基于决策树的P2P流量识别方法研究

李晟锴
(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

摘 要:针对新型P2P业务采用净荷加密和伪装端口等方法来逃避检测的问题,提出了一种基于决策树的P2P流量识别方法。该方法将决策树方法应用于网络流量识别领域,以适应网络流量的识别要求。决策树方法通过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。实验结果验证了C4.5决策树算法相比较Naïve Bayes、Bayes Network算法,处理相对简单且计算量不大,具有较高的数据处理效率和分类精度,能够提高网络流量分类精度,更适用于P2P流量识别。

关键词:决策树;流量识别;特征选择;分类精度
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2011)12-0029-04

P2P Network Traffic Classification Based on Decision Tree

LI Sheng-kai
(School of Computer Science & Engineering, Anhui University of Science & Technology, Huainan 232001, China)

Abstract: To solve the question of new P2P application with payload encryption and camouflage to evade detection port, propose P2P network traffic classification based on decision tree. This method applies decision tree into the areas of network traffic to accommodate Internet traffic identification requirements. Decision tree method builds a classification model using information entropy in training data and classifies flows just by a simple search of the decision tree. Compared with Naïve Bayes, Bayes network algorithm, experimental results demonstrate the C4.5 decision tree can achieve high classification accuracy with faster computational time by relatively simple and small calculation processing. It is more suitable to P2P traffic identification.

Key words: decision tree; traffic classification; feature selection; classification accuracy

0 引 言

网络流量的精确分类是分析网络用户行为、检测网络异常行为和提高服务质量等行为的前提和基础[1]。P2P应用的飞速发展,其流量爆发式的增长和不加限制的带宽使用,极大地增加了网络负担,使网络拥塞现象日趋严重[2]。随着P2P各种业务的剧增,以P2P流量为主要对象的网络业务感知已成为目前研究的热点。近期涌现的新型P2P业务采用净荷加密、伪装端口和分块传输等方法来逃避检测识别,增加了P2P业务感知的难度[3]。由于IANA对端口号规定的非强制性和有限性,越来越多的应用采用非规范的端口,有些应用甚至使用动态端口和冒充特定端口的方法来伪装自己,使得端口识别法不再有效[4]。净荷特征识别法表现出了很好的网络流量识别能力,但由于涉及到能够窥视个人隐私的问题而受到质疑,同时对于加密后的数据包,净荷特征识别法也无能为力[5]。

近年来,基于机器学习的网络流量识别方法表现出了较高的准确率,因此得到了越来越多研究者的青睐[6]。文献[7~9]研究神经网络技术应用于网络流量分类,其中文献[7]研究了BP神经网络技术,分析P2P流量的特征,构建BP网络,通过对该网络的足够训练,得到相关的测试结果;文献[8]研究了基于自组织网映射网络的流量分类算法,自组织映射网络算法模拟生物神经元,通过自组织行为对数据进行分类学习,实验表明,该无监督型算法能够对新流量进行自动识别,提高了流量识别的准确率;文献[9]利用神经网络的自学习能力和模糊逻辑的动态性和及时性等特点,将模糊理论和神经网络相互混合,研究在线识别P2P流特征的方法,

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已成为当前网络测量领域内一个新兴的研究热点[10]。在使用机器学习方法处理流量分类的问题时,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

收稿日期:2011-05-12;**修回日期**:2011-08-21
基金项目:安徽省高等学校自然科学基金重点项目(KJ2009A093)
作者简介:李晟锴(1985-),男,硕士研究生,研究方向为计算机网络、人工智能。

(flow)[11]。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[12]。

针对 P2P 业务采用净荷加密、伪装端口和分块传输等方法来逃避检测识别的问题,文中提出一种基于决策树的 P2P 流量识别方法。

1 决策树

决策树学习是应用广泛的数据挖掘算法之一。它是一种逼近离散值函数的方法,对噪声数据有很好的健壮性且能够学习析取表达式。它是以实例为基础的归纳学习算法,通常用来形成分类器和预测模型,着眼于从一组无次序、无规则的事例中推理出决策树表示形成的分类规则。到目前为止决策树有很多实现算法。例如由 Quinlan 提出的 ID3 算法和 C4.5 算法等。

