软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(6):1500−1516 [doi: 10.3724/SP.J.1001.2012.04074]

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

分类不平衡协议流的机器学习算法评估与比较*

张宏莉, 鲁 刚+

(哈尔滨工业大学 计算机科学与技术学院 计算机网络与信息安全技术研究中心,黑龙江 哈尔滨 150001)

Machine Learning Algorithms for Classifying the Imbalanced Protocol Flows: Evaluation and Comparison

ZHANG Hong-Li, LU Gang+

(Computer Network and Information Security Technology Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: lgang198202@126.com

Zhang HL, Lu G. Machine learning algorithms for classifying the imbalanced protocol flows: Evaluation and comparison. *Journal of Software*, 2012,23(6):1500−1516. http://www.jos.org.cn/1000-9825/4074.htm

Abstract: In the case of the imbalanced protocol flows, the changes of flow distribution have a huge impact on the accuracy and stability of traffic classifiers that use machine learning algorithms. It is very important to select a suitable machine learning algorithm to classify the imbalanced protocol flows on line. By means of single-factor experiment design, this paper verifies that it is possible for C4.5 decision tree, Naïve Bayes with kernel density estimation (NBK) and support vector machine (SVM) to classify traffic with the first four packets of the TCP connection. After comparing the performances of the three classifiers abovementioned, the study finds that the testing time of C4.5 decision tree is the shortest and SVM is the most stable. Finally, Bagging algorithm is applied to classify traffic. The experimental results show that, the stability of Bagging is similar to SVM and the testing time and modeling time of Bagging is close to C4.5 decision tree. Therefore, Bagging classifier is the most suitable to classify traffic on line.

Key words: imbalance; feature selection; traffic classification; ensemble learning; single-factor experiment

摘 要: 网络协议流不平衡环境下,流样本分布的变化对基于机器学习的流量分类器准确性及稳定性有较大的影响.选择合适的机器学习算法以适应网络协议流不平衡环境下的在线流量分类,显得格外重要.为此,首先通过单因子实验设计,验证了 C4.5 决策树、贝叶斯核估计(NBK)和支持向量机(SVM)这 3 种分类算法统计 TCP 连接开始的前 4 个数据包足以分类流量.接着,比较了上述 3 种分类算法的性能,发现 C4.5 决策树的测试时间最短,SVM 分类算法最稳定.然后,将 Bagging 算法应用到流量分类中.实验结果表明,Bagging 分类算法的稳定性与 SVM 相似,且测试时间与建模时间接近于 C4.5 决策树,因此更适于在线分类流量.

关键词: 不平衡;特征选择;流量分类;集成学习;单因子实验

中图法分类号: TP181 **文献标识码**: A

* 基金项目: 国家自然科学基金(60903166); 国家重点基础研究发展计划(973)(2007CB311101, 2011CB302605); 国家高技术研究发展计划(863)(2010AA012504, 2011AA010705)

