基于决策树集成的P2P流量识别研究

刘三民[1,4] 孙知信[1,2,3] 刘余霞[5]

(南京航空航天大学信息科学与技术学院 南京 210016)[1]

(南京邮电大学计算机技术研究所 南京 210003)[2]

(南京大学计算机软件新技术国家重点实验室 南京 210093)[3]

(安徽工程大学计算机与信息学院 芜湖 241000)[4] (安徽工程大学电气工程学院 芜湖 241000)[5]

摘 要 为提高分类模型的稳定性,提出基于决策树分类器集成方案用以识别流量。模型首先利用特征选择方法(FCBF)提取最优分类特征信息,按Bagging随机抽样原理形成5个子分类器,依少数服从多数原则生成决策模型。利用两种实验方案在公开数据集上进行测试,结果表明提出的方案比贝叶斯、基于核密度估计贝叶斯方案具有更好的稳定性、模型分类准确率和P2P流量识别准确率,并对此现象进行了解释。

关键词 流量识别,集成学习,决策树,贝叶斯分类,稳定性

中图法分类号 TP393 **文献标识码** A

Research of P2P Traffic Identification Based on Decision Tree Ensemble

LIU San-min[1,4] SUN Zhi-xin[1,2,3] LIU Yu-xia[5]

(College of Information Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)[1]

(College of Computer of Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[2]

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[3]

(College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China)[4]

(College of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China)[5]

Abstract A novel P2P traffic identification method based on decision tree ensemble was proposed for improving the model stability. First, the most optimal feature set was extracted by using fast correlation based filter(FCBF), and then the decision model based on five sub-classifier formed by Bagging was developed by the principle of the majority. Through test result comparison based on the open data set in the two distinct experiment scheme among the proposed model, naïve bayes and naïve bayes based on kernel density estimation, it shows the proposed model owns a better stability, the high classification accuracy and P2P traffic identification accuracy and gives the explanation about this phenomenon.

Keywords Traffic identification, Ensemble learning, Decision tree, Bayes classification, Stability

1 引言

网络流量识别对于科学规划网络、链路堵塞分析、QoS调度、降低运营成本等具有十分重要的意义。目前P2P技术广泛应用于文件共享、流媒体、即时通讯、VoIP等方面,成为网络应用的主要方式之一,据统计其流量占网络流量的70%左右[1],成为真正网络"流量杀手",已经严重影响到传统的网络服务,因此P2P流量识别成为研究热点。

早期流量识别研究主要是基于端口识别方法,Sen等[2]采用被动测量方式,基于端口识别原理对当时流行的3种P2P文件共享系统进行流量识别。国内学者张云飞等[3]首次使用基于固定端口的方法对国内骨干网上的P2P流量进行测量与分析,取得较好的识别效率。但各P2P应用采用端口跳变、HTTP隧道技术躲避流量检测和过滤,同时网络中大量采用地址转换技术,使得基于端口检测的方法识别效果较差。

鉴于P2P应用有其具体的协议特征,Sen等[4]通过完全提取包负载的方法,重点研究5种流行的P2P流量分类,实验结果表明5种应用的误报率和漏报率均小于5%。Moore等[5]详细分析基于端口识别方法的缺陷,采用9种识别途径构建出基于内容特征流量识别体系,实验表明它在10种常见网络服务上具有较好的识别精度,同时对识别精度和复杂性提出了折衷方案。Satoshi等[6]通过设置诱饵节点方法,结合应用层特征识别技术识别Winny系统流量。但协议加密和

到稿日期:2010-12-26 返修日期:2011-03-03 本文受国家自然科学基金(60973140),江苏省自然基金(BK2009425),江苏省高校自然科学基础研究项目(08KJB520005),江苏省六大才高峰项目,安徽工程大学校青年基金项目(2008YQ041)资助。

刘三民(1978-),男,博士生,讲师,主要研究方向为模式识别、流量检测,E-mail:aqlsm@163.com;孙知信(1964-),男,博士,教授,博士生导师,主要研究方向为网络安全、对等计算。