C4.5 算法是 Quinlan JR 在 1993 年提出的,是以 ID3 算法为核心的完整的决策树生成系统。它通过两个步骤来建立决策树:树的生成阶段和树的剪枝阶段。C4.5 采用基于信息增益率(Information Gain Ratio)的方法选择测试属性。信息增益率通过加入一个被称为分裂信息量(Split Information)的项,分裂信息用来衡量属性分裂数据的广度和均匀性:

$$\text{SplitInformation}(S,A) = -\sum_{i=1}^{c}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}$$

其中:S_1 到 S_c 是 c 个值的属性 A 分割 S 而形成的 c 个样例子集。

信息增益率等于信息增益对分裂信息量的比值:

$$\text{GainRatio}(S,A) = \frac{\text{gain}(S,A)}{\text{SplitInformation}(S,A)}$$

2 基于决策树的 P2P 流量识别方法

2.1 基于决策树的 P2P 流量识别方法

网络流量分类是一种典型的多元分类问题,可以抽象为:已知网络流样本集合 S,样本 $X=\{x_1,x_2,\cdots,x_n\}$,其中 $x_1,x_2,\cdots,x_n$ 分别是集合 S 上定义的对应网络流特征 $A=\{A_1,A_2,\cdots,A_n\}$ 的值,类变量 C 的取值范围为 $\{c_j \mid 1\leqslant j\leqslant m, m$ 是正整数$\}$,目的在于利用机器学习算法构建网络流量分类模型 $f:X\rightarrow C$,并根据此模型对类型未知的网络流进行分类。基于机器学习的网络流量识别方法主要是首先捕获大量的网络流量,根据这些流量的各种属性特征,运用合适的算法进行分类,得到流量属于何种应用。如图 1 所示,选择流的属性特征和运用决策树算法实现是两个关键环节。

2.2 属性选择

据相关资料研究表明[13],可以用于流量识别的属性多达 249 项。使用所有流的属性进行学习分类是非常不现实的,因为其中超过 100 项的属性通过傅里叶变换技术得来,如若全部计算,则负载过于沉重。此外,在实际网络环境中大部分的属性与分类的结果关联较小。因此,选择合适的流的属性显得至关重要。

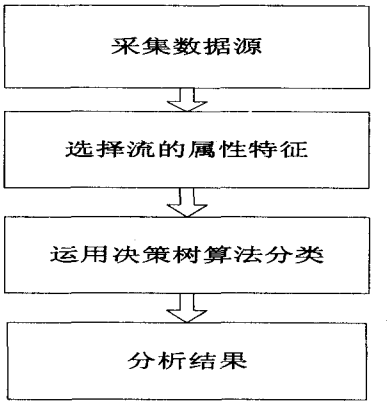

图 1 基于决策树算法 P2P 流量识别方法流程图

经分析,P2P 流的数据包长度变化与其它非 P2P 流有明显的区别,可以将流的数据包长度作为流的属性。这是因为 P2P 文件共享过程可分成信令和数据传输两个过程,信令包一般只有几字节,如建立连接时的 UDP 包较多且包长较短,而 TCP 数据包则以兆字节为单位传输,使得 P2P 流的数据包数量较多,包长变化较大。而非 P2P 流的数据包主要使用 TCP 连接,数据包数量有限,包长大小变化不大。系统从属性易于获取且对分类结果关联性较大的角度出发,选择了流的一些属性以探求潜在价值的信息,如表 1 所示。其中,最后一项 flow_type 为类别属性,指明了业务流所处的应用类型,其余项为流的统计属性。

表 1 选择的流的属性

序号	属性类型	描述
1	tot_tm	流的持续时间
2	tot_pktnum	流的总分组数
3	tot_bytes	流的总字节数
4	avr_pktm	流的平均包长度
5	tot_payload	流的净荷长度
6	flow_type	类别属性

