# 一种主动学习式 P2P 流识别方法

戴 磊, 王 源, 刘科科
(中国电子科学研究院, 北京 100041)

**摘 要**: P2P 流的识别对于网络的维护与运营都具有重要意义,基于机器学习的流识别技术是目前研究的热点和难点内容,但目前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以及因此而导致的工作量及难度过大和实用性不强等问题,而当前的研究工作很少涉及到这些问题的解决办法。针对这一问题,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 SVM 分类算法提出了一种基于锦标赛选择的样本筛选方法。实验结果表明,其相对于已有的流识别方法,能够在仅依赖少量高质量训练样本的前提下,保证较高召回率及较低误报率,更适用于现实网络环境。

**关键词**: 支持向量机; 主动学习; 机器学习
**中图分类号**: TP393 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0717-05
doi:10.3969/j.issn.1001-3695.2012.02.085

# Peer-to-peer traffic identification using active learning

DAI Lei, WANG Yuan, LIU Ke-ke
(*China Academy of Electronics & Information Technology*, *Beijing* 100041, *China*)

**Abstract**: The identification of peer-to-peer(P2P) traffic is essential to network maintenance and management. Traffic identification using machine learning (ML) techniques has been an active and difficult research topic for recent years. However, there are still some unresolved and scarcely addressed problems such as the difficulties in obtaining adequate qualified training data for the supervised classifiers to model traffic patterns, the data acquisition task is always time-consuming and greatly relies on the domain experts. To solve the problem, this paper introduced active learning method to select the most qualified data for training and proposed a tournament selection based instance selection method for SVM. The experimental results demonstrate that the proposed method is able to guarantee 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**: SVM(support vector machine); active learning; machine learning

## 0 引言

P2P 技术能够提供快速高效的文件共享、强大的网络连通性、低成本高可用的计算资源共享,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。P2P 技术的大量使用也带来了相应的问题,已有研究表明,P2P 流量已占据网络中业务总量 70%[1],其业务抢占了大量网络带宽,增大网络设备负载,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,对 P2P 流量进行识别和控制对于网络维护和运营具有重要意义。

P2P 流的检测问题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换为相应的二分类或多分类问题处理,如基于 SVM 算法的流识别方法[2,3]、基于贝叶斯算法的流识别方法[4]等。它们在一定程度上取得了较好的识别效果,但这些方法的效果都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标注是具有一定难度并耗费大量人力的工作。因此,如何利用少量的训练样本,保证 P2P 流的检测效率,具有重要研究价值。

## 1 相关工作

随着互联网的快速发展,传统的流识别方法呈现出许多弊端,难以满足应用需求。近年来,研究者逐渐把目光聚集在基于机器学习技术的流分类研究。

对流分类使用机器学习技术的思想在文献[5]的入侵检测中初次提出。文献[6]的作者使用主成分分析(principal component analysis, PCA)和密度估计(density estimation)方法把流分类到不同的应用。文献[7]选取流的四个属性作为特征,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

**收稿日期**: 2011-05-14; **修回日期**: 2011-08-08

**作者简介**: 戴磊(1979-),男,江苏徐州人,工程师,博士,主要研究方向为数据挖掘、计算机网络(daidaoxin@163.com);王源(1968-),男,高级工程师,博士,主要研究方向为网络安全、软件工程;刘科科(1981-),男,工程师,硕士,主要研究方向为计算机网络.

到不同的 QoS 类别。文献[8]提出了一种使用神经网络技术进行流分类的方法。在文献[9]中,作者关注了流分类的计算性能问题,他们抽取 26 个流特征,并在特征选择后比较了五种分类算法。在与 P2P 流识别相关的研究中,文献[4]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, NBK)对于包括 P2P 流在内的多种业务流进行分类研究。在文献[2,3]中,作者都使用SVM 分类算法根据不同的流特征对 P2P 流的识别问题进行研究。在文献[10]中,作者提取流内容,使用多指纹方法区分 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,产生了主动学习的思想[11,12]。国内外相关研究表明,主动学习方法已在多个领域成功应用[13,14]。主动学习算法可以由以下五个组件进行建模:

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

其中:$C$ 为一个或一组分类器;$L$ 为一组已标注的训练样本集;$Q$ 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。主动学习算法主要分为两阶段: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声数据对于分类器的负面影响。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

在实际网络环境中,搜集 P2P 数据并对其类别进行正确标注,是具有一定难度并耗费大量人力的工作。而精简训练样本则有助于减少样本标注工作量,降低算法的计算开销。为了筛选高质量流样本建立流分类器模型,本文将主动学习方法引入到流分类研究中,并分析了主动学习方法中存在的早熟收敛现象,提出了一种基于锦标赛选择的查询函数构造方法,能够有效避免早熟收敛现象的发生,提高算法的稳定性。

### 3.1　分类器的选择与样本不确定性的评估

在主动学习算法中,查询函数建立在分类器模型之上,因此查询函数是分类器相关的。对于不同类型的分类器,查询函数的具体设计方法也有所不同,确定分类器类型是查询函数设计的首要条件。在分类器方面,考虑到支持向量机是目前处理分类问题的最佳工具之一,它以能够处理非线性、小样本、高维度和局部最小点等实际问题而被广泛使用[3,9],因此本文采用 SVM 分类器作为主动学习算法的分类器。SVM 建立在结构风险最小化原则基础上。对于训练样本为 $\{(x_i,y_i)\}_{i=1}^{l}\subset R^n\times\{1,-1\}$ 的二分类问题,根据统计学理论,可建立如下标准的线性 SVM 模型 $A$:

$$\begin{cases}\min\ \dfrac{1}{2}(\omega^{\mathrm T}\omega)+C\sum\limits_{i=1}^{l}\xi_i\\ \text{s.t.}\quad y_i((\omega^{\mathrm T}x_i)+b)\geqslant 1-\xi_i;i=1,2,\cdots,l\\ \xi_i\geqslant 0;i=1,2,\cdots,l\end{cases} \tag{2}$$

其中:$C>0$ 为正则化参数,$\xi_i(i=1,2,\cdots,l)$ 为松弛变量,$\omega\in R^n$ 为分类超平面的法向量,$b\in R$ 为阈值。利用优化理论中的 KKT 条件和对偶理论,可得对偶优化模型 $A'$:

$$\begin{cases}\max\ \sum\limits_{i=1}^{l}\alpha_i-\dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i{}^{\mathrm T}x_j)\\ \text{s.t.}\quad \sum\limits_{i=1}^{l}y_i\alpha_i=0\\ 0\leqslant\alpha_i\leqslant C;i=1,2,\cdots,l\end{cases} \tag{3}$$

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

$$\omega^*=\sum_{i=1}^{l}\alpha_i{}^*y_ix_i \tag{4}$$

根据 KKT 互补条件,最优解必满足

$$\alpha_i(y_i(\omega^{\mathrm T}x_i+b)-1+\xi_i)=0;i=1,2,\cdots,l \tag{5}$$

$$(C-\alpha_i)\xi_i=0;i=1,2,\cdots,l \tag{6}$$

由式(5)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 $\omega^*$ 起作用,从而决定分类结果。

UBS 查询策略是目前最可行且使用最广泛的,本文的查询函数设计也基于 UBS 策略。该策略认为分类器越不确定的样本其质量越高,因此对样本的不确定性进行评估是基于 UBS 策略查询函数设计的基础,结合 SVM 分类器,研究者们通常使用样本到 SVM 分类超平面的代数和距离计算样本的不确定性。样本 $x_i$ 到分类超平面的代数和距离计算方法如下:

$$d=\frac{g(x_i)}{\|w\|} \tag{7}$$

其中:$g(x)=0$ 为分类超平面,$w$ 为法向量,当 $d$ 值越小时,样本的不确定性越大。

### 3.2　早熟收敛现象

早熟收敛是进化算法中的常见现象,它是指进化过程中出现的群体多样性减少,个体分布在一个或若干个点附近,导致搜索停滞不前。在主动学习方法中,同样存在早熟收敛现象,主动学习方法与进化算法有相似之处,其本质都是搜索问题,主动学习方法依照查询函数提供的启发式规则在样本空间中搜索样本点,并根据这些样本点构造分类模型。其循环查询过程可以描述为

