ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.20, No.10, October 2009, pp.2692−2704
doi: 10.3724/SP.J.1001.2009.03444

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 基于 C4.5 决策树的流量分类方法*

徐 鹏[1,2+], 林 森[1,2]

[1](中国科学院 软件研究所,北京 100190)

[2](中国科学院 研究生院,北京 100049)

## Internet Traffic Classification Using C4.5 Decision Tree

XU Peng[1,2+], LIN Sen[1,2]

[1](Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: xupeng@iscas.ac.cn

**Xu P, Lin S. Internet traffic classification using C4.5 decision tree. *Journal of Software*, 2009,20(10): 2692−2704.** http://www.jos.org.cn/1000-9825/3444.htm

**Abstract**: In recent years, Internet traffic classification using machine learning has become a new direction in network measurement. Being simple and efficient Naïve Bayes and its improved methods have been widely used in this area. But these methods depend too much on probability distribution of sample spacing, so they have connatural instability. To handle this problem, a new method based on C4.5 decision tree is proposed in this paper. This method builds a classification model using information entropy in training data and classifies flows just by a simple search of the decision tree. The theoretical analysis and experimental results show that there are obvious advantages in classification stability when C4.5 decision tree method is used to classify Internet traffic.

**Key words**: traffic classification; network measurement; decision tree; flow; statistical attribute

**摘 要**: 近年来,利用机器学习方法处理流量分类问题成为网络测量领域一个新兴的研究方向.在现有研究中,朴素贝叶斯方法及其改进算法以其实现简单、分类高效的特点而被广泛应用.但此类方法过分依赖于样本在样本空间的分布,具有潜在的不稳定性.为此,引入 C4.5 决策树方法来处理流量分类问题.该方法利用训练数据集中的信息熵来构建分类模型,并通过对分类模型的简单查找来完成未知网络流样本的分类.理论分析和实验结果都表明,利用 C4.5 决策树来处理流量分类问题在分类稳定性上均具有明显的优势.

**关键词**: 流量分类;网络测量;决策树;网络流;统计属性

**中图法分类号**: TP181 **文献标识码**: A

随着互联网用户规模的日益增大,互联网的拥塞状况也日益加剧.为了解决这一问题,网络研究人员提出了容量规划、流量调度等一系列策略来提高网络的运营效率.然而,无论是根据用户需求对网络资源进行 QoS 调度,还是根据网络应用的发展趋势对现有网络进行扩容改造,都必须对网络流量中各种应用进行准确的分类与识别.此外,在网络安全、流量计费、应用趋势分析等研究领域,准确的流量分类也具有极其重要的意义.

* Supported by the National Basic Research Program of China under Grant No.2007CB307100 (国家重点基础研究发展计划(973))
Received 2007-10-23; Revised 2008-03-27; Accepted 2008-08-07

为了适应 Internet 流量数据庞大、应用属性动态变化的特点,利用机器学习方法处理流量分类问题成为当前网络测量领域内一个新兴的研究热点.在使用机器学习方法处理流量分类问题时,研究的对象是一组具有相同 5 元组(源 IP、目的 IP、源端口、目的端口、传输层协议)取值的分组序列,即网络流(flow).研究人员通过提取网络流的统计属性,将网络流抽象为由一组统计属性值构成的属性向量,实现由流量分类向机器学习问题的转化.因此,从机器学习的角度来看,流量分类问题可以抽象为:在已知网络流类型集合 $T=\{T_1,T_2,...,T_k\}$ 和类型已知的网络流集合 $X=\{X_1,...,X_n\}$ 的情况下,如何利用机器学习方法通过对该网络流集合的"学习",来构建流量分类模型 $f:X\to T$,并以此模型对类型未知的网络流进行分类.

利用机器学习方法处理流量分类问题的核心工作主要包括两个方面:(1) 选择适当的网络流属性集合构建属性向量;(2) 选择适当的机器学习算法构建分类模型.从目前的研究成果来看,选择改进的朴素贝叶斯方法 NB(naïve Bayes)[1]进行流量识别,不仅分类准确率较高,而且实现简单,处理高效,整体性能相对较好.朴素贝叶斯方法是一种基于概率的参数估计方法,该方法应用的前提在于:待分类样本的先验概率在样本空间保持稳定.然而,在真实的网络环境中,网络流样本的分布是动态变化的,应用朴素贝叶斯方法的前提条件无法满足.为此,本文首先通过理论分析指明利用朴素贝叶斯及其改进算法处理流量分类问题的不足,然后提出了一种基于 C4.5 决策树(decision tree)的流量分类方法.该方法根据训练数据集中的信息熵(information entropy)来构建分类模型.在使用该模型进行流量分类时仅需根据网络流属性值自顶向下进行比较,最终找到标明网络流类型的叶子节点.实验结果表明,利用 C4.5 决策树方法来处理流量分类问题,在分类稳定性**和数据处理效率上具有明显的优势.

本文第 1 节综述网络流量分类研究的现状,简要介绍各种主要的流量分类方法.第 2 节分析朴素贝叶斯方法的不足,引入基于 C4.5 决策树的分类方法.第 3 节简要说明实验环境,给出实验数据集和分析平台的说明.第 4 节简要介绍评估策略.第 5 节对实验结果进行分析.第 6 节总结全文并展望未来的工作.

## 1 研究现状

近年来,随着互联网技术的不断发展,网络应用的快速增长和变化给流量分类带来了一系列挑战.最初的流量分类方法是根据互联网地址指派机构 IANA(Internet assigned numbers authority)规定的端口映射表,将特定端口的网络流量划分到相应的网络应用.然而,随着 P2P 和被动 FTP 等新型网络应用的日益流行,大量的随机端口被用于数据传输,从而导致这种基于端口的流量分类方法被迅速淘汰[2].

2005 年,剑桥大学的 Moore 等人[3]提出了基于特征字段的流量分类方法,主要是通过分析数据分组的应用层负载,检测不同应用的特征字段来划分网络流量.该方法能够有效地识别现有的互联网应用,已为大多数商用流量识别系统所采用.然而该方法不仅依赖于数据分组的应用层负载,而且依赖于应用的特征字段.分析完整的应用层负载不仅计算开销较大,而且还可能带来不必要的用户隐私权纠纷.此外,对于使用负载加密技术或者特征字段保密的网络应用,该方法通常也是无能为力的.

针对负载加密等问题,加州大学河滨分校的 Karagiannis 等人[4]又提出了一种基于传输层行为的流量分类方法 BLINC(blind classification),该方法利用不同网络应用在传输层连接模式的差异来划分网络流量,不依赖于负载的内容,具有良好的可扩展性.然而该方法利用了网络应用的行为属性,不仅容易随着网络应用自身的改进而逐步失效,而且也会因为网络环境的不同而导致分类性能出现显著的变化.

