机器学习在P2P流量检测中的研究

吴 敏,王汝传,蔡涛涛
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:P2P流量逐渐成为了互联网流量的重要组成部分,在对Internet起巨大推动作用的同时,也带来了因资源过度占用而引起的网络拥塞以及安全隐患等问题,妨碍了正常的网络业务的开展。文中提出了基于机器学习的P2P流量识别方案,并运用FCBF(Fast Correlation-Based Filter)特征选择算法形成了流量特征子集,构建了机器学习P2P流量识别模型并对比了几种常见的机器学习算法在流量识别方面的性能。测试实验结果表明,C4.5算法和贝叶斯网络算法都适合于P2P流量检测,其个别模型达到了90%以上的识别率。

关键词:对等网络;流量识别;机器学习算法;特征选择
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2010)11-0133-04

Study on Applications of Machine Learning in P2P Traffic Identification

WU Min, WANG Ru-chuan, CAI Tao-tao
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: P2P traffic has taken great portions in the network traffic. While having a significant impact on the Internet, it brings serious problems such as network congestion and traffic hindrance caused by the excessive occupation in the bandwidth. Proposes a P2P traffic identification based on machine learning. Firstly the FCBF(Fast Correlation-Based Filter) feature selection algorithm is used to select the attribute features subset, then P2P flows identification model is built and several machine learning algorithms are compared. The result showed that in P2P traffic identification based on machine learning algorithms, C4.5 and Bayesian network was feasible and the identification accuracy of some models can reach above 90 percent.

Key words: P2P; identification of traffic; machine learning algorithm; feature selection

0 引 言

随着P2P网络[1]技术的兴起,P2P流量逐渐成为了互联网流量的重要组成部分。精确地识别P2P流量对于有效地管理网络和合理地利用网络资源都具有重要意义。但是越来越多的P2P应用使用任意端口及采用加密的方法,使得利用端口识别和关键字检测等P2P流量识别方法已经逐步遭到淘汰[2]。而且,人们需要实时地识别出P2P流量以后才能够实现对P2P流量的控制,从而提高网络的性能。因此有研究人员将机器学习[3,4]应用到流量识别中[5~7],将流量作为分析粒度,每个流都由一系列相同的统计特征来代表,运用机器学习算法来找出流量之间的异同从而进行分类[8,9],然而现阶段对基于机器学习的流量分类算法的评判标准还未明确。文中使用英国剑桥大学计算机实验室的网络公开数据源[10]作为训练数据集,采用了属性选择算法FCBF选择出适用于P2P应用分类的最佳特征子集,并应用了4种机器学习算法对之进行测试比较,分析各种机器学习方法的结果并评估各种机器学习算法在识别P2P流量时的性能和效果。

收稿日期:2009-10-01;**修回日期**:2010-01-11
基金项目:国家自然科学基金(60973139,60773041);江苏省自然科学基金(BK2008451);省级现代服务业发展专项资金;江苏高校科技创新计划项目(CX09B_153Z, CX08B-086Z);南京邮电大学青蓝工程项目(NY206034, NY208011);江苏省六大高峰人才项目(2008118)
作者简介:吴 敏(1976-),女,江苏泰州人,讲师,博士研究生,研究方向为移动代理技术、分布式计算、计算机密码学和网格计算等;王汝传,教授,博士生导师,研究方向为计算机软件、计算机网络和网格、对等计算、信息安全、无线传感器网络、移动代理和虚拟现实技术等。