收稿时间: 2010-06-24; 定稿时间: 2011-06-20

准确的网络流量分类是确保网络安全的关键,有助于保护网络资源、强化机构策略,例如带宽限制、网络计费、恶意流量检测等等.早期使用固定端口号识别流量,但目前许多 P2P 应用常使用随机端口号.Mandukar 等人[1]发现,端口识别技术已不能识别出 30%~70%的互联网流量.目前,绝大多数机构常采用深度数据包检测(deep packet inspection,简称 DPI)技术分类流量.DPI 技术利用协议分析和还原技术提取协议载荷特征,通过模式匹配算法搜索载荷特征识别流量.DPI 技术准确性高且可靠性好,但主要缺点在于其触犯用户隐私,且对协议负载加密的流量识别能力有限.为此,Karagiannis 等人[2]开发了基于应用行为的流量分类系统 BLINC.BLINC 系统关注于主机的会话模式并使用图加以描述,其流量分类准确率可达 90%以上.但 BLINC 系统只能粗粒度地识别 P2P 流量,不能够将 P2P 流量细分到具体的协议.此外,Kim 等人[3]指出,BLINC 流量分类系统适合于部署到单宿主边缘网络(single-homed edge network)的边界连接处,而不适合于部署到骨干节点.基于机器学习的流量分类技术也不依赖于应用层负载,它分类的对象是流,流在本文里被定义为使用相同的五元组模式〈源 IP,目的 IP,源端口,目的端口,传输层协议〉进行通信的双向数据包集合,即通信双方的一次会话.基于机器学习的流量分类技术需要统计数据包层(packet-level)和数据流层(flow-level)信息,例如包大小、包到达时间间隔和流的大小等等.相比于基于应用行为的流量分类技术,机器学习流量分类技术识别的粒度更细.目前,基于机器学习的流量分类技术常作为 DPI 技术的辅助部分,流量分类工具 Tstat2.0[4]就结合使用 DPI 技术和机器学习技术.但 Tstat2.0 仅利用文献[5]提出的机器学习算法识别 skype 流量,还不能够利用该方法分类网络中所有流量.其主要原因在于,基于机器学习的流量分类器受网络环境的影响较大,网络流样本的统计分布发生变化、网络延迟和拥塞等因素都会对基于机器学习的流量分类器产生影响.本文结合实际的网络环境,比较分析基于监督学习的流量分类器的性能,尤其关注以下两点:(1) 网络流样本分布的动态变化对基于机器学习的流量分类器性能的影响;(2) 网络流样本分布的不平衡性对基于机器学习的流量分类器性能的影响.流样本分布的不平衡性在这里是指某一种网络协议的流样本数远远超过其他网络协议.

与以往研究工作不同的是,本文的实验数据集体现出协议流样本分布不平衡性,即数据集中 http 协议类别的样本数远远大于其他类别的样本数.本文的研究贡献在于以下几点.

(1) 尝试细粒度地将流量按照不同的协议进行区分,而不是粗粒度地将网络流量分类成 BULK,P2P 和 Services 等等.之所以细粒度地识别,是因为同一种应用可以运行多个协议.例如,emule 应用即支持 eDonkey 协议下载方式,又支持 HTTP 协议下载方式.而不同协议流的统计特征不尽相同;

(2) 根据 C4.5 决策树、SVM 和 NBK 这 3 种不同分类算法的准确性,本文进行了单因子方差分析实验,以确定统计 TCP 流开始时合适的数据包数目.统计分析 TCP 流开始的前若干个数据包来决定整个 TCP 流的类别,这可以提高流量分类器的分类速度,为机器学习流量分类器在线分类提供依据;

(3) 本文在协议流样本分布不平衡数据集上,从流的准确性、字节的准确性、召回率和精度这 4 个指标详细地比较分析 C4.5 决策树、NBK 和 SVM 这 3 种分类算法的性能,分析了协议流样本分布的不平衡性对 3 种分类算法性能的影响;

(4) 本文将 Bagging 集成学习算法应用到流量分类中,解决了 C4.5 决策树分类不稳定问题.本文实验验证了 Bagging 集成学习分类算法更适合于在线分类流量.

本文第 1 节阐述机器学习流量分类的相关研究.第 2 节分析流量分类算法的原理.第 3 节介绍实验环境,给出实验数据集和特征选择的方法.第 4 节比较分析 C4.5 决策树、NBK 和 SVM 这 3 种分类算法的实验结果.第 5 节介绍基于 Bagging 集成学习的流量分类器.第 6 节总结全文.

1 相关研究

近年来,国内外关于机器学习的流量分类技术已经有大量的研究,这些研究的侧重点主要在提高分类器的准确性、实时性以及提取和选择有效的网络流特征上.Moore 等人[6]利用 NBK 算法分类流量,将网络流粗粒度地分成 10 种不同业务,分类准确性为 95%左右.Li 等人[7]在 Moore 工作的基础上,采用基于相关的快速过滤算法(fast correlation-based filter,简称 FCBF)进行特征选择,选择了 12 种时空稳定的特征.Este 等人[8]利用互信息量比