2004 年,澳大利亚阿德莱德大学的 Roughan 等人[5]率先引入了 *K*-近邻(*K*-nearest neighbor)和线性判别式分析(linear discriminant analysis,简称 LDA)这两种最简单的机器学习方法来处理流量分类问题.*K*-NN 方法在处理测试样本时需要逐个地计算测试样本和训练样本之间的相似度,不仅具有较大的计算开销,而且分类性能极易受到噪声数据的干扰.LDA 方法通常需要复杂的预处理过程以将多维信息映射到一维空间,在处理流量分类问题时,会引入大量的额外开销.

** 在本文中,分类稳定性用以描述分类准确率对样本先验概率分布的依赖程度,依赖程度越低,稳定性越好.

2005年,牛津大学的Zuev和剑桥大学的Moore等人[6]引入了基于概率模型的朴素贝叶斯方法.该方法要求参与分类的各项属性条件独立而且遵循高斯分布,然而在流量分类问题中,原始的网络流属性集合很难满足上述条件,因此,该方法的整体准确率只有65%左右.为了克服条件独立假设和高斯分布假设带来的消极影响,Moore等人[1]采用基于关联的快速过滤机制FCBF(fast correlation-based filter)算法和核估计(kernel estimation)技术这两种策略对原始的朴素贝叶斯方法进行了改进.从他们的实验结果来看,采用其中任意一项改进策略都能将分类准确率提高到90%以上,采用两种改进策略后的分类准确率达到了95%左右.

为了进一步提高流量分类模型的实时性,2006年澳大利亚斯温伯恩大学的Nguyen等人[7]提出了多子流模型(multiple sub-flows model).该方法首先将网络流根据协议通信的不同阶段划分为若干条子流,然后分别为每条子流构造属性向量,并以此作为基本单元构造训练数据集.该方法通过子流属性提取摆脱了对网络流进行处理时必须等待网络流结束的限制,极大地提高了分类模型的实时性.但子流持续时间相对较短,其属性特征容易受到网络运行状态的影响而发生变化,因此,该方法的健壮性还有待于进一步验证.

2006年,加拿大卡尔加里大学的Erman等人[8,9]引入了聚类方法来处理流量分类问题,此类方法在聚类过程中无须使用训练样本的类型,因而能够识别部分类型尚未定义的新型网络流量,然而他们在聚类结束后必须进行手工标记以实现网络流量的分类,分类效率偏低.

此外,2006年,法国LIP6实验室的Bernaille等人[10]还提出利用TCP连接中前5个数据分组的长度序列来代表相应的网络流.他们采用了$K$均值、高斯混合模型(Gaussian mixture model)和谱聚类(spectral clustering)这3种聚类方法来处理这些由分组长度序列构成属性向量.他们的实验结果表明,这些聚类算法不仅能够以超过90%的准确率有效地分类知名应用的网络流量,而且能够将部分类型未定义的新型网络流标记出来.由于Bernaille等人采用网络流属性依赖于数据分组的到达顺序,而在实际网络环境中数据分组由于路由动态性、网络路径的拥塞等原因,通常无法保证顺序到达[11],因此,Bernaille等人的方法,其稳定性和实用性无法保证.

国内方面,2006年,国防科学技术大学的王锐等人[12]率先将支持向量机SVM(support vector machine)方法应用到P2P流的识别领域.他们利用网络连接数相关的统计属性将网络流简单划分为P2P流和非P2P流,然而他们所用的统计属性依赖于应用的连接模式,因此,该方法与基于传输层行为的流量识别方法相似,分类结果的稳定性极易受到网络环境的影响.同年北京航空航空大学的Li等人[13]提出了一种基于粗糙集理论和遗传算法的流量分类方法.但这一算法受条件所限,仅在包含2 254条流的数据集上进行了小规模验证,没有给出与现有方法的进一步比较与分析.

## 2 流量分类方法

### 2.1 朴素贝叶斯方法的不足

朴素贝叶斯方法是一种利用贝叶斯定理对属性集合和类型变量之间概率关系进行建模的方法.因此,在使用NB方法处理流分类问题[5]时,对于任意网络流$X$,其属于类型$Y$的概率为

$$P(Y \mid X)=\frac{P(X \mid Y)P(Y)}{P(X)} \tag{1}$$

其中,先验概率$P(Y)$表示在网络流集合中流量类型为$Y$的网络流所占的比例;类条件(class-conditional)概率$P(X|Y)$表示在类型为$Y$的网络流中出现网络流$X$的概率;分母$P(X)$表示网络流$X$出现的概率.

由于网络流$X$被抽象为属性向量$(A_1,...,A_m)^T$,因此,式(1)可以扩展为

$$P(Y \mid A_1,...,A_m)=\frac{P(A_1,...,A_m \mid Y)P(Y)}{P(X)} \tag{2}$$

在使用NB方法所得到的模型进行流量分类时,需要为每一条待测网络流$X$分别计算其取不同$Y$值的后验概率$P(Y|X)$,从中选择后验概率最大的类型作为分类模型的判定结果.此时,分母$P(X)$总是相等,因此在比较时可以忽略不计.为了进一步简化式(2)的计算,原始的NB方法提出了条件独立假设和高斯分布假设,即假定属性$A_1,...,A_m$相互独立且遵循高斯分布,通过计算训练数据集上各类属性的统计值来获取高斯分布的各项参数值.

然而在实际的流量分类问题中,条件独立假设和高斯分布假设通常难以满足,以分组首部长度和分组长度的统计值为例,整个分组的长度等于分组首部长度加上分组负载长度,两者之间具有明显的加法关系,无法实现相互独立.此外,由于部分网络流统计属性符合多重模态分布(multi-modal distribution)[1],直接使用高斯分布假设将无法有效拟合网络流属性的统计分布.因此,在直接使用原始的 NB 方法来处理流量分类问题时,分类准确率只有 65%左右[6],必须对原始的 NB 方法进行改进.

针对原始的网络流属性集合难以满足条件独立假设和高斯分布假设的情况,Moore 等人[1]采用 FCBF 算法来选择符合条件独立假设的属性子集,并采用了基于核估计技术的朴素贝叶斯 NBK(naïve Bayes using kernel density estimation)方法来拟合网络流属性的多重模态分布.从他们的实验结果来看,使用上述两种改进策略后,可以将流量分类模型的分类准确率从 65%提高到 95%左右.

Moore 等人的改进策略虽然有效提高了朴素贝叶斯方法处理流量分类问题时的整体分类准确率,然而,从朴素贝叶斯方法的基本原理式(1)来看,在比较同一网络流属于不同类型的后验概率 $P(Y|X)$时,问题的关键不仅在于类条件概率 $P(X|Y)$,还包括先验概率 $P(Y)$.然而,无论是原始的朴素贝叶斯方法还是 NBK 方法都只关注了类条件概率 $P(X|Y)$的计算,且都只是用训练数据集中各类样本的比例来简单预估各类样本的先验概率 $P(Y)$.这种简单的预估实际上是使用特定条件下的采样值来预估实际环境中的动态变化值,这必然会导致分类准确性的明显变化.

