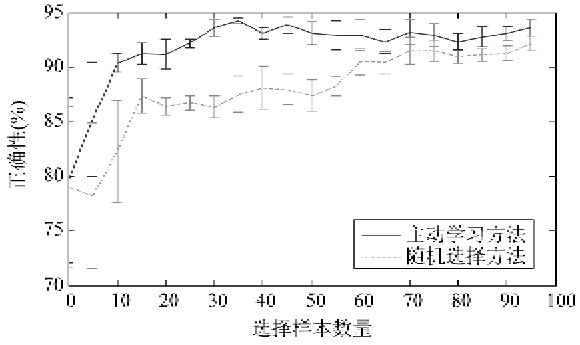

图2 主动学习方法与随机选择方法的实验对比

图2显示了实验中分类准确率的均值与标准差变化。结果很明显,随着选择样本数量的增加,采用基于TCM的主动学习方法后,分类准确率迅速提升,而使用随机选取训练样本的方法分类正确率的提升相对缓慢。这表明,使用所提出的主动学习方法进行样本选择能够有效减少构造P2P检测模型所需训练样本的规模。

4.4 实验结果分析

上述实验充分说明了本文所述方法的有效性:在对比试验中,SVM相对于其它几种常用算法,提供了较高的检测率与较低的误报率,而且使用FCBF算法进行特征选择优化后,在计算性能获得极大提升的同时,检测效果受到的负面影响极为微小;采用基于TCM的主动学习方法选择样本后,算法在使用相同数量训练样本情况下获得了更高的检测性能,从而能够有效控制所需训练样本的规模。

另外,互联网在不断发展变化,需要周期性地获取训练样本,实验表明,采用本文所述的主动学习方法后,能够有效控制需要标注的样本数量,减轻训练样本标注的工作量。

5 结 论

快速准确地识别P2P类型的流对于网络的维护与运营都具有重要意义。利用机器学习技术识别流是当前重要的研究方向之一,该类方法的有效性严重依赖于训练样本的数量与质量,而训练样本的标注是非常困难且耗费人力的工作。本文引入主动学习方法筛选训练样本,能够以少量的训练样本,保证较好的召回率,极大地降低了由样本标注所耗费的工作量以及使用大量训练样本给算法带来的过大计算开销。实验表明:该方法行之有效,具有较高的召回率和较低的误报率,与同领域的有指导P2P识别方法相比也具有相当的优势。但从实用的角度看,此方法还有待进一步改进,如何将其应用于实际网络环境,并依据实际情况加以优化以及将其试用于它类别流的识别,是下一步工作的重点。

参考文献

[1] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification. In: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation, Monterey, USA, 2006. 179-188

[2] David M, Denis Z. Internet traffic classification using bayesian analysis techniques. In: Proceedings of the 2005

ACM SIGMETRICS international conference on Measurement and modeling of computer systems. NY, USA: ACM New York, 2005. 50-60

[3] Yang A M, Jiang S Y, Deng H. A P2P network traffic classification method using SVM. In: Proceedings of the 2008 the 9th International Conference for Young Computer Scientists, Zhangjiajie, China, 2008

[4] Frank J. Machine learning and intrusion detection: current and future directions. In: Proceedings of the National 17th Computer Security Conference. MD, USA, 1994

[5] Dunnigan T, Ostrouchov G. Flow characterization for intrusion detection. Oak Ridge National Laboratory. Tech Rep: 2000. http://www.csm.ornl.gov/~ost/id/tm.ps

[6] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification. In: Proceedings of the 4th ACM SIGCOMM conference on Internet Measurement. NY, USA: ACM New York, 2004. 135-148

[7] Zander S, Nguyen T, Armitage G. Self-learning IP traffic classification based on statistical flow characteristics. In: Proceedings of the 6th Passive and Active Measurement Workshop. Berlin: Springer, 2005. 41-54

[8] Nigel W, Sebastian Z, Grenville A. A preliminary comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 2006, 36(5): 5-16

[9] Deng H, Yang A M, Liu Y D. P2P traffic classification method based on SVM. *Computer Engineering and Application*, 2008, 44(14): 122-126

[10] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Proceedings of The 2nd ADETTI/ISCTE CoNEXT Conference. Lisboa, Portugal, 2006

[11] Herber A S, Glenn L. Problem solving and rule reduction, a unified view. In: Knowledge and Cognition. Erbuam, 1974

[12] Valiant L G. A theory of learnable. *Communications of the ACM*. 1984. 27(11): 1134-1142

[13] Shai F, Gilad-Bachrach R, Eli S. Query by committee, linear separation and random walks. *Theoretical Computer Science*, 2002, 284(1): 25-51

[14] Herbrich R, Graepel T, Campbell C. Bayes point machines. *Journal of Machine Learning Research*, 2001, 1:245-279

[15] Barbara D, Domeniconi C, Rogers J P. Detecting outliers Using transduction and statistical testing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM New York, 2006, 55-64

[16] Proedru K, Nouretdinov I, Vovk V, et al. Transductive confidence machine for pattern recognition. In: Proceedings of the 13th European conference on Machine Learning. London, UK: Springer Berlin, 2002. 381-390

[17] Gammerman A, Vovk V. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 2002, 287: 209-217

[18] Moore A W, Zuev D. Discriminators for use in flow – based classification. Intel Research, Tech Rep: 2005

[19] Yu L, Liu H. Feature selection for high – dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2003. 856-863

Peer-to-Peer traffic identification using TCM based active learning

Dai Lei*,**, Yun Xiaochun*, Zhang Yongzheng*,**, Wu Zhigang*,**

(* Institute of Computing technology, Chinese Academy of Sciences, Beijing 100190)

(** Graduate University of Chinese Academy of Sciences, Beijing 100039)

Abstract

To solve the problem that present identifications of Peer-to-Peer (P2P) traffic based on machine learning are still too difficult and time-consuming and are still unpratical due to the need for obtaining adequate qualified training data for the supervised classifiers to model traffic patterns and the great dependence on the domain experts, in marking the training data, the authors introduce the active learning method to select the most qualified data for training and propose a transductive confidence machine (TCM) based instance selection method for support vector machines (SVM). The experimental results demonstrate that the proposed method is able to guarantee a high recall rate and low false positives by using a small quantity of high qualified data. Therefore, it is more suitable for the real network applications than the traditional ones.

Key words: support vector machines (SVM), active learning, transductive confidence machines (TCM), machine learning, uncertainty based sampling