$$f_n(X_n)\xrightarrow{x_{n+1}}f_{n+1}(X_{n+1}) \tag{8}$$

其中:$X_n$ 是第 $n$ 轮查询后获得的训练样本集,$f_n(X_n)$ 是根据样本集 $X_n$ 构造的分类模型,$x_{n+1}$ 是依据分类模型 $f_n(X_n)$ 以及查询函数中的启发式规则筛选的样本。在每轮循环中,算法根据分类模型 $f_n(X_n)$ 及查询函数在样本池中查找样本 $x_{n+1}$,将 $x_{n+1}$ 标注后加入训练样本集,形成新的训练样本集 $X_{n+1}$,然后根据 $X_{n+1}$ 构建新的分类模型 $f_{n+1}(X_{n+1})$。可以看出,待选择样本由 $X_n$ 以及查询策略确定,由于样本池中的样本在特征空间中分布有可能极度不均,而 $f_n(X_n)$ 则是根据相对少量的样本 $X_n$ 建立的分类器,依托于分类模型 $f_n(X_n)$ 的查询函数所获得的样本 $x_{n+1}$ 可能仅携带极少的新增分类信息,以致分类模型 $f_{n+1}$

$(X_{n+1})$与$f_n(X_n)$近似,如图 1 所示。如果存在大量的样本点集中在$x_{n+1}$附近,由于$f_{n+1}(X_{n+1})$与$f_n(X_n)$之间差异微小,在下轮依托分类模型$f_{n+1}(X_{n+1})$进行的查询中获得的样本点仍会落在$X_{n+1}$附近,进一步导致分类模型变化微小,循环往复下去,会导致搜索的样本点长期停滞在局部范围,发生早熟收敛现象。早熟收敛现象发生时,分类器在每次查询中都未能获得更多的改进信息,因此分类准确率会维持在很小的范围内波动,并且远低于正常水平。在实验中,首先使用 5 折交叉验证方法对最常用的 SIMPLE[15] 算法进行评估,发现该现象对算法稳定性造成较大影响。

图 2 是 SIMPLE 算法最坏情况下的分类准确率与五次评估的平均分类准确率比较。图中横轴表示选取样本的数目,纵轴表示分类准确率,曲线表示选择训练样本数目对应的分类准确率,曲线附近的悬垂线段则是准确率的标准误差。可以看出,随着选择样本数目的增加,评估中最坏情况下的准确率一直维持在 70% 附近,且与五次评估的平均准确率及标准误差覆盖的范围存在较大距离,出现了显著的早熟收敛现象。除去最坏的一次评估,SIMPLE 算法最后获得的分类准确率都在 96.8% 以上。

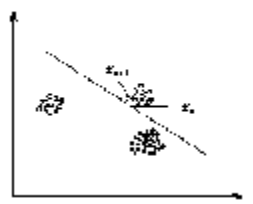

图 1　早熟收敛现象示意图

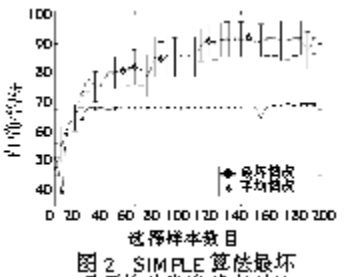


图 2　SIMPLE 算法最坏及平均分类准确率对比

### 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:

$$f_{\text{error}} = \frac{\text{MSE}(f^*(x,D))}{f(x,D)} \tag{9}$$

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 $\varepsilon$,当该值小于 $\varepsilon$ 时,终止样本选择循环。

### 3.4　基于锦标赛选择的主动学习算法

避免早熟收敛现象的主要方法是增加选择的多样性,而增加选择多样性则有可能对算法的收敛速度造成负面影响。为了适度增加样本选择的多样性,本文采用锦标赛选择方法改善查询函数的设计。锦标赛选择是进化算法中常用的搜索策略,其思想是在种群中取出一定规模的个体进行比较,选择其中最优的进入下一轮。采用该方法选择样本不仅有利于增加选择的随机性,还可以大量减少样本不确定性计算与比较的次数,降低主动学习方法的计算开销。根据上面构造的选择函数以及终止策略,本文给出了针对该算法的主动学习方法伪代码,如下所示:

基于锦标赛选择与 SVM 的主动学习方法如下:

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

```
将训练集初始化为 L 并将 Q_min 置 1;
repeat
{
    for each sample j in U do
    {
      if(! select(j)) continue;
      选取样本 j 并计算其 d 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)
```

假定有一个少量样本已标注好的训练集 $L$ 和一个未标注的样本池 $U$。当算法进入循环阶段,在每一轮中,首先按顺序选取样本池 $U$ 中的每一个样本,并以一定的概率丢弃该样本,以控制锦标赛规模,如果该样本未被丢弃,则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已标记好的训练集 $L$ 中再进行训练,直到满足终止条件均方预测误差率$f_{\text{error}}$小于阈值 $\varepsilon$ 为止。最后得到的训练集 $l$ 即为选择与精简后的训练集。

## 4　实验及分析

为了评估所提出流分类方法的有效性,本文进行了大量的实验。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:

a)检测率(true positive rate,TP)。对待检测类别,被正确检测出为该类别的样本数量与测试集这一类样本的总数量的比值。

b)误报率(precision,false positive rate,FP)。错误判为待检测类别的样本数量与测试集中非待检测样本数量的比值。

### 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[2]。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[2]中有对这些特征更详细的描述。

表 1　流特征示例

| Feature |
|---|
| 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[2]。

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,从中抽取 1 676 条 P2P 类别样本以及 6 006 条其他类别样本构成实验数据。同时,将其他类别样本归作一类,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 | 6 006 |
| P2P | 1 676 |

## 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 FCBF(fast correlation-based filter)算法对独立于实验数据的训练数据进行特征选择。表 4 中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的公平性,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 的参数为:类型为 C-SVM,核函数为 RBF(radial basic function),gamma 取值 0.005,$C$ 取值 2048;KNN 算法的近邻个数为 3。不难看出,相对于其他算法,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 算法在性能上大幅提升,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 | 96.5 | 1.6 | 73.16 | 22.78 |
| 特征选择后数据集 | 96.4 | 1.3 | 4.11 | 4.83 |

## 4.3　采用主动学习方法的实验结果

在对于主动学习方法的评价方面,一般采用两种方法评价效果:a)对于某个定量的训练集采用主动学习方法相对于被动学习方法所提升的性能提升程度;b)为了达到某个性能指标相对于被动学习方法所需要的训练样本集的精简程度。为了验证基于锦标赛选择的主动学习方法的有效性,本文在相同数量样本的情况下,关注了采用主动学习方法与 SIMPLE 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 组作为训练样本,剩下 1 组作为测试,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 3 显示了实验中分类准确率的均值与标准差变化。结果很显著:随着选择样本数量的增加,采用基于锦标赛选择的主动学习方法后,准确率迅速提升,并且分类准确率的标准差显著收敛,而使用 SIMPLE 算法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提升算法稳定性,进而减少构造 P2P 检测模型所需训练样本的规模。

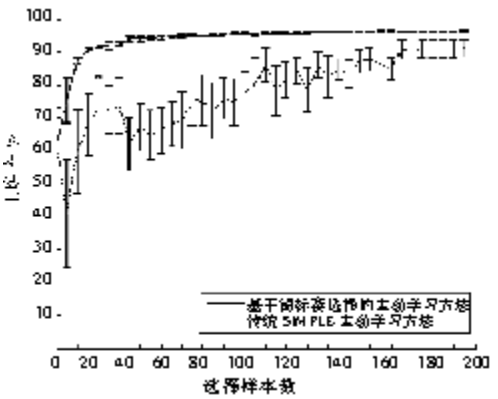


图 3　主动学习方法与随机选择方法的实验对比

## 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:在对比实验中,SVM 相对于其他几种常用算法,具有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;采用基于锦标赛选择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,并且算法的耗时更少。

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,减少训练样本标注的工作量。

## 5　结束语

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,但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,在减少建模成本,降低计算开销方面具有显著优势。

本文所提出的方法在实践应用中还需根据实际情况作进一步改进,以提高性能。如何将所述的算法应用于实际网络环境,并依据实际情况加以优化以及将该方法试用于其他类别流的识别,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[C]//Proc of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation. 2006:179-188.