### 2.2 C4.5决策树方法

决策树方法是以实例为基础的归纳学习算法,它从一个无次序、无规则的实例集合中归纳出一组采用树形结构表示的分类规则.自 20 世纪 60 年代以来,决策树方法在分类、预测、规则提取等领域得到广泛应用.

利用决策树处理分类问题通常分为两步:第 1 步是通过训练集合的学习,形成决策树分类模型;第 2 步是利用生成的决策树模型对类型未知的样本进行分类.在使用决策树模型对类型未知样本进行分类时,从根节点开始逐步对该样本的属性进行测试,并沿着相应的分支向下行走,直至到达某个叶节点,此时叶节点所代表的类型即为该样本的类型[14].由此可见,利用决策树方法进行分类的关键是根据训练集合构建决策树分类模型.

本文采用了目前广泛应用于金融、医疗领域的 C4.5 决策树算法,该算法根据信息增益率(information gain ratio)来选择测试属性.以实际的流量分类问题为例:已知网络流样本集 $S=\{X_1,X_2,...,X_n\}$,其中每个样本可以由一个包含 $m$ 项网络流属性的属性向量 $(A_1,...,A_m)^T$ 来表示,假设类别属性 $A_m$ 具有 $k$ 个不同取值,那么根据 $A_m$ 的不同取值可以将样本集 $S$ 划分为 $C_1,C_2,...,C_k$ 共 $k$ 个子集,由此可以得出样本集 $S$ 对分类的平均信息量:

$$H(S)=-\sum_{p=1}^{k}P(C_p)\log_2 P(C_p) \tag{3}$$

其中,$P(C_p)=|C_p|/|S|$ $(1\le p\le k)$.决策树的构建过程就是使划分后不确定性逐渐减小的过程.以任意的离散属性 $A_i(1\le i\le m-1)$ 为例,假设 $A_i$ 存在 $t$ 个不同的取值 $a_q$ $(1\le q\le t)$,那么根据 $A_i$ 的取值,不仅可以将 $S$ 划分为 $S_1,S_2,...,S_t$ 共 $t$ 个子集,还可以将 $C_1,C_2,...,C_k$ 这 $k$ 个子集进一步划分为 $k\times t$ 个子集,每个子集 $C_{pq}$ 表示在 $A_i=a_q$ 的条件下属于第 $p$ 类的样本集合,其中,$1\le p\le k,1\le q\le t$.由此,选择离散的非类别属性 $A_i$ 进行划分后,样本集 $S$ 对分类的平均信息量为

$$H(S/A_i)=-\sum_{q=1}^{t}P(C_q)\left[-\sum_{p=1}^{k}P(C_{pq})\log_2 P(C_{pq})\right] \tag{4}$$

其中,$P(C_q)=\sum_{p=1}^{k}|C_{pq}|/|S|$,$P(C_{pq})=|C_{pq}|/|S|$,那么利用 $A_i$ 对 $S$ 进行划分的信息增益量(information gain)$f_G(S,A_i)$则等于使用 $A_i$ 对 $S$ 进行划分前后,不确定性下降的程度,即:

$$f_G(S,A_i)=H(S)-H(S/A_i) \tag{5}$$

由于使用属性 $A_i$ 对 $S$ 进行划分的信息增益率等于信息增益量与分割信息量(split information)之比,那么可以得到:

$$f_{GR}(S,A_i)=\frac{f_G(S,A_i)}{f_{sp}(S,A_i)} \tag{6}$$

其中,分割信息量 $f_{sp}(S,A_i)=-\sum_{l=1}^{t}\left(|S_l|/|S|\right)\log_2\left(|S_l|/|S|\right)$.对于非离散的网络流属性,C4.5 决策树算法采用离散化其取值空间的策略[14],将其转化为离散属性进行计算.通过选择具有最大信息增益率的属性作为测试属性,C4.5 决策树方法自上而下地完成决策树的建树过程.为了去除噪声点和孤立点引起的分支异常,C4.5 决策树方法利用训练数据集中剩余的样本,对生成的初始决策树进行了剪枝,进而得到最终的 C4.5 决策树.

通过对 C4.5 决策树原理的分析可以看出,与 NB 和 NBK 方法相比,利用 C4.5 决策树方法处理流量分类问题具有以下几个优势:

(1) C4.5 决策树方法在模型构建和样本预测过程中都不依赖于网络流样本的分布,因此,该方法能够有效地避免网络流样本分布变化所带来的影响,具有良好的分类稳定性.

(2) 在利用 C4.5 决策树模型对待分类样本进行处理时,仅需要根据网络流样本属性值自顶向下地进行比较,找到相应的叶节点即可,处理相对简单,具有更高的数据处理效率.

## 3 实验环境

### 3.1 实验数据

为了便于对比和分析,本文采用了两个实验数据集,第 1 个是 Moore 等人在文献[1]中所用的实验数据集.为了描述方便,称其为 Moore_Set.第 2 个是由本文采集网络流量记录(trace)提取而来,简称为 CAS_Set.

Moore_Set 数据集采自 3 个生物学研究所共享的网络出口.这 3 个研究所总共拥有大约 1 000 名研究人员、管理人员和技术人员,通过共用 1 条千兆全双工以太网链路连接到互联网.由于原始的流量记录集合中包含了 2003 年 8 月 20 日 0 时开始 24 小时内流经该网络出口所有双向网络流量,数据量过于庞大.为此,Moore 等人采用抽样的方法,分别提取 10 个包含双向网络流量的流量子集.这 10 个流量子集的平均抽样时间大约是 1 680s 左右.为了将主要精力放在流量分类模型的构造上,Moore 等人在构造实验数据集时,只选用语义完整的 TCP 双向流作为网络流样本.在本文中 TCP 流语义完整的判定条件是:以完整的握手过程(SYN-ACK)开始,且以完整的握手过程(FIN-ACK)结尾的 TCP 双向流[1].

在 Moore_Set 数据集中共包含 377 526 个网络流样本,被分为 10 种类型.每种类型所包含的应用名称、每类网络流的数量和所占的比例见表 1.Moore_Set 中每条网络流样本都是从一条完整的 TCP 双向流抽象而来,包含 249 项属性,其中最后一项属性是目标属性,指明了该双向流的类型;而其中的第 1 项属性和第 2 项属性分别是该 TCP 流的源端口号和目的端口号.为了避免对应用端口信息的依赖,本文的所有实验都未使用这两个属性.剩余 246 种网络流属性的具体描述可查阅文献[15].