对于流的统计属性,需要计算属性的最大值和最小值,得到平均值和方差值,以利于形成统计特性。其中流的属性的方差值计算较为复杂,计算公式如下:

$\delta^2 = D(p_{\text{flow}}) = E\ (p_{\text{flow}} - E(p_{\text{flow}}))^2$;得 $\delta = \sqrt{E\ (p_{\text{flow}} - E(p_{\text{flow}}))^2}$

考虑到实际存储空间大小及数据处理效率的需要,本系统只对每个流的前 5 个数据包进行属性统计,这种方法称之为 5-flow 检测。经验证,这种方法统计出的结果与对流的每个包进行属性统计所获得的结果非常接近。

3　实验验证

3.1　实验数据集

文中采集实际的校园网出口流量作进一步的试验和验证。该实验数据集分为四组,共包含了 161430 个网络流样本,每组实验数据集都被分成 2 种类别,即 P2P 流和 Non-P2P 流,其中 P2P 流包括 Kazaa,Gnutella,eMule,BitTorrent,Skype 等,Non-P2P 流包括 HTTP,SMTP,POP3,FTP,BIOS,GAME,DNS,Streaming 等。每组数据集的网络流的数量和 P2P 流所占比例见表 2。采用分类精度相对较高的净荷匹配法,为采集的训练与测试数据预先分类并打上标签,作为分类基准,用来评估各分类算法。由于存在分类错误,所以不可避免地存在评估偏差,但并不妨碍实验验证。

表 2　实验数据集统计信息

数据集	P2P 流所占比例%	总流数
D1	32.4	25467
D2	43.1	45534
D3	39.0	32587
D4	36.6	57842

对上述获得的数据集作如下处理:

1)分流:根据五元组(源地址、目的地址、源端口、目的端口和传输层协议)来划分流。对于 TCP 流,由 TCP 的三次握手来识别流的开始,由 TCP 的 FIN/RST 分组作为流结束的标志。此外,若连接持续空闲了 t 秒(取 90s),则假定该流已经结束。同样,对于 UDP 流,若 t 秒(取 90s)时间内没有分组到达,则认为该 UDP 流结束。

2)统计流的相关特征如流持续时间、分组到达时间间隔、分组净荷长度等。

3.2　评估策略

在机器学习的流量分类中,评价指标表示了根据测试数据集的分类结果,预估在未知数据集上处理流量分类的能力。为了评估文中所提出方法的性能,同时便于与其他方法进行对比,文中采用正确肯定率(TP Rate)、错误肯定率(FP Rate)、精确率(Precision)、反馈率(Recall)、F-Measure 和 ROC 域等,分别定义为:

$$\text{TP Rate}=\text{tp}=\frac{TP}{TP+FN}\times 100\%$$

$$\text{FP Rate}=\text{fp}=\frac{FP}{FP+TN}\times 100\%$$

$$\text{Precision}=\frac{TP}{TP+FP}\times 100\%$$

$$\text{Recall}=\text{tp}=\frac{TP}{TP+FN}\times 100\%$$

$$\text{F-Measure}=\frac{2\cdot TP}{2\cdot TP+FP+FN}\times 100\%$$

ROC 域值等于 ROC 曲线下方区域的面积。ROC(Receiver Operating Characteristic,接收者操作特性)曲线是一种评价数据挖掘方案的图形技术,它在不考虑分类分布或误差成本的情况下描绘分类器的性能。

3.3　实验结果与分析

为更好地验证 C4.5 算法是否适用于 P2P 流量识别,实验尝试用 WEKA 平台提供的另外两种常用的机器学习分类方法朴素贝叶斯(Naïve Bayes)算法和贝叶斯网络(Bayes Network)算法进行分类测试,将分类结果的评价指标与 C4.5 算法分类评价指标相比较。