1 特征选择

特征选择,也叫属性约简,是指在不丧失特定的应用数据原有价值的基础上去除不相关和冗余的属性,选择最小部分的属性,形成子集。这种方式能够提高

数据的质量,并能够加快学习的速度,特征选择是机器学习过程中的重要的一部分。

从广义上可将属性选择的算法分为过滤器(Filter)和嵌入方式(Wrapper)两种算法,FCBF(Fast Correlation-based Feature Selection)属于后者,所以在处理数据量较大的网路数据上很有优势。一般来说,如果一个特征和某个类的相关性足够,同时它与其它任意特征的相关性又都没达到某一水平,则认为这个特征对这个类来说是好的特征。FCBF 用对阵不确定性(Symmetrical Uncertainty,SU)作为衡量指标,利用了 SU 的值来进行特征选择,SU 取值在[0,1]之间,1 表示两个随机变量可以相互完全预测对方的值,0 则表示两个随机变量彼此独立。SU 的值越大,越能代表其特征的优越性越大。SU 定义如下:

$$SU = 2\left[\frac{IG(X \mid Y)}{H(X) + H(Y)}\right] \tag{1}$$

其中 $IG(X \mid Y)$ 表示信息增益(information gain)。公式如下:

$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{2}$$

$H(X)$ 表示随机变量 X 的熵:

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{3}$$

这里,$P(x_i)$ 为 X 单个分量的先验概率,$P(x_i \mid y_j)$ 为 Y 值已知条件下 X 的条件概率。

在流量识别中进行属性选择不仅能够找到最合适于流量分类的最小属性集合,也能够提高算法性能。实验结果表明在识别的准确率上使用全部的流属性只稍微高于利用属性选择算法选择出来的属性集的准确率,但是在算法效率上后者高出很多。因此属性选择是流量识别过程中的关键的一步。

2 应用机器学习方法识别 P2P 流量

2.1 应用机器学习方法识别 P2P 流量的实现过程

机器学习的方法一般是结合数据挖掘理论用于观测数据(样本),发现数据之间的规律,预测新数据。应用机器学习进行流量识别时,网络流量就是样本,通过学习流量的各种属性特征而发现的规律可以预测流量属于何种应用[10~12]。

对流量识别来说,每个流都可以由一系列相同的属性特征来代表,但是这些属性值不尽相同,所以使用机器学习算法利用这些不同的属性值将网络中的流量匹配到各自应用中。图 1 表明了机器学习在 P2P 流量识别中的实现过程。文中使用了 FCBF 的属性选择方法从文献[3]列举的 248 个流量属性识别中可挑选出最适用于 P2P 应用和非 P2P 应用准确分类的相关属性,并去除掉不相关的属性和冗余的属性,得到一个合适的属性集,然后结合具体的机器学习算法构建出不同的模型。最后利用不同的模型对流量进行分类,并根据分类结果分析算法的准确率和性能。

2.2 算法选择

应用在流量识别中的机器学习算法通常分为有监督学习、无监督学习以及半监督学习三类。根据文献[4],文中只研究有监督学习算法在 P2P 流量检测方面的性能。使用了最常用的 4 种有监督机器学习算法[8]:朴素贝叶斯(Naive Bayes,NB)、贝叶斯网络(Bayesian Networks)、朴素贝叶斯决策树(NBTree)、决策树算法。

1)朴素贝叶斯(Naive Bayes,NB)。

朴素贝叶斯分类器是基于贝叶斯理论的。假设每个数据样本用一个 n 维特征向量来描述其 mn 个属性的值,即:$X = \{x_1, x_2, \cdots, x_m\}$,假定有 n 个类,分别用 $C_1, C_2, \cdots, C_n$ 表示。对一个未知类别的样本 X,先分别计算出 X 属于每一个类别 C_i 的概率 $P(X \mid C_i)P(C_i)$,然后选择其中概率最大的类别作为其所属类别。

2)贝叶斯网络(Bayesian Networks)。

贝叶斯网络是一种基于概率推理的图形化网络,其中贝叶斯公式是该概率网络的基础。它是在 1988 年由 Pearl 提出后,日渐成为近几年来研究热点。贝叶斯网络是一个有向无环图(Directed Acyclic Graph,DAG),包括变量节点和连接这些节点的有向边。其中变量节点代表了随机变量,有向边代表了节点间的相互关系(由父节点指向其后代节点),通过条件概率表达关系强度,对于没有父节点则用先验概率进行信息表达。任何问题都可以用变量节点进行抽象,例如:测试值,观测现象,意见征询等。该方法比较适用于分析和表达不确定性和概率性的事件,能够应用于有条件地依赖多种控制因素的决策,该方法能够从不完全、不精确或不确定的知识或信息中做出推理。