**Table 1** Statistics of Moore_Set

**表 1** Moore_Set 数据集的统计信息

| Type of flow | Application names | Num of flow | Percent (%) |
|---|---|---|---|
| WWW | Http,Https | 328 091 | 86.91 |
| MAIL | Imap, Pop2/3, Smtp | 28 567 | 7.567 |
| BULK | Ftp | 11 539 | 3.056 |
| DB | Postgres, Sqlnet, Oracle, Ingres | 2 648 | 0.701 |
| SERV | X11, Dns, Ident, Ldap, Ntp | 2 099 | 0.556 |
| P2P | Kazaa, Bittorrent, Gnutella | 2 094 | 0.555 |
| ATT | Internet Worm And Virus Attacks | 1 793 | 0.475 |
| MULT | Windows Media Player,Real | 1 152 | 0.305 |
| INT | Ssh, Klogin, Rlogin, Telnet | 110 | 0.029 |
| GAME | Half-Life | 8 | 0.002 |
| Total | 26 applications | 377 526 | 100 |

CAS_Set 数据集采自中国科学院某研究所的网络出口.该研究所拥有大约 1 000 多名员工与学生,通过共享

一个上/下行速率为 20Mbps 的网络接口接入 Internet.本文捕获了 2006 年 12 月 12 日下午 14:00~14:59 之间经过该网络接口的所有网络流量.为了便于对比,本文参考 Moore 等人的工作仅对语义完整的 TCP 流进行分析.在本文中所用 TCP 流是双向网络流,以网络流第 1 个分组的转发方向作为该网络流的前向转发方向.

CAS_Set 数据集包含 684 551 条完整的双向 TCP 流,主要分为 7 种类型.每种类型所包含的应用名称、每类网络流的数量以及在 CAS_Set 中所占的比例见表 2.

**Table 2** Statistics of CAS_Set

**表 2** CAS_Set 数据集的统计信息

| Type of flow | Application names | Num of flow | Percent (%) |
|---|---|---|---|
| WEB | Http, Https | 401 338 | 58.63 |
| BT | BitTorrent | 106 975 | 15.63 |
| ED | eDonkey2000,eMule | 76 291 | 11.15 |
| BULK | FTP | 70 194 | 10.25 |
| PP | Pplive | 19 994 | 2.92 |
| MAIL | IMAP, POP2/3, SMTP | 9 454 | 1.38 |
| INT | SSH, Telnet | 305 | 0.05 |
| Total | 12 applications | 684 551 | 100.00 |

在 Moore_Set 的 249 项网络流属性中有 100 多项属性是通过傅里叶变换技术得来的.在实际的网络环境中,网络速度越来越高,待分类的网络流数目通常达到数十万乃至数百万条流,如果对每条网络流都进行傅里叶变换,则计算负载过于沉重.为此,我们从属性易于获取的角度出发,利用工具软件包 Netmate[16]将 CAS_Set 中的每条双向 TCP 流都抽象为仅包含 34 项网络流属性的属性向量,具体的属性说明见表 3.

**Table 3** Description of flow attributes

**表 3** 网络流属性描述

| No. | Abbreviation | Description in English |
|---|---|---|
| 1 | total_fpackets | Total number of packets in forward direction |
| 2 | total_fvolume | Total number of bytes in forward direction |
| 3 | total_bpackets | Total number of packets in backward direction |
| 4 | total_bvolume | Total number of bytes in backward direction |
| 5 | fpsh_cnt | Number of Push packet in forward direction |
| 6 | bpsh_cnt | Number of Push packet in backward direction |
| 7 | furg_cnt | Number of Urgent packet in forward direction |
| 8 | burg_cnt | Number of Urgent packet in backward direction |
| 9 | min_fpktl | Minimum forward packet length |
| 10 | mean_fpktl | Mean forward packet length |
| 11 | max_fpktl | Maximum forward packet length |
| 12 | std_fpktl | Standard deviation of forward packet length |
| 13 | min_bpktl | Minimum backward packet length |
| 14 | mean_bpktl | Mean backward packet length |
| 15 | max_bpktl | Maximum backward packet length |
| 16 | std_bpktl | Standard deviation of backward packet length |
| 17 | min_fiat | Minimum forward inter-arrival time |
| 18 | mean_fiat | Mean forward inter-arrival time |
| 19 | max_fiat | Maximum forward inter-arrival time |
| 20 | std_fiat | Standard deviation of forward inter-arrival time |
| 21 | min_biat | Minimum backward inter-arrival time |
| 22 | mean_biat | Mean backward inter-arrival time |
| 23 | max_biat | Maximum backward inter-arrival time |
| 24 | std_biat | Standard deviation of backward inter-arrival time |
| 25 | duration | Duration of the flow |
| 26 | min_active | Minimum of active time |
| 27 | mean_active | Mean of active time |
| 28 | max_active | Maximum of active time |
| 29 | std_active | Standard deviation of active time |
| 30 | min_idle | Minimum of idle time |
| 31 | mean_idle | Mean of idle time |
| 32 | max_idle | Maximum of idle time |
| 33 | std_idle | Standard deviation of idle time |
| 34 | flow_type | Type of flow |

在这 34 项属性中,最后一项称为类别属性,指明了该网络流所属的应用类型;剩余的 33 项网络流统计属性,主要分为以下 3 类:

(1) 分组数量相关属性.这是指与网络流中与分组数目相关的统计属性,主要有前向转发分组的总数(total number of packets in forward direction)和后向转发分组的总数(total number of packets in backward direction)等.不同网络应用在分组数量相关属性上可能存在较大差异,以 FTP 应用和 SSH 应用为例,前者通常用于块文件数据传输,而后者通常用于实时消息传送,因此,前者的前向(后向)转发分组数目将远远大于后者.

(2) 分组长度相关属性.这是指与网络流中分组长度相关的统计属性,主要包括前向转发分组的平均长度(mean forward packet length)和后向转发分组的平均长度(mean forward packet length)等.Erman 等人[9]的研究指出,P2P 流数据流与被动 FTP 数据流的本质区别就在于,前者是双向数据传输,而后者是单向数据传输.由此可见,利用前、后向转发分组平均长度的差异可以有效地区分被动 FTP 数据流和 P2P 数据流.

(3) 时间相关属性.这是指与网络流中时间相关的统计属性,例如流持续时间(flow duration)以及前向转发分组的平均到达间隔(mean forward inter-arrival time)等.时间相关属性是网络流属性中相对易于波动的一组属性,然而它们也在一定程度上能够有效地区分不同的网络应用.以网络流持续时间为例,以文件传输为主要目的 FTP 流,其持续时间就远大于以网页浏览为目的的 Web 流.