较网络流特征的时空稳定性,发现 TCP 连接建立后的第一个数据包大小所含的信息量最大,并在不同的时间和地点采集的数据上进行测试,验证了数据包大小是最稳定的特征.Bernaille 等人[9]尝试统计 TCP 流开始的前 5 个数据包大小,使用 *K* 均值聚类算法分类流量.Soysal 等人[10]评估 Bayes 网络、决策树和多层感知机(multilayer perceptrons)这 3 种机器学习算法,发现 Bayes 网络和决策树更适合于高速分类网络流.Pietrzyk 等人[11]在不同地点采集的数据集上,用 NBK 算法、Bayes 网络和 C4.5 决策树这 3 种机器学习算法进行地点交叉(cross-site)测试,发现训练数据集和测试数据集都是同一地点采集时,3 种分类算法都表现出最好的性能;而训练数据集和测试数据集都是在不同地点采集时,3 种分类算法的性能都有所下降.他们认为,造成这种现象的原因是分类算法对训练数据的过拟合(data overfitting).他们同时也发现,相比于其他两种分类算法,C4.5 决策树的分类性能最好.国内关于流量分类的研究中,徐鹏等人[12]也发现 C4.5 决策树分类算法要优于 Naïve Bayes 分类算法和 NBK 分类算法.文献[10−12]仅仅比较了流量分类的流准确性,并没有给出分类的字节准确性,Erman 等人[13]发现,0.1%的大象流占据了整个流量传输字节数的 46%.如果流量分类器未识别出这 0.1%的大象流而得到 99.9%的流准确性,那么流量分类器将损失 46%的字节准确性.表 1 比较分析了目前国内外基于机器学习的流量分类技术.

Table 1 Comparison of traffic classification technologies

表 1 流量分类技术的比较

文献	机器学习算法	流量特征	流量类别	分类粒度	是否考虑字节准确性
Moore 等人[6]	Naïve Bayes 和 NBK 分类算法	从完整数据流中提取 248 种流量特征,包括流持续时间、TCP 端口、包到达时间间隔的统计信息、负载大小的统计信息等	BULK,Database P2P,Mail,Services 等	粗粒度	否
Li 等人[7]	C4.5 决策树和 Naïve Bayes	TCP 端口、TCP 标识位的统计信息、TCP 分段大小的统计信息	与文献[6]相同	粗粒度	是
Bernaille 等人[9]	*K* 均值聚类	TCP 流开始的前 5 个数据包大小和包方向	eDonkey,ftp,http, kazaa 等	细粒度	否
Soysal 等人[10]	Bayes 网络、决策树和多层感知机	完整的 TCP 流统计特征	P2P,Web,Bulk 等	粗粒度	否
Pietrzyk 等人[11]	NBK,Bayes 网络和 C4.5	数据包层面:数据包大小和数据包传输方向 数据流层面:TCP 端口和标识位的统计信息	Web,eDonkey, Mail,Chat,Gnutella 等	粗粒度	否
徐鹏等人[12]	C4.5 决策树	与网络流相关的 34 个统计特征	Web,BT,BULK 等	粗粒度	否
徐鹏等人[14]	SVM	与文献[6]相同	与文献[6]相同	粗粒度	否

由表 1 可见,目前基于机器学习的流量分类技术要么粗粒度地分类流量,要么忽略了分类的字节准确性.此外,文献[9,15]的流量分类工作没有考虑识别流样本数小于 500 的网络应用.然而我们的实验结果表明,即使网络应用传输的数据流非常少,但他们传输的字节数却很大,丢掉这些数据流将会导致分类器的字节准确性下降(详见第 3.2 节).与以往工作相比,本文按照不同协议细粒度地分类流量.从在线分类流量的角度考虑,在本文的实验数据集上通过单因子方差实验,验证了 C4.5 决策树、NBK 和 SVM 统计 TCP 连接开始的前 4 个数据包足以分类流量.本文还从字节准确性、流准确性、各协议类别的召回率和精度这 4 个指标,详细地比较上述 3 种分类算法的性能,分析了协议流不平衡性对这 3 种分类算法性能的影响.最后,本文将 Bagging 集成学习算法应用到流量分类中,并在协议流不平衡数据集上与上述 3 种分类算法做了实验比较分析.

2 流量分类算法

本节简要介绍 NBK、C4.5 决策树和支持向量机 SVM 这 3 种分类算法.