新应用的出现使得基于协议特征识别方法表现出较多局限性。

Karagiannis 等[7]提出两条启发规则：(1)TCP/UDP 启发式：若一 IP 对之间同时存在 TCP/UDP 协议，则这 IP 对间除基于知名端口的非 P2P 流量外，其他流量均视为 P2P 流量；(2){IP,Port}启发式：若某{IP,Port}所连不同 IP 数和不同端口数之差小于特定阈值，则该条连接是 P2P 应用。该方案实现基于 P2P 流的连接模式识别，具有较高的识别效率，同时对未知 P2P 流量具有鉴别能力。由于国内网络差异，中科院徐鹏等[8]对文献[7]提出的算法进行改进并提出 3 条针对性的启发规则：基于非 P2P 知名端口的过滤机制；基于有效数据流的计数机制；基于反向流的 FTP 过滤机制。改进后的识别算法在实验中表现出优异性能。Karagiannis 等[9]实现基于主机行为模式的盲分类算法(BLINC：Blind Classification)，算法从社会层、功能层和应用层 3 个层面反映主机行为特点。实验结果表明，BLINC 算法能够达到 95%的精度，但由于模型较为复杂难以保证实时性要求。

目前 P2P 流量识别领域的热点是运用机器学习方法对 P2P 流量进行识别。2005 年 Zuev 等[10]首次提出基于概率模型的朴素贝叶斯方法，流识别率达 67%，该方法要求属性特征独立和遵循高斯分布，但这两点难以保证。在此基础上 Moore 等[11]采用基于关联的快速过滤 FCBF 算法[12]提取流量特征和核密度估计拟合分布函数方法，使得识别效率显著提高。2007 年 Moore 等[13]构建三层贝叶斯神经网络的结构，采用最小风险决策策略，实现流量识别，但该类方法具有局限性：网络结构层数和神经元个数难确定；容易陷入局部极小值；有过学习现象存在。文献[14,15]分别采用决策树模型和支持向量机进行流量识别研究。文献[16]采用多维支持向量机模型，其表现出较好的流量识别能力。决策树模型和支持向量机不需要较多先验知识和数据分布信息，在上述 3 文献中均表现出较好的识别精度。

分类模型具有时间效应属性，表现为模型的稳定性。现有基于机器学习的 P2P 流量识别方法，基本上是基于单一的分类器模型的识别方案，其模型识别准确率有待提高，模型的稳定性均未考虑。所以本文在现有研究基础上，结合集成学习理论，提出基于对数据较敏感的决策树模型进行集成，实现 P2P 流量识别。通过在公开的 Moore 数据集上的实验分析发现，模型具有较好稳定性，识别效率有明显改善。本文第 2 节介绍集成学习基本概念；第 3 节介绍基于决策树分类器的集成模型设计；第 4 节介绍实验仿真；最后是总结。

2 集成学习基本概念

Valiant 提出学习性问题[17]后，在机器学习领域，很多研究者致力于 PAC(probably approximately correct)学习模型的研究。Kearns 和 Valiant 引入强可学习和弱可学习概念[18,19]，在此基础上 Schapire 提出著名的弱学习定理[20]：在无法或较难找到强学习算法时，可以将一个比随机猜测好的弱学习算法提升成为强学习算法，并通过构造性方法证明这一结论。Thomas G. Dietterich 从统计、计算和表示 3 个方面证明集成学习的有效性和可行性[21]，集成学习开始成为机器学习的研究热点。按照集成基分类器模型是否相同分为异态集成学习和同态集成学习。常用的集成方案有 Bagging (Bootstrap Aggregating)[22]和 AdaBoost[23]。Bagging 是有放回地抽取训练样本，从而为基本分类器构造出一个同样大小但各不相同的训练集，它可训练出不同的基本分类器。AdaBoost 最早由 Robert T. Schapire 提出，基本思想是对分错训练样本加强学习，执行过程描述如下：首先给每一个训练样本赋予相同的权重，然后训练第一个基本分类器并用来测试训练集，对分类错误的样本提高权重(实际算法中是降低分类正确样本的权重)，然后用调整后的带权训练集训练第二个基本分类器，重复该过程直到获得足够好的学习器。为保证子分类器的相异度，本文采用 Bagging 方法通过多次随机取样形成相异的训练集。

3 基于决策树分类器的集成模型设计

基于机器学习的流量识别方案大多以模型的准确度作为关键指标，而单一的分类模型难以全面地反映出全局信息，其稳定性和准确率有待改善。稳定性是指某时刻训练所得的分类模型能够在未来一段时间内仍然具有较好的分类性能。这点现有的分类模型难以做到，因为无法用少量局部信息完整反映出全局信息。

3.1 子分类器分类原理

决策树分类器具有几个优点：对训练集敏感，适合集成学习；训练模型时不需要过多先验知识；对需分类样本处理较简单，有较好的时间优势。因此本文选择决策树子分类器进行集成实现流量识别。