由于 CAS_Set 所使用的网络流属性相对较少,为了避免 FCBF 算法等属性过滤机制利用局部信息过滤网络流属性所带来的局部最优性问题,在使用 CAS_Set 实验中,没有采用任何属性过滤机制.此外,由于 CAS_Set 使用的网络流属性大多属于离散取值,因而在进行相关实验时,没有使用 NBK 方法,而使用离散的朴素贝叶斯(naïve Bayes using discretisation,简称 NBD)方法来改进朴素贝叶斯方法对网络流属性的拟合.

### 3.2 分析工具和平台

本文所使用的数据挖掘工具是 Weka-3.5.6[17].该工具是由新西兰怀卡托大学 Witten 教授等人开发的开源工作平台.该平台利用 Java 语言实现了决策树、朴素贝叶斯等多种机器学习方法.该软件工具包可以通过 Internet 直接从相关的网站上获取.本文所用实验平台为一台普通 PC 机,其 CPU 为 Intel Pentium-4 2.66G Hz,内存为 DDR-667 2G Bytes;运行 Windows XP 操作系统.

## 4 评估策略

在基于机器学习的流量分类研究中,模型评估是指在测试数据集上运行流量分类模型,并根据运行结果预估分类模型在未知数据集上处理流量分类问题的能力.以一个 $m$ 元流量分类问题为例:假设在测试集中存在 $N$ 条网络流样本,分别属于 $m$ 种网络应用类型,那么对应的混淆矩阵(confusion matrix)见表 4.

**Table 4** Confusion matrix for $m$ classification

**表 4** $m$ 元流量分类问题的混淆矩阵

| | | Prediction class | | | | |
|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | … | Class $m$ |
| Actual class | Class 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | … | $n_{1m}$ |
| | Class 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | … | $n_{2m}$ |
| | Class 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | … | $n_{3m}$ |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | Class $m$ | $n_{m1}$ | $n_{m2}$ | $n_{m3}$ | … | $n_{mm}$ |

上表中任意表项 $n_{ij}$(其中 $1\le i\le m$, $1\le j\le m$, $\sum_{i=1}^{m}\sum_{j=1}^{m}n_{ij}$)表示实际类型为 $i$ 而被分类模型预测为类 $j$ 的样本个数.由此可以对任意类型 $i$ 定义如下概念:

(1) 真正 TP(true positive):实际类型为 $i$ 的样本中被分类模型正确预测的样本数,$TP_i=n_{ii}$.

(2) 假负 FN(false negative):实际类型为 $i$ 的样本中被分类模型误判为其他类型的样本数, $FN_i=\sum_{j\neq i}n_{ij}$.

(3) 假正 FP(false positive):实际类型为非 $i$ 的样本中被分类模型误判为类型 $i$ 的样本数量, $FP_i=\sum_{j\neq i}n_{ji}$.

基于以上概念,下面给出衡量分类模型准确性的 3 个常用指标:类准确率(accuracy)、类可信度(trust)和整体准确率(overall accuracy)的形式化描述.

$$\text{类 } i \text{ 的准确率 } A_i=\frac{TP_i}{TP_i+FN_i} \tag{7}$$

$$\text{类 } i \text{ 的可信度 } T_i=\frac{TP_i}{TP_i+FP_i} \tag{8}$$

$$\text{整体准确率 } OA=\frac{\sum_{i=1}^{m}TP_i}{\sum_{i=1}^{m}(TP_i+FH_i)} \tag{9}$$

在上述 3 项评估指标中,分类模型的整体准确率应用最广,几乎为所有研究人员所采纳,它反映了分类模型正确预测样本数在预测总数中的比例.类 $i$ 的准确率度量在类 $i$ 的所有样本中被分类模型正确预测的样本所占的比例.类 $i$ 的准确率越高,表明分类模型将类型为 $i$ 的样本错判为其他类型的错误数越少.类 $i$ 的可信度表示在被分类模型判定为 $i$ 类的样本中实际为 $i$ 类的样本所占的比例.类 $i$ 的可信度越高,表明分类模型将其他类型的样本误判为 $i$ 类的错误数越少.单个类型的准确率和可信度反映了分类模型针对特定类型的分类能力,对于评估分类模型小类样本(即此类样本数量在整个样本集合相对较少)的处理能力具有重要的意义.

## 5 实验结果与分析

### 5.1 准确性指标

#### 5.1.1 Moore_Set 的结果

为了对比分析决策树和朴素贝叶斯方法分类的稳定性,首先将 Moore_Set 数据集均分为两个数据子集,分别是 Set1 和 Set2,在这两个数据子集中每类样本的比例与 Moore_Set 保持一致.再从 Set1 中分别抽取每类应用 0.1%(至少 1 个)的样本构成训练集.由于在 Moore_Set 包含的 249 项网络流属性中,存在众多的冗余属性和无关属性,这些属性的存在不仅会降低分类模型的准确率,而且会极大地加重分类模型的计算负载.因此,首先使用 FCBF 方法对训练数据集进行过滤,然后在完成过滤训练数据集上分别运行 NB,NBK,C4.5 这 3 种机器学习算法以获得相应的流量分类模型,并使用这些模型在测试数据集 Set2 上进行验证.随后,再将训练数据集的规模逐步提高到 1%,10%,50%的 Set1.重复上述实验 10 次,所得实验结果如图 1 所示.

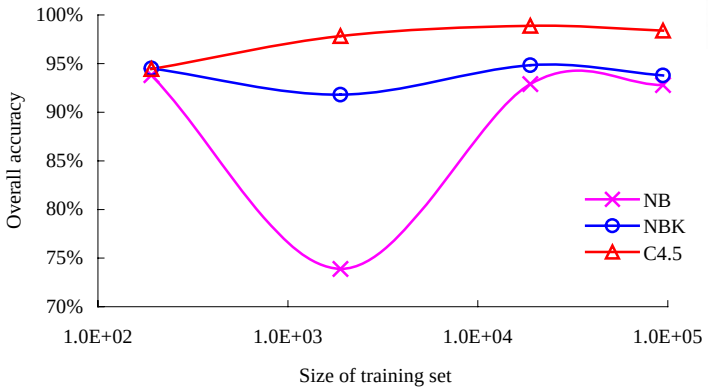


Fig.1 Overall accuracy with stratified sampling

图 1 分层抽样下的整体分类准确率

图 1 中 $X$ 轴采用对数坐标表示训练数据集中样本的总数,$Y$ 轴表示分类模型的整体分类准确率.3 条曲线分别表示训练样本集从 191 开始逐步增加到 1 889,18 876,94 382 时,NB,NBK 和 C4.5 算法在 10 次实验中的整体分类准确度的均值.