2.1 NBK分类算法

本节先介绍 Naïve Bayes 算法,因为 NBK 分类算法是 Naïve Bayes 算法的一般形式.

假定有 n 个类别 $C_i(i=1,\ldots,n)$,任意网络流样本 y 属于类别 C_i 的概率为

$$P(C_i \mid y)=\frac{f(y \mid C_i)P(C_i)}{P(y)}=\frac{f(y \mid C_i)P(C_i)}{\sum_{j=1}^{n} f(y \mid C_j)P(C_j)} \tag{1}$$

我们将流样本 y 指派到使得 $P(C_i|y)(i=1,\ldots,n)$最大的类别中,其中,$f(.|C_j)$为概率密度函数,Naïve Bayes 算法假设它服从正态分布,但实际上总体的分布常常不是正态分布.NBK 分类算法采用核函数来拟合总体分布,如公式(2)所示.

$$f(t \mid C_j)=\frac{1}{n_{c_j} h}\sum_{x_i:C(x_i)=c_j} K\left(\frac{t-x_i}{h}\right) \tag{2}$$

这里,h 是核带宽(kernel bandwidth),K 是核函数.核函数的值非负且满足 $\int_{-\infty}^{+\infty} K(x)\mathrm{d}x=1$.由于高斯函数有很好的平滑特性,NBK 分类算法常采用高斯函数作为核函数.从公式(1)、公式(2)可以看出,NBK 分类算法分类结果依赖于训练样本集的分布状况.

2.2 C4.5决策树分类算法

C4.5 算法采用自顶向下递归分治的方式构造分类模型,其构造的分类模型是一种树的结构,每个分裂节点(非树叶节点)表示一个属性上的测试,每个分枝代表一个测试输出,而每个树叶节点存放一个类标号.决策树构建的原则就是使每次划分后的不确定性尽可能地小.C4.5 算法采用信息增益率作为属性选择的度量,每次划分都是选择信息增益率最大的属性作为分裂节点.假定有 n 个类别 $C_i(i=1,\ldots,n)$,D 为网络流的训练样本集.在训练集 D 中,分类所需的期望信息(即 D 的熵)为

$$H(D)=-\sum_{i=1}^{n} p_i \log_2 p_i \tag{3}$$

其中,p_i 为训练集 D 中的流样本属于类别 C_i 的概率.假定选择属性 A 作为分裂节点,属性 A 根据训练数据集的测试有 m 个不同的输出,将 D 划分为 m 个子集 $D_j(j=1,2,\ldots,m)$.划分后的训练集再分类所需的信息量为

$$H_A(D)=\sum_{j=1}^{m}\frac{|D_j|}{|D|}\times H(D_j) \tag{4}$$

于是,由公式(3)、公式(4)得到此次划分所获得的信息增益

$$Gain(A)=H(D)-H_A(D) \tag{5}$$

C4.5 算法使用分裂信息值将信息增益规范化.分裂信息定义为

$$SplitH_A(D)=-\sum_{j=1}^{m}\frac{|D_j|}{|D|}\times\log_2\frac{|D_j|}{|D|} \tag{6}$$

因此,信息增益率可定义为

$$gainratio(A)=\frac{Gain(A)}{SplitH(A)} \tag{7}$$

由公式(3)~公式(7)可见,C4.5 决策树算法实际上采用了熵最小选择策略.该算法每选择一次分裂节点就要计算信息增益率,而信息增益率的计算仍旧依赖于训练样本集中各类别的概率分布.

C4.5 算法采用剪枝法解决训练数据过拟合问题.剪枝后的分类树更小,复杂度更低,因此在实际网络流分类时,剪枝后的分类树分类速度更快、更好.