决策树分类器原理可作如下描述：设训练样本集为 S，样本共有 n 类，记为 $C=\{C_1,C_2\cdots C_n\}$。则样本集合 S 的不确定程度即信息熵(Entropy)如式(1)所示。

$$I(S)=-\sum_{i=1}^{n}p_i\log_2(p_i) \tag{1}$$

式中，p_i 表示训练样本集合中属于第 i 类的概率。若 A 为流量对象属性，具有 v 个不同值$\{a_1,a_2,\cdots,a_v\}$，则相应可将 S 划分为 v 个子集$\{S_1,S_2,\cdots,S_v\}$；其中 S_j 由 S 中样本属性 A 值为 a_j 的样本构成。设 S_{ij} 是子集 S_j 中类 C_i 的样本数。根据属性 A 划分的熵为：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{nj}}{s}I(s_{1j},\cdots,s_{nj}) \tag{2}$$

项$\frac{s_{1j}+\cdots+s_{nj}}{s}$充当第 j 个子集的权。对于给定的子集 S_j，有：

$$I(s_{1j},s_{2j}\cdots,s_{nj})=-\sum_{i=1}^{n}p_{ij}\log_2 p_{ij} \tag{3}$$

式中，p_{ij} 表示 S_j 中样本属于类 C_i 的概率。则属性 A 的信息增益(Gain(A))：$Gain(A)=I(S)-E(A)$。每次选择增益值最大的属性作分支节点即可。为了避免属性多值情况所占优势，常以信息增益率作为选择分支属性的标准。

$$Ratio(S,A)=Gain(S,A)/Split(S,A) \tag{4}$$

式中，$Split(S,A)=-\sum_{j=1}^{v}\frac{|S_j|}{|S|}\log_2\frac{|S_j|}{|S|}$。

训练过程按照上述分支属性选择方法，自顶向下形成决策树分类器。内部节点表示分支属性，叶节点代表类。决策树分类器形成后，从根到叶节点提取合取范式，形成分类规则。

3.2 基于决策树分类器的集成模型

决策树分类器对训练数据较敏感，基于多个决策树分类器集成方案能够较全面地反映出数据分布信息，集成后能够显著提高模型的稳定性和准确率。基于决策树集成的流量识别方案如图 1 所示。

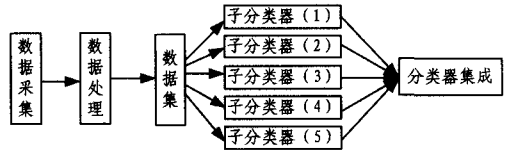


图 1　基于决策树集成的流量识别流程图

训练过程：按集成学习原理，若各子分类器相异度较高，错误率小于 0.5 则集成后将会显著提高泛化能力，这是本文集成模型的理论支撑。集成模型设置了数据采集模块，它可以从网络出口处获取实时流数据。数据处理模块是对采集好的原始数据集进行特征提取、空缺值处理和属性归一化或离散化，并在此基础上形成所需训练数据集。对采集好的数据集按照重复采样原理（Bootstrap Sampling），构建相异训练子数据集，形成 5 个基于决策树的子分类模型。

识别过程：从网络出口处采集到网络流，按要求进行数据处理后输入 5 个子分类器进行决策，对决策结果按照少数服从多数的原则进行集成，输出识别结果。

4　实验仿真

4.1　实验环境

为便于对比和验证，本文采用 Moore 数据集[24]。该数据集采自 3 个研究所总共大约有 1000 名员工，通过千兆全双工方式与网络连接。原始数据记录于 2003 年 8 月 20 日 0 时开始 24 小时内流经网络出口的所有网络流量。构建数据集时，选用语义完整的 TCP 双向流作为网络流样本，包含 377526 个网络流样本，分为 10 种类型，表 1 详细列举了应用类型和各自所包括的具体应用。根据时间顺序按区段划分为 10 个数据子集，每样本流共有 249 项属性，最后一项是网络应用类型属性，详细描述可参考文献[24]。

本文仿真环境是 Weka3.6.2 平台、新西兰怀卡托大学 Witten 教授等人基于 Java 开发的开源平台，本身集成常见的分类模型有决策树、贝叶斯和支持向量机等机器学习算法，用以实现分类、关联规则、聚类分析和特征过滤等功能。

表 1　应用类型

类型	具体应用
WWW	www
MAIL	Imap,pop,smtp
BULK	ftp
DB	Postgres,sqlnet,oracle,ingres
SERV	X11,dns,ident,ldap,ntp
P2P	KaZaA,BitTorrent,Gnutella
ATT	Internet worm and virus attack
MULT	Windwos Media Player,Real
INT	Ssh,klogin,rlogin,telnet
GAME	Half-Life