从图 1 中可以看出,NB 方法因直接使用高斯分布假设无法有效拟合网络流属性分布,致使分类结果明显较差.而 NBK 方法和 C4.5 方法在这种先验概率保持不变的情况下,虽然都能够保持较高的分类准确率,但分类模型的准确率却随着训练数据集的增大而出现轻微抖动.这主要是因为 FCBF 算法是根据训练数据集的局部信息对样本属性进行过滤,无法从全局角度对属性向量进行优化,必然会导致属性选择的局部最优性,从而引起分类准确率出现显著变化.由此可见,NBK 方法既需要使用属性过滤机制来满足条件独立假设,同时又会因为属性过滤机制的局部最优性而导致分类结果不够稳定.因此,如何优化网络流属性子集的选择,避免由于局部最优性而导致的分类结果不稳定,这将是一个需要进一步深入研究的问题.

图 1 虽然给出了分类模型整体分类准确率伴随训练数据集大小的变化的情况,但由本文第 4 节的讨论可知,分类模型准确性的相关指标还包括每类的准确率和可信度.为了进一步比较 3 种机器学习算法在分层抽样下的分类准确性,表 5 和表 6 分别给出了训练数据集大小为 94 382 时,各类网络应用的类准确率和类可信度.

**Table 5** Accuracy of all type with stratified sampling

**表 5** 分层抽样下各种方法的类准确率

| Method | WWW | MAIL | ATT | P2P | DB | BULK | MUL | SERV | INT | GAME |
|---|---|---|---|---|---|---|---|---|---|---|
| NB (%) | 98.21 | 87.73 | 0.00 | 2.31 | 1.05 | 22.96 | 11.57 | 8.83 | 1.62 | 0.00 |
| NBK (%) | 99.36 | 90.87 | 2.58 | 3.09 | 8.85 | 14.88 | 3.97 | 0.01 | 0.00 | 0.00 |
| C4.5 (%) | 99.70 | 96.77 | 68.52 | 41.07 | 92.15 | 90.17 | 72.01 | 66.90 | 40.73 | 0.00 |

**Table 6** Trust of all type with stratified sampling

**表 6** 分层抽样下各种方法的类可信度

| Method | WWW | MAIL | ATT | P2P | DB | BULK | MUL | SERV | INT | GAME |
|---|---|---|---|---|---|---|---|---|---|---|
| NB (%) | 95.22 | 82.04 | 0.00 | 6.90 | 20.43 | 66.06 | 13.36 | 1.29 | 5.59 | 0.00 |
| NBK (%) | 95.68 | 77.97 | 12.50 | 20.64 | 49.93 | 96.13 | 9.35 | 10.00 | 0.00 | 0.00 |
| C4.5 (%) | 98.82 | 98.19 | 97.08 | 79.53 | 94.90 | 92.45 | 91.37 | 66.94 | 71.32 | 0.00 |

从表 5 和表 6 中的实验结果可以看出,在比较小类样本的类准确率和类可信度时,C4.5 决策树方法明显优于 NB 和 NBK 方法.这主要是因为朴素贝叶斯及其改进算法都是依赖样本先验概率的方法,从式(1)明显可以看出,在其他条件相同的情况下,先验概率 $P(Y)$越大,后验概率 $P(Y|X)$也越大.利用朴素贝叶斯及其改进算法所得分类模型在处理分类问题时,将明显有利于大类样本(即此类样本数量在整个样本集合相对较多);而 C4.5 决策树方法不依赖于样本先验概率分布,则可以有效地避免这种情况.

为了进一步分析朴素贝叶斯和决策树两类方法对样本先验概率的依赖程度,本文从 Set1 的网络流中抽取每种应用的 100 条随机样本(由于 MUL,INT,GAME 这 3 种类型的样本数相对较少,因此在本实验中去除这 3 种类型的样本),合成一个各类样本均等的训练集.同样,事先采用 FCBF 算法对训练数据集进行过滤,再分别运行 NB,NBK 和 C4.5 这 3 种机器学习算法获得相应的流量分类模型,并以此模型在测试数据集 Set2 上进行验证.再依次将训练数据集中每类样本的个数提高到 300,500,700 和 897(由于在 Set1 中 ATT 类型的网络流样本只有 897 个).重复上述实验 10 次,所得实验结果如图 2 所示.

图 2 中使用 $X$ 轴代表训练集合中每类样本的数量,$Y$ 轴表示分类模型的准确率.3 条曲线分别表示训练数据集中各类样本数量从 100 开始逐步增加到 300,500,700,897 时,NB,NBK 和 C4.5 算法在 10 次实验中的整体分类准确率的均值.

从图 2 可以看出,在上述 3 种机器学习方法的分类结果中,只有 C4.5 决策树算法的整体分类准确率随着训练集合的逐步增大而保持相对稳定的增加,而 NB 和 NBK 方法的分类准确率不仅没有随着训练数据集的增大而提高,相反却随着训练数据集的增大而出现显著下降.这不仅是由于 FCBF 方法的局部最优性所带来的抖动,更主要的是因为训练数据集和测试数据集中各类样本的分布存在较大差异,而基于贝叶斯定理的 NB 和 NBK 方法都是假设先验概率保持不变.当这一假设条件无法满足时,NB 和 NBK 方法也就随之失效.

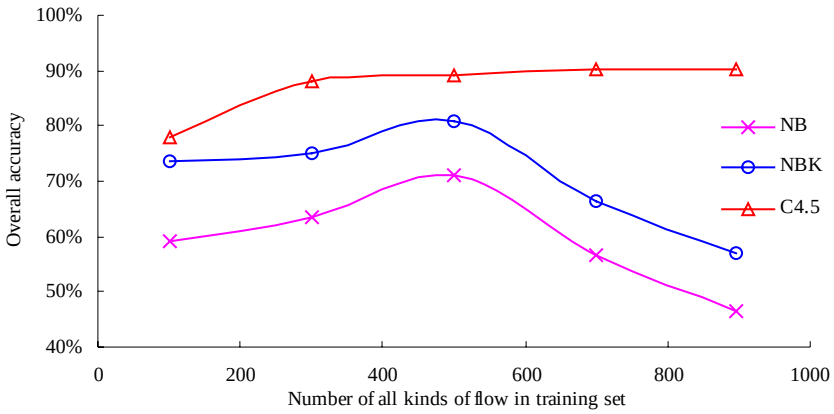


Fig.2 Overall accuracy with uniform sampling

图 2 均匀抽样下各种方法的平均分类准确率

为了进一步分析均匀抽样条件下 3 种机器学习方法的准确性,我们分别在表 7 和表 8 中给出了训练数据集中各类样本数为 897 时,各类网络应用的类准确率和类可信度.