[2] ESTEA A, GRINGOLIA F, SALGARELLI L. Support vector machines for TCP traffic classification [J]. Computer Networks, 2009, 53(14):2476-2490.

[3] 邓河, 阳爱民, 刘永定. 一种基于 SVM 的 P2P 网络流量分类方法[J]. 计算机工程与应用, 2008, 44(14):122-126.

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

[5] FRANK J. Machine learning and intrusion detection: current and future directions [C]//Proc of the the 17th Computer Security Conference. 1994.

[6] DUNNIGAN T, OSTROUCHOV G. Flow characterization for intrusion detection[R]. [S. l.]: Oak Ridge National Laboratory, 2000.

[7] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[C]//Proc of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004:135-148.

[8] 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.

[9] WILLIAMS N, ZANDER S, ARMITAGE G. A preliminary comparison of five machine learning algorithms for practical IP traffic flow classification[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(5):5-16.

[10] PARK B, HONG J W, WON Y J. Toward fine-grained traffic classification [J]. Communications Magazine, 2011, 49(7):104-111.

[11] SIMON H A, LEA G. Problem solving and rule reduction, a unified view[J]. Knowledge and Cognition, 1974, 15(2):63-73.

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

[13] BALDRIDGE J, OSBORNE M. Active learning for HPSG parse selection[C]//Proc of the 7th Conference on Natural Language Learning. 2003:100-108.

[14] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(11):45-66.


(上接第 713 页)

## 5　结束语

本文研究的主要目的是将经典的控制论应用于网络控制中的主动队列管理算法研究中,通过在网络参数与可变裕度之间建立的关系,来适应网络实时动态变化,同时引入灰色预测来补偿由于可变裕度 PID 调节所引起的网络时滞性。从仿真的结果来看,灰色预测的可变裕度 PID 网络自适应 AQM 算法改进了 AQM 算法的动态适应能力,同时补偿了网络的时滞性问题。

**参考文献:**


[1] BRADEN B, CLARK D, CROWCROFT J, *et al*. RFC 2309, Recommendations on queue management and congestion avoidance in the Internet[S]. 1998.

[2] FLOYD S, JACOBSON V. Random early detection gateways for congestion avoidance[J]. IEEE/ACM Trans on Networking, 1993, 1(4):397-413.

[3] FENG W, KANDLUR D, SAHA D, *et al*. A self-configuration RED gateway[C]//Proc of the INFOCOM'99. New York: IEEE Computer Society, 1999:1320-1328.

[4] OTT T J, LAKSHMAN T V, WONG L H. SRED: stabilized RED [C]//Proc of the INFOCOM'99. New York: IEEE Computer Society, 1999:1346-1355.

[5] HOLLOT C V, MISRA V, TOWSLEY D, *et al*. A control theoretic analysis of RED[C]//Proc of IEEE/INFOCOM. 2001:1235-1339.

[6] 任丰原,王福豹,任勇,等. 主动队列管理中 PID 控制器[J]. 电子与信息学报, 2003, 25(1):94-99.

[7] 任丰原,林闯,王福豹,等. ATM 网络拥塞控制中 PID 控制器的设计[J]. 软件学报, 2002, 25(10):1024-1029.

[8] 王军平,王安,敬忠良,等. Fuzzy-Gray 预测控制算法及应用[J]. 系统工程理论与实践, 2002(8):132-135.

[9] 孙雁飞,张顺颐,周雷. 基于模糊免疫 PID 的时滞网络自适应主动队列管理[J]. 通信学报, 2005, 26(8):36-42.

[10] WU Wei, REN Yong, SHAN Xiu-ming. A self-configuring PI controller for active queue management[C]//Proc of the 7th Asia-Pacific Communication Conference(APCC'2001). 2001.

[11] FUNG H K, WANG Q G, LEE T H. PI tuning in terms of gain and phase margins[J]. Automatica, 1998, 34(9):1145-1149.

[12] 汪浩,严伟,黄明和,等. 稳定裕度与网络特征参数无关的 AQM 算法[J]. 计算机工程, 2010, 36(4):12-15.