4.2　结果分析

原数据集属性特征较多，计算量大而且可能存在大量冗余属性和无关属性，因此本文同文献[13]首先利用 FCBF 算法选择排名前 10 位属性特征作为流量输入特征空间。所选特征不包含端口信息，以克服基于端口识别的局限性。

在 3 种分类模型中贝叶斯分类（NB：Narve Bayes）、基于密度估计的贝叶斯分类（NBK：Naïve Bayes using kernel density estimation）和集成学习方案（ELS：Ensemble Learning Schema）均采用如下策略，标记方案 1：在各自训练集建立分类模型，用数据子集（DS10）作为统一的测试集，其目标是使训练集与测试集采集时间相距最大化，以便于验证分类模型的稳定性。主要从 3 方面对模型进行评估：整体的识别准确率；P2P 流量识别真正率；P2P 流量假正率。模型准确率指样本正确分类所占的比例，用来反映整体性能好坏，是一个重要的评价指标，被大多数研究人员认可。真正率（TPR，True Positive Rate）指某类样本正确判断所占的比例值，能够反映出该模型对此类应用的识别好坏。假正率（FPR：False Positive Rate）指分类模型把其它应用类型错误地判为 P2P 应用，在涉及风险决策最小化时具有重要作用。

定义两个统计量平均值（med）和偏差（var），用来刻画 3 种分类模型的性能好坏和稳定性，其中 $x[i]$ 表示每次实验的结果。

$$med=\frac{1}{10}\sum_{i=1}^{10}x[i]$$
$$\mathrm{var}=\frac{1}{10}\sum_{i=1}^{10}(x[i]-med)^2 \tag{5}$$

从表 2 可以看出集成方案在各个数据子集上的准确率平均好于另外两种方案，而且 NBK 方法的模型准确率比 NB 好。NB 的分类模型建立在各属性特征相互独立基础之上并服从高斯分布，因此难以满足。在方案中事先进行了特征选择，在一定程度上降低了属性间的依赖性，从准确率反映明显比文献[10]要好。NBK 否定属性特征服从高斯分布的假设，利用训练集拟合分布空间，所以性能好于 NB 方案，但它基于数据子集拟合的局部结果，不能完整反映出全局分布信息。基于贝叶斯定理分类模型，除了类条件概率分布外，还有先验概率问题，在 NB 和 NBK 两方案中，都是简单地用训练数据集中各类样本所占比例来拟合先验概率，其值具有一定的随机性。集成方案利用 5 个子分类器，在数据集上采用抽样形成 5 个训练子集，充分反映出分布样本的分布信息。表 2 中均值反映上述信息。从偏差值可得集成方案平稳特性最好，具有较长的使用寿命，NBK 次之。而且在实验中发现与测试子集（DS10）在时间相邻近的子集上建立的分类模型，其准确性较好，呈现出非严格递增现象，这也证实分类模型具有时间效应。

表 2　方案 1 准确率（x%）

方案	均值	偏差
ELS	95.36	2.65
NB	87.02	6.00
NBK	92.47	5.70

表 3 包含两个指标值：TPR 和 FPR，这两个指标能够反映分类模型在 P2P 应用中识别的好坏。通常追求高 TRP 值和低 FPR 值，特别在有决策风险的环境中，因为错误的肯定所带来的风险比错误的否定要严重。从表中数据明显说明集成方案优于另外两种，NBK 次之。表 4 列举了各样本集合大小和所含 P2P 样本个数，其中 P2P 样本所占比例较小，导致表 3 数据值偏小。在实验中凡是 P2P 样本数较多的子集所得分类模型，在测试时 P2P 识别效果较好，这说明由于样本

数过少，在建立模型时缺乏样本有效的分布信息。样本不平衡是流量识别中面临的关键问题。

表3 P2P流量识别率

方案		均值	偏差
ELS	TPR	0.336	0.052
	FPR	0.001	2.2E-6
NB	TPR	0.042	0.012
	FPR	0.009	5.1E-5
NBK	TPR	0.059	0.007
	FPR	0.006	2.4E-5

表4 数据集样本数

子集	总样本数	P2P样本数	子集	总样本数	P2P样本数
DS1	24863	339	DS6	19384	94
DS2	23801	94	DS7	55835	116
DS3	22932	100	DS8	55494	289
DS4	22285	114	DS9	66248	249
DS5	21648	75	DS10	65036	624