3)朴素贝叶斯决策树(NBTree)。

NBTree 结合了决策树分类器和朴素贝叶斯分类器方法。起初是被设计用来适应大训练数据集的 NBTree,在某些数据集上的准确度已经超越了 C4.5 和朴素贝叶斯算法。

4)决策树算法。

决策树学习是一种归纳学习算法,以实例为基础,主要是着眼于从一组无次序、无规则的事例中推理出分类规则,并用决策树表示。该方法通常用来形成分类器和预测模型,并且可以对未知数据进行分类或预测、数据挖掘等。20 世纪 60 年代以来,该算法广泛应用于分类、预测、规则提取等领域,特别是 1986 年 Qul-

ulna J R 提出 ID3 算法以后,该算法在机器学习、知识发现领域得到了巨大的发展。

3　实验结果分析

3.1　基于 FCBF 的最佳数据集

测试方案使用英国剑桥大学计算机实验室的网络公开数据源[7]作为训练数据,其中 Day1. TCP～Day3. TCP,SiteB. TCP,Day. PF 这 5 个数据集采集自两个不同的站点(拥有 1000 个左右的本地用户和全双工的千兆以太网),采集时间分别是其所属年份的某工作日(持续采集整 24 小时)。如表 1 所示。

表 1　网络公开流量数据集

数据集名称	采集时间	流总数	P2P 流/非 P2P 流	P2P 流比率
Day1. TCP	2003 年	324277	2085/322192	0.64%
Day2. TCP	2004 年	175662	2762/172900	1.57%
Day3. TCP	2006 年	260023	22287/260023	8.57%
Day. PF	2006 年	130572	10871/119701	8.33%
SiteB. TCP	2007 年	248362	17851/230511	7.19%

由于 TCP 协议定义完整,且 TCP 流有清晰的首部和结尾,所以文中只分析 TCP 流量,相应的,这 5 个数据数据集也都是 TCP 流量数据。

文中采用 Weka 平台的 FCBF 模块对上述网络公开数据源进行最佳数据集的筛选。结果 FCBF 选择出 10 个特征,但是由于端口对也是重要的网络流量特征,所以把服务器端口和客户端端口也添加进最佳特征集中,组成了如表 2 所示的特征集。

表 2　最佳特征集

缩写	描述	SU 值
push - pkts - serv	TCP 首部设置的所有数据包总数(服务器到客户机)	0.3165
init - win - bytes - clnt	被送到初始窗口的总字节数(客户机到服务器)	0.2070
init - win - bytes - serv	被送到初始窗口的总字节数(服务器到客户机)	0.3422
avg - seg - size - serv	平均段大小(服务器到客户机)	0.3390
IP - bytes - med - clnt	IP 数据包的平均字节(客户机到服务器)	0.2011
act - data - pkt - clnt	包含大于 1 字节的数据部分的 TCP 数据包总数(客户机到服务器)	0.1722
data - bytes - var - serv	数据包字节数的方差(服务器到客户机)	0.2605
min - seg - size - clnt	最小段大小(客户机到服务器)	0.2131
RTT - samples - clnt	RTT 样本的总数(客户机到服务器)	0.2434
push - pkts - clnt	TCP 首部设置的所有数据包总数(客户机到服务器)	0.2138
serv - port	服务器端口	0.8378
clnt - port	客户机端口	0.0760

3.2　4 种机器学习算法对 P2P 流量识别的性能效果比较

对该 P2P 流量识别模型通过以下参数进行描述:识别准确度(Accuracy),P2P 识别精确率(Precision)和反馈率(Recall)。

识别准确率定义如下:

$$\text{Accuracy}=\frac{TP+FN}{TP+TN+FT+FN}\times 100\% \quad (4)$$

P2P 识别精确率和反馈率定义如下:

$$\text{Precision}=\frac{TP}{TP+FP}\times 100\% \quad (5)$$

$$\text{Recall}=\frac{TP}{TP+TN}\times 100\% \quad (6)$$

式中 TP(True Positive)正确的肯定表示正确分类 P2P 数据包个数;FP(False Positive)错误的肯定表示将 Non－P2P 分类为 P2P 数据包个数;FN(False Negative)错误的否定表示将 Non－P2P 数据包分类成 Non－P2P 的数据包个数;TN(True Negative)正确的否定表示将 P2P 分类为 Non－P2 的数据包个数。