评价一个算法是否适用于流量识别,不仅在于分类器建立的完好,更在于其测试性能的表现。实验对 HTTP、AMULE、BT 三种业务流进行分类识别,由于实验结果相近,下文仅以 BT 流为例,比较三种机器学习的分类器的分类结果。图 2 和图 3 分别显示了三种分类器在训练阶段和测试阶段的分类评价指标差异。

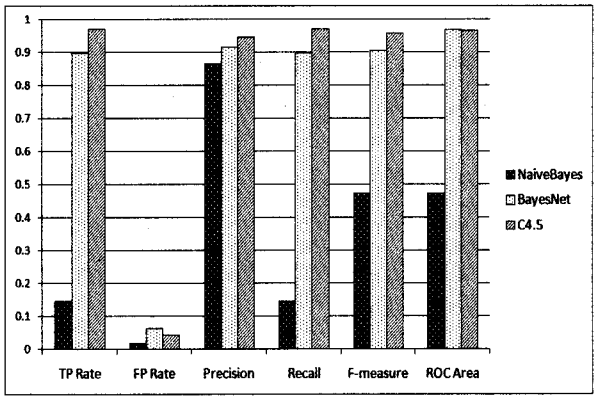

图 2　训练阶段 BT 流分类性能比较

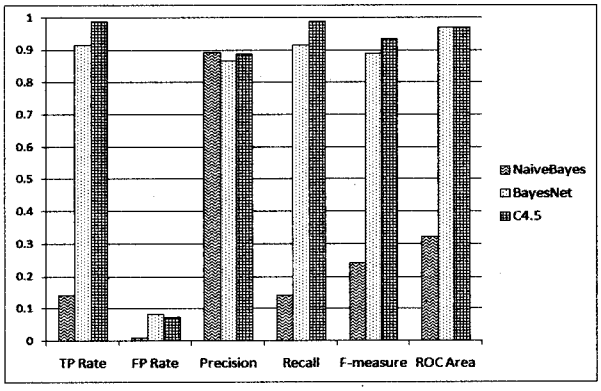

图 3　测试阶段 BT 流分类性能比较

从图 2 和图 3 中可以看出:不论是在训练阶段还是在测试阶段,Naïve Bayes 分类器的各项评价指标均相对较低,尤其是 TP Rate 还不到 0.2,明显不适用于 P2P 流量识别。而 C4.5 与 Bayes Network 两种分类器的各项评价指标均相对较高,比较适用于 P2P 流量识别,其中 Bayes Network 分类器的各项评价指标基本达到 90% 以上,而 C4.5 分类器的评价指标则更高一筹。

为了更全面地分析 C4.5 算法的性能,表 3 从总体

上比较了三种分类器的分类性能。可以看出:分类准确率由高到低依次为 C4.5、Bayes Network、Naïve Bayes,所消耗的分类时间由短到长依次为 Naïve Bayes、Bayes Network、C4.5,因此,三种分类器的总体性能由高到低依次为 C4.5、Bayes Network、Naïve Bayes。虽然 Naïve Bayes 分类器的分类时间消耗最小,但是其准确率仅在 60% 左右,性能明显最差。而 Bayes Network、C4.5 两种分类器以时间消耗为代价,均取得了不错的准确率,其中 Bayes Network 分类器的准确率达到90%左右,C4.5 分类器的准确率更是达到95%左右,且分类测试时所消耗的时间相对更小。由此可见,在三种分类器中,C4.5 算法不论是在建模训练阶段还是在分类测试阶段都具备了最高的分类准确率,虽然其建模过程较为复杂,但是其分类处理过程较为简单,数据处理速度较快,因此,C4.5 算法最为适用于 P2P 流量识别。综上所述,在理论和实验结果上均说明了 C4.5 决策树算法相比较 Naïve Bayes、Bayes Network 算法而言,处理相对简单且计算量不大,具有较高的数据处理效率和分类稳定性,更适用于 P2P 流量识别。

表3 分类器总体性能比较

分类器	准确率(%)		所耗时间(s)	
	建模训练	分类测试	建模训练	分类测试
Naïve Bayes	58.65	63.33	0.02	0.01
Bayes Network	90.625	89.79	0.05	0.03
C4.5	95.91	93.75	0.05	0.02