实验过程及数据证明了基于决策树集成分类器具有较好的识别性能和稳定性。

为说明分类模型具有时间效应，本文重新制定实验策略，标记方案2：在各自数据子集上训练模型，在其相邻的下一个数据集上测试，这样保证训练集和测试集在时间上邻近。表5中数据均值大于表2中数据，而偏差值小于表2中偏差，说明分类模型具有时间属性，具有一定的生命周期，在一段时间后需要重新训练新模型。从表5中可以进一步说明集成方案具有更好的稳定性和性能。

表5 方案2准确率

方案	均值	偏差
ELS	97.76	1.05
NB	92.27	12.12
NBK	95.55	2.45

结束语 本文实现了基于决策树集成的流量识别模型，实验证明该方案具有较好的稳定性和准确性。在流量识别过程中，通常面临的是多类识别问题，按集成学习理论，其子分类器个数和类别多少具有一定的关系。在本次仿真实验中，采用的数据集样本均是具备完整语义的网络流样本，而在实时网络环境中，通常存在大量不完整、较短的网络流量。在实时流量识别环境中，除了准确性和稳定性外，其时间因素也是至关重要的。这些问题在本文中均未涉及，如何改善时间效率和确定子分类器个数，并将其处理对象扩展至语义不完整的网络对象，将是下一步研究重点。

参考文献

[1] Cache logic. Peer-to-Peer in 2005 [OL]. http://www.cachelogic.com/home/pages/research/p2p2005.php,2005

[2] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks[J]. IEEE/ACM Trans on Networking, 2004, 12(2): 219-232

[3] 张云飞，雷连虹，陈常嘉. Internet中peer-to-peer应用流量测量与分析[J]. 铁道学报，2004，26(5)：55-60

[4] Sen S, Spatscheck O, Wang D. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures[C]// Proceedings of the 13th international conference on WWW 2004. New York, USA, 2004: 512-521

[5] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]// PAM 2005. LNCS 3431. Heidelberg: Springer-Verlag, 2005: 41-54

[6] Satoshi O, Yoichi H, Matsuaki T, et al. a traffic identification method and evaluations for a pure p2p application[C]// PAM 2005, LNCS 3431. 2005: 55-68

[7] Karagiannis T, Broido A, Faloutsos M, et al. transport layer identification of p2p traffic[C]// Proceedngs of the 4th ACM SIGCOMM conference on Internet measurement. Taormina, Sicily, Italy: ACM Press, 2004: 121-134

[8] 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

[9] Karagiannis T, Papagiannaki K, Faloutsos M. Blinc: Multilevel Traffic Classification in the Dark[C]// ACM SIGCOMM. 2005: 229-240

[10] Zuev D, Moore A W. Traffic classification using a statistical approach[C]// Dovrolis C, ed. Proc of the PAM 2005. LNCS 3431. Heidelberg: Springer-Verlag, 2005: 321-324

[11] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]// Proc of the 2005 ACM SIGMETRICS Int'l Conf on Measurement and Modeling of Computer Systems. Banff, 2005: 50-60

[12] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]// Proc. of the 20th International Conference on Machine Learning (ICML 2003). 2003

[13] Auld T, Moore A W, Gull S F. Bayesian Neural Networks for Internet Traffic Classification[J]. IEEE Transactions on Neural Networks, 2007, 18(1): 223-239

[14] 徐鹏，林森. 基于C4.5决策树的流量分类方法[J]. 软件学报，2009，20(10)：2692-2704

[15] 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

[16] 孙知信，张玉峰. 基于多维支持向量机的P2P网络流量识别模型[J]. 吉林大学学报，2010，40(5)：1298-1302

[17] Valiant L G. A theory of learnable[J]. Communication of the ACM, 1984, 27(11): 1134-1142

[18] Kearns M, Valiant L G. Learning Boolean formulate or factoring [R]. TR-1488. Cambridge, MA: Havard University Aiken Computation Laboratory, 1988

[19] Kearns M, Valiant L G. Crytographic limitation on learning Boolean formulae and finite automata[C]// Proc. of the 21st Annual ACM Symp on Theory of Computing. New York: ACM, 1989: 433-444

[20] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2): 197-227

[21] Dietterich T G. Ensemble learning. The Handbook of Brain Theory and Neural Networks (Second Edition)[M]. Ensemble Methods in Machine Learning. In Multiple classier Systems, Cagliari, Italy, 2000

[22] Breiman L. Bagging Predictors [J]. Machine learning, 1996, 24 (2): 123-140

[23] Freund Y, Schapire R E. A Decision theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Scienses, 1997, 55(1): 119-139

[24] Moore A W, Zuev D, Crogan M. Discriminators for use in flow-based classification[R]. RR-05-13. London: Queen Mary University of London, 2005