图 1 是各个模型在 10 折交叉验证下的 P2P 流分类情况,其中 X 轴中的 P 表示 P2P 流的精确度 Precision,R 表示反馈率 Recall。从图中可以看出,四种机器学习算法在不同数据集上建模时的反馈率都差不多;但是精确度有很大的差异,比如朴素贝叶斯的精确度就很差,从这一点上就可以看出朴素贝叶斯并不太适合 P2P 流量识别,虽然其拥有很快的建模时间;同时可以看到,C4.5,BN,NBTree 算法在 P2P 流精确度上都维持在很高的水准。

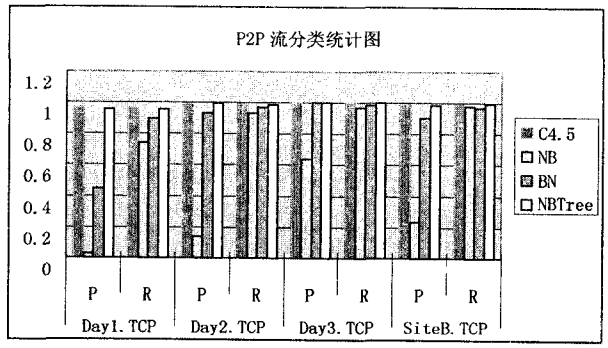

图 1　P2P 流分类统计图

3.3　模型分类情况

每种机器学习算法的四个模型为一组测试,共进行四组分类测试。结果如图 2 和图 3。

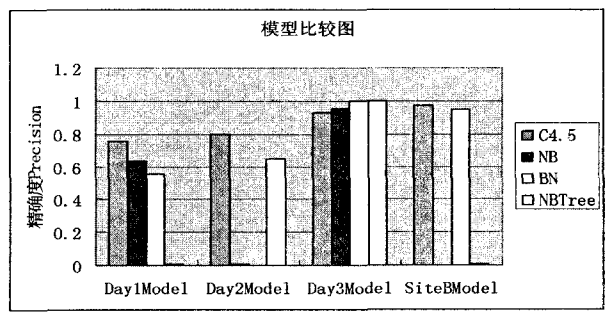

图 2　模型比较图

从模型比较图 2 中可以很清晰看到四种机器学习算法的 Day3 模型都表现很好，SiteB 模型除了两个已经排除的算法外，另外两个都达到了很高的识别率；四个算法的 Day1 模型表现一般，都在 80% 以下，而 Day2 即使排除两个算法(NB,NBTree)也只达到平均 70%。

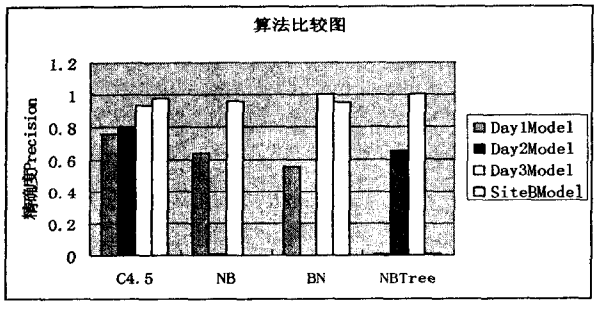

图 3 算法比较图

算法比较图 3 显示出 C4.5 算法在 P2P 流量识别方面的优越性，同时加上其较短的建模时间，所以说 C4.5 算法是很适合 P2P 的流量识别的(至少从本课题比较的这四种算法来说是如此)。其次是贝叶斯网络算法，虽然在一个数据集上表现不佳，但凭借其在其他三个数据集上的良好表现加上它拥有比 C4.5 更快的建模时间(平均提高 58%)，所以还是能肯定这个算法的。