**Table 7** Accuracy of all type with uniform sampling

**表 7** 均匀抽样下各种方法的类准确率

| Method | WWW | MAIL | BULK | ATT | P2P | DB | SERV |
|---|---|---|---|---|---|---|---|
| NB (%) | 44.33 | 82.67 | 17.38 | 23.38 | 22.64 | 38.29 | 88.50 |
| NBK (%) | 55.74 | 87.30 | 13.57 | 37.54 | 40.11 | 65.29 | 96.16 |
| C4.5 (%) | 89.81 | 93.87 | 91.70 | 73.99 | 84.13 | 98.54 | 97.29 |

**Table 8** Trust of all type with uniform sampling

**表 8** 均匀抽样下各种方法的类可信度

| Method | WWW | MAIL | BULK | ATT | P2P | DB | SERV |
|---|---|---|---|---|---|---|---|
| NB (%) | 98.17 | 85.02 | 28.43 | 0.27 | 2.80 | 13.10 | 16.37 |
| NBK (%) | 98.32 | 88.40 | 67.98 | 0.63 | 4.15 | 16.23 | 27.96 |
| C4.5 (%) | 99.69 | 97.88 | 55.80 | 16.87 | 9.79 | 83.88 | 60.83 |

从表 7 和表 8 所示的实验结果来看,在均匀抽样的条件下,3 种机器学习算法所得分类模型在处理网络流样本时,大类样本 WWW 的类准确率都有所下降,其中 C4.5 决策树方法下降了 10%左右.这主要是因为训练数据集仅有 897 条 WWW 样本,相对于整个数据集中 20 多万条 WWW 流来说,采样相对不足,从而导致分类结果的准确性略有下降.然而,NB 和 NBK 方法则有所不同,它们的 WWW 类准确率下降了 50%左右,这不仅是因为采样相对不足而带来一定的信息损失,更由于均匀采样失去了训练数据集中的先验概率信息,从而导致 WWW 的类准确率显著下降.由此可见,基于贝叶斯定理的机器学习方法在处理流量分类问题时,过于依赖先验概率信息,具有明显的不足.

5.1.2 CAS_Set 上的结果

为了进一步对比分析 C4.5 决策树和 NB 类方法的分类准确率,本文在 CAS_Set 数据集上分别进行了分层抽样和均匀抽样的实验.由于实验结果相近,为节省篇幅,下文仅对两种实验场景下整体分类准确率进行比较.

我们首先进行分层抽样的实验,将 CAS_Set 数据集均匀划分为两个数据子集:CAS1 和 CAS2,在这两个数据子集中每类样本的比例与 CAS_Set 数据集保持一致.从 CAS1 中分别抽取每类应用 0.1%的样本构成训练数据集(由于 CAS_Set 中 INT 型的样本数量相对较少,不足 0.1%,因此在使用 CAS_Set 的所有实验中,都将 INT 型样本忽略不计).然后,在训练数据集上分别运行 NB,NBD,C4.5 这 3 种机器学习算法获得相应的流量分类模型,并使用这些分类模型来处理测试数据集 CAS2;随后再将训练数据集的规模逐步提高到 CAS1 的 1%,10%,50%.重复上述实验 5 次,所得实验结果见表 9.

**Table 9** Overall accuracy of 3 methods using stratified sampling

**表 9** 利用分层抽样构造训练数据集时,3 种分类方法的整体分类准确率

| Training set | Testing set | NB (%) | NBD (%) | C4.5 (%) |
|---|---|---|---|---|
| 0.1% Set1 | Set2 | 55.60 | 80.80 | 86.08 |
| 1% Set1 | Set2 | 45.91 | 84.86 | 93.26 |
| 10% Set1 | Set2 | 49.87 | 87.79 | 97.22 |
| 50% Set1 | Set2 | 58.45 | 92.23 | 98.80 |

从表 9 中明显可以看出 NB 方法由于直接使用高斯分布假设而无法有效地拟合网络流属性分布,因而分类结果明显较差;NBD 方法和 C4.5 方法不仅具有较高的分类准确率,而且分类准确率随着训练数据集规模的增大而平稳上升.但是,对比 NBD 和 C4.5 方法分类结果,明显可以看出,C4.5 方法的分类结果始终优于 NBD 方法.这主要是因为在本文的实验数据集中,网络流属性存在一定程度的相关性,因而导致 NBD 方法的条件独立假设难以保证,从而导致分类准确率相对较低.

与第 5.1.1 节中的实验一样,在进行均匀抽样实验时,本文首先从 CAS1 的网络流中抽取每种应用的 500 条随机样本,合成一个各类样本均等的训练数据集,再分别运行 NB,NBD 和 C4.5 这 3 种机器学习算法以获得相应的流量分类模型,并以此模型在测试数据集 CAS2 上进行验证.再依次将训练数据集中每类样本的条数提高到 1 500,2 500,3 500 和 4 500(由于 CAS1 中 Mail 型的网络流样本最多只有 4 727 条).重复上述实验 5 次,所得实验结果见表 10.

**Table 10** Overall accuracy of 3 methods using uniform sampling

**表 10** 利用均匀抽样构造训练数据集时,3 种分类方法的分类准确率

| Training set | Testing set | NB (%) | NBD (%) | C4.5 (%) |
|---|---|---|---|---|
| Every type flow num=500 | Set2 | 59.36 | 83.34 | 89.72 |
| Every type flow num=1500 | Set2 | 50.39 | 85.55 | 93.81 |
| Every type flow num=2500 | Set2 | 48.84 | 85.52 | 94.62 |
| Every type flow num=3500 | Set2 | 44.86 | 86.04 | 95.35 |
| Every type flow num=4500 | Set2 | 42.89 | 86.50 | 95.85 |

从表 10 可以看出,在上述 3 种机器学习方法的分类结果中, NB 方法的整体分类准确率最差,不仅没有随着训练数据集的增大而提高,相反,却随着训练数据集的增大而明显下降;NBD 方法虽然整体上分类准确率略有上升,但训练集规模在从每类样本 1 500 条增加到每类样本 2 500 条时,分类准确率出现了轻微的下降.C4.5 方法的整体分类准确率最好,不仅始终优于 NB 和 NBD 方法,而且随着训练数据集的增大而稳步上升.由此可见,利用 C4.5 决策树处理流量分类问题,在分类稳定性上具有明显的优势.

### 5.2 效率性指标

由于本文实验使用了 Moore_Set,每次实验都必须使用 FCBF 算法对当前的训练数据集进行属性选择.训练数据集的差异决定了每次使用 FCBF 算法的结果各不相同.通过直接计算每次实验的训练时间和测试时间来衡量分类模型的效率没有可比性.为此,我们选择一组固定的网络流属性集合上的流量分类模型来分析不同机器学习算法的效率.为了保证所选网络流属性集合的代表性,在整个 Moore_Set 数据集上运行 FCBF 算法得到 7 个属性.对整个 Moore_Set 数据集进行属性过滤后,随机选择其中 10%的样本作为训练数据集,再利用剩余 90%的样本作为测试数据集.在训练数据集上分别运行 NB,NBK 和 C4.5 算法来获得不同的流量分类模型.随后,使用这些流量分类模型来处理相应的测试数据集.重复上述实验 10 次,得到的实验结果见表 11.