4 结束语

文中提出了一种基于决策树的 P2P 流量识别方法。该方法利用 C4.5 算法能够综合考虑各个特征属性间的关系对 P2P 网络流量进行识别。实验结果表明,该方法相比于朴素贝叶斯(Naïve Bayes)算法和贝叶斯网络(Bayes Network)算法的分类精度有了一定的提高,具有较好的分类精度。如何减小该方法的计算开销将是下一步研究的重点。

参考文献:

[1] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark[C]//Proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications. New York, NY, USA: ACM, 2005: 229-240.

[2] Xu Ke, Zhang Ming, Ye Mingjiang. Identify P2P traffic by inspecting data transfer behavior[J]. Computer Communications, 2010, 33(10): 1141-1150.

[3] Yuan Ruixi, Li Zhu, Guan Xiaohong, et al. An SVM-based machine learning method for accurate Internet traffic classification[J]. Information Systems Frontiers, 2010, 12(2): 149-156.

[4] Horng Shi-Jinn, Su Mingyang, Chen Yuan-Hsin. A novel intrusion detection system based on hierarchical clustering and support vector machines[J]. Expert Systems with Applications, 2011, 38(1): 306-313.

[5] Zainab Z, Sara H, Bjorn L, et al. Real-time detection of traffic anomalies in wireless mesh networks[J]. Wireless Networks, 2010, 16(6): 1675-1689.

[6] Huang Shijun, Chen Kai, Liu Chao. A statistical-feature-based approach to internet traffic classification using machine learning[C]//2009 International Conference on Ultra Modern Telecommunications and Workshops. [s.l.]: [s.n.], 2009.

[7] 沈富可,常 潘,任肖丽. 基于 BP 神经网络的 P2P 流量识别研究[J]. 计算机应用,2007,27(12): 44-45.

[8] 潘亚东,周 健,孙海霞. 基于自组织映射网络的流量分类算法[J]. 合肥工业大学学报:自然科学版,2009,32(8): 1142-1145.

[9] Nogueira A, Salvador P, Couto A, et al. Towards the On-line Identification of Peer-to-peer Flow Patterns[J]. Journal of Networks, 2009, 4(2): 108-118.

[10] Alice E, Francesco G, Luca S. Support Vector Machines for TCP Traffic Classification[J]. Computer Networks, 2009, 53(14): 2476-2490.

[11] Constantinou F, Mavrommantis P. Identifying known and unknown peer-to-peer traffic[C]//IEEE NCA'06 Conference. [s.l.]: [s.n.], 2006: 93-102.

[12] Yu Jaehak, Lee Hansung, Kim Myung-Sup, et al. Traffic flooding attack detection with SNMP MIB using SVM[J]. Computer Communications, 2008, 31(17): 4212-4219.

[13] Manuel C, Francesco G, Luca S. Optimizing statistical classifiers of network traffic[C]//Proceedings of the 6th International Wireless Communications and Mobile Computing Conference. [s.l.]: [s.n.], 2010: 758-763.

(上接第28页)

Plants[M]. New York: Spinger-Velag, 1990.

[9] 韩向峰,刘希玉. 基于 L 系统的三维分形图的生成算法[J]. 计算机应用,2004,24(10): 86-91.

[10] 王春华,杨克俭,韩 栋. 基于分枝类型和空间点的三维树木建模方法[J]. 计算机应用研究,2009,26(4): 1592-1597.

[11] Chen Peng, Li Xiang. Imitation of Plants Inflorescence Based on Fusion of L System and IFS[C]//2009 WRI World Congress on Computer Science and Information Engineering. [s.l.]: [s.n.], 2009: 647-650.

[12] Prusinkiewicz P, Hmmel M, Hanan J, et al. Visual models of plant development[M]//Hand-book of Formal Languages. [s.l.]: Springer-Verlag, 1996.