4 结束语

提出了基于机器学习的 P2P 流量检测系统的实现方案，设计实现了基于 FCBF 方法的流量属性方案，并对网络公开数据源使用 4 种流行的机器学习算法进行了离线的分析和比较。在所比较的四种机器学习算法中，决策树算法 C4.5 在各方面的都表现较好。文中只比较了四种分类算法的性能，如前文所述，目前应用到 P2P 流量检测的机器学习算法除了分类算法外，还有无监督的学习算法，其中一个很重要的应用就是聚类分析，所以探讨聚类分析在 P2P 流量检测方面的性能是下一步研究的问题，同时利用另外的属性选择算法选取特征集再结合各种机器学习算法进行类似测试也是很有价值的研究方向。

致谢：英国剑桥大学计算机实验室的 Li Wei 为本文提供了实验数据集，在此表示由衷的感谢。

参考文献：

[1] 吴国庆.对等网络技术研究[J]. 计算机技术与发展,2008,18(7):100－104.

[2] 蒋海明,张剑英,王青青,等.P2P 流量检测与分析[J]. 计算机技术与发展, 2008,18(7):74－76.

[3] McGraw－Hill. 机器学习[M].曾华军,张银奎,等译.北京:机械工业出版社,2003.

[4] 威 滕.数据挖掘:实用机器学习技术[M]. 北京:机械工业出版社,2006.

[5] Zuev D, Moore A. Traffic classification using a statistical approach[J]. Lecture Notes in Computer Science,2005,3431:321－324.

[6] Constantinou F,Mavrommatis P. Identifying Known and Unknown Peer－to－Peer Traffic[C]// Proceedings of Fifth IEEE International Symposium on Network Computing and Applications.[s.l.]:[s.n.], 2006:93－102.

[7] Karagiannis T, Broido A, Faloutsos M, et al. Transport Layer Identification of P2P Traffic[C]//In: Proc. of ACM SIGCOMM IMC. Taormina, Sicily, Italy:[s.n.],2004:121－134.

[8] 吴 敏,王汝传. 基于主机的 P2P 流量检测与控制方案[J]. 计算机技术与发展,2009,19(10):26－30.

[9] 王 锐. P2P 流量检测技术研究[D]. 长沙:国防科学技术大学,2006.

[10] Yu Lei, Liu Huan. Feature selection for high－dimensional data: a fast correlation－based filter solution[C]// in: Proceedings of the 20th International Conference on Machine Learning (ICML'03). Washington, D.C.:[s.n.], 2003:856－863.

[11] Erman J, Mahanti A, Arlitt M, et al. Offline/realtime traffic classification using semi－supervised learning[C]//26th International Symposium on Computer Performance, Modeling, Measurements, and Evaluation.[s.l.]:[s.n.], 2007:1194－1213.

[12] Li Wei, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks,2009,53:790－809.

(上接第 132 页)

[6] 张建辉.K－means 聚类算法研究及应用[D].武汉:武汉理工大学,2007:4－13.

[7] 张建萍,刘希玉.基于聚类分析的 K－means 算法研究及应用[J].计算机应用研究,2007(5):23－24.

[8] 吴 涛,吴 涛,尚 丽,等.一种基于聚类的交叉覆盖算法[J].计算机技术与发展,2008,18:113－116.

[9] 李士勇.蚁群优化算法及其应用研究进展[J].计算机测量与控制,2003(12):45－47.

[10] 吴 斌,傅伟鹏,郑 毅,等.一种基于群体智能的 Web 文档聚类算法[J].计算机研究与发展,2002(11):63－65.

[11] 杨黎刚,苏宏业,张 英,等.基于 SOM 聚类的数据挖掘方法及其应用研究[J].计算机工程与科学,2007(8):40－43.

[12] 袁 方,周志勇,宋 鑫.初始聚类中心优化的 k－means 算法[J].计算机工程,2007(3):76－77.