**Table 11** Training time and testing time of 3 methods using Moore_Set

**表 11** Moore_Set 上 3 种方法的模型训练时间和测试时间

| Time | NB | NBK | C4.5 |
|---|---|---|---|
| Training time (s) | 0.47 | 0.48 | 8.66 |
| Testing time (s) | 19.47 | 51.74 | 2.46 |

再将 CAS_Set 按照分层抽样的方法每次抽取 10%的样本作为训练数据集,而使用剩余的 90%作为测试数据集.利用 Weka 软件包分别运行上述 3 种机器学习方法.重复上述实验 10 次,所得实验结果见表 12.

**Table 12** Training time and testing time of 3 methods using CAS_Set

**表 12** CAS_Set 上 3 种方法的模型构造时间和测试时间

| Time | NB | NBK | C4.5 |
|---|---|---|---|
| Training time (s) | 8.74 | 24.37 | 103.11 |
| Testing time (s) | 166.72 | 18.74 | 8.14 |

对比 3 种方法的结果可以看出,C4.5 决策树方法的模型构建过程相对复杂,所消耗的模型训练时间相对较长,但 C4.5 决策树方法在数据处理速度上具有明显的优势.这主要是因为利用 C4.5 决策树模型进行流量分类时,仅需根据网络流样本的属性值在决策树上进行自上而下的简单比较,处理相对简单.而 NB,NBK,NBD 模型在处理待分类网络流样本时,首先需要计算出样本属于每种类型的概率,再从中选择最大概率最大的应用类型作为网络样本的所属类型;计算过程相对复杂.在实际应用场景中,流量分类问题通常无须频繁重建分类模型,只要求分类模型能够在短时间内处理大量的网络流.因此,与 NB 及其改进方法相比,C4.5 决策树方法更适合处理大规模流量分类问题.

## 6 小 结

利用机器学习方法处理流量分类问题是近年来网络流量分类领域一个新兴的研究热点,其中朴素贝叶斯方法及其改进算法以其实现简单、分类高效而被大多数研究人员所接受.本文通过系统分析朴素贝叶斯方法的原理,从理论上证明了以贝叶斯定理为基础的机器学习方法在处理流量分类问题时具有潜在的不稳定性,并引入 C4.5 决策树方法来处理的流量分类问题.在 Moore_Set 和 CAS_Set 数据集上进行对比实验可以看出:(1) C4.5 决策树方法不依赖于网络流样本的先验概率,能够有效地避免网络流样本分布变化所带来的消极影响.(2) C4.5 决策树模型在处理待测网络流样本时,仅需进行属性值比较,处理相对简单,在处理大规模流量分类问题时具有明显的性能优势.

由于本文目前所采用的网络流属性大部分必须等待网络流结束后才能获取,因此本文的方法还不能实现真正意义上的网络流在线识别.因此,如何利用网络流初始的若干数据分组实现网络流属性的实时提取,将是本文下一步的主要研究工作之一.

**References**:

[1] Moore AW, Zuev D. Internet traffic classification using Bayesian analysis techniques. In: Proc. of the 2005 ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. Banff, 2005. 50−60. http://www.cl.cam.ac.uk/~awm22/publications/moore2005internet.pdf

[2] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification. In: Proc. of the 14th IEEE Int'l Symp. on Modeling, Analysis, and Simulation. Monterey, 2006. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1698549

[3] Moore AW, Papagiannaki K. Toward the accurate identification of network applications. In: Dovrolis C, ed. Proc. of the PAM 2005. LNCS 3431, Heidelberg: Springer-Verlag, 2005. 41−54.

[4] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark. In: Proc. of the ACM SIGCOMM. Philadelphia, 2005. 229−240. http://conferences.sigcomm.org/sigcomm/2005/paper-KarPap.pdf

[5] Roughan M, Sen S, Spatscheck O, Dutfield N. Class-of-Service mapping for QoS: A statistical signature-based approach to IP traffic classification. In: Proc. of the ACM SIGCOMM Internet Measurement Conf. Taormina, 2004. 135−148. http://www.imconf.net/imc-2004/papers/p135-roughan.pdf

[6] Zuev D, Moore AW. Traffic classification using a statistical approach. In: Dovrolis C, ed. Proc. of the PAM 2005. LNCS 3431, Heidelberg: Springer-Verlag, 2005. 321−324.

[7] Nguyen T, Armitage G. Training on multiple sub-flows to optimise the use of Machine Learning classifiers in real-world IP networks. In: Proc. of the 31st IEEE LCN 2006. Tampa, 2006. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4116573

[8] Eerman J, Mahanti A, Arlitt M. Internet traffic identification using machine learning techniques. In: Proc. of the 49th IEEE GLOBECOM. San Francisco, 2006. http://pages.cpsc.ucalgary.ca/~mahanti/papers/globecom06.pdf

[9] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms. In: Proc. of the ACM SIGCOMM Workshop on Mining Network Data (MineNet). Pisa, 2006. http://conferences.sigcomm.org/sigcomm/2006/papers/minenet-01.pdf

[10] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Proc. of the Conf. on Future Networking Technologies 2006 (CoNEXT 2006). Lisboa, 2006. http://portal.acm.org/citation.cfm?id=1368445

[11] Paxson V. Measurements and analysis of end-to-end Internet dynamics [Ph.D. Thesis]. Berkeley: University of California, 1997.

[12] Wang R, Liu Y, Yang YX, Zhou XY. Solving the app-level classification problem of P2P traffic via optimized support vector machines. In: Proc. of the 6th Int'l Conf. on Intelligent Systems Design and Applications (ISDA 2006). Ji'nan, 2006. 534−539. http://portal.acm.org/citation.cfm?id=1173623

[13] Li N, Chen ZL, Zhou G. Network traffic classification using rough set theory and genetic algorithm. In: Huang DS, Li K, Irwin GW, eds. Proc. of the ICIC 2006. LNCIS 344, Berlin, Heidelberg: Springer-Verlag, 2006. 945−950.

[14] Guan XQ. Research on the classifying algorithm based on decision tree [MS. Thesis]. Taiyuan: Shanxi University, 2006 (in Chinese with English abstract).

[15] Moore AW, Zuev D, Crogan M. Discriminators for use in flow-based classification. Technical Report, RR-05-13, Queen Mary University of London, 2005.

[16] NetMate. 2007. http://sourceforge.net/projects/netmate-meter/

[17] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed., San Francisco: Elsevier Inc., 2005.

**附中文参考文献**:

[14] 关晓蔷.基于决策树的分类算法研究[硕士学位论文].太原:山西大学,2006.

**徐鹏**(1981－),男,博士,主要研究领域为网络测量,数据挖掘.

**林森**(1983－),男,硕士,主要研究领域为网络测量,数据挖掘.