2.3 SVM分类算法

SVM 算法将实际问题通过非线性变换到高维的特征空间,并在高维空间中构造线性判别函数来实现分类.在构建 SVM 分类模型时,首先选择多项式、高斯函数等核函数将输入数据映射到高维空间,然后在高维空间中

搜索线性最佳分类超平面.分类超平面可描述为

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{k}w_i^*y_iK(x_i,x)+b^*\right) \tag{8}$$

其中,w^*和 b^*是超平面的参数,$K(x_i,x)$是核函数,$y_i\in\{-1,1\}$是类别标记.本文选择高斯径向基函数作为核函数,x_i 是第 i 个训练样本,而 x 是测试样本.由公式(8)可见,SVM 算法根据 $f(x)$的符号进行分类,即 $f(x)>0$ 的样本归为一类,而 $f(x)<0$ 的样本归为另一类.当要分 M 个类别时,SVM 算法将多类分类问题转换为成对分类,即建立 $M\times(M-1)/2$ 个二分类器.当判别一个测试样本 x 属于 M 类中的哪一类时,我们考虑所有二分类器对 x 的分类结果.即一个二分类器判别 x 属于第 i 类时,就意味着第 i 类获得 1 票,票数最多的类别就是最终判定 x 所属的类别.

SVM 算法不依赖于网络流样本的先验概率,它通过将流量分类问题转换为二次寻优以保证经验风险和真实风险最小,进而防止过拟合现象的发生.

通过对上述 3 种分类算法原理的分析可以看出,NBK 分类算法最依赖于网络流样本的概率分布,而 SVM 最不依赖于网络流样本的概率分布,即网络协议流样本分布的动态变化对 SVM 的影响最小.

3 实验环境

3.1 数据集

在 2010 年 3 月和 4 月不同的时间段内,我们在哈尔滨工业大学某实验室出口,用 tcpdump 工具采集 6 次流量,每次采集都包含数据包负载的完整信息.实验环境采用 100Mbps 的以太网,大约有 80 台主机.数据集的详细情况见表 2.由表 2 可见,本文已经采集了实验室所有工作时间的流量.

Table 2 Description of the data sets

表 2 数据集描述

数据集名	采集时间	数据集大小
dump1	3 月 23 日 15:00~16:00	3.28G
dump2	3 月 25 日 16:00~17:40	2.82G
dump3	4 月 8 日 10:00~11:40	241M
dump4	4 月 8 日 13:20~17:40	11G
dump5	4 月 13 日 9:10~10:10	2.02G
dump6	4 月 14 日 18:30~20:30	0.99G

3.2 流量标注

目前,流量标注的方法主要有两种:一种是利用基于负载和端口匹配的流量分类工具标注流量;另一种是在可控的网络环境下标注流量,例如单独搭建 web 服务器、FTP 服务器、emule 服务器等,并在每台服务器上抓取特定应用的流量.对于粗粒度地分类流量,可以采用第 2 种方案.但本文的研究内容在于细粒度地识别协议,由于同一种应用可以运行多个协议,例如,emule 1.1.8 版本即支持 eDonkey 协议下载方式,又支持 HTTP 协议的下载方式,如果在服务器上只运行 emule 应用,我们也只能标注应用而无法标注协议.因此,我们采用第 1 种方案标注流量.

目前,基于负载匹配的流量分类工具 L7-filter[16]应用较为广泛,但它并不能准确识别所有协议,文献[17]给出目前 L7-filter 对不同协议的识别程度.本文选择 L7-filter 识别准确的 9 种常用协议,分别为 HTTP,eDonkey,xunlei(迅雷),ssl,ftp,pop3,qq,ssh 和 smtp.由于 TCP 报文在采集的数据集中占了大部分,本文仅分类 TCP 数据流.流量标注后每个数据集的样本数见表 3.各协议类别样本数所占比例如图 1 所示.

由图 1 可见,http 协议流样本数占据了所有类别样本数的大部分.各种协议流量所占的字节比例如图 2 所示.

对比图 1 和图 2,我们发现对于 dump4、dump5 数据集,http 流比例虽然很大,但字节比例却不是最大.dump4 数据集中,ssh 协议流的字节比例占总字节数的 60%左右;而 dump5 数据集中,eDonkey 协议流的字节比例达到 78%左右.也就是说,dump4 和 dump5 数据集中,eDonkey 和 ssh 协议的数据流虽然少,但很可能包含大流.可见,

即使流样本数很少的协议,我们也不能忽略对它们的识别.

Table 3 Number of samples in each data set

表 3 每个数据集中的样本数

数据集名	http	eDonkey	xunlei	ssl	ftp	pop3	qq	ssh	smtp
dump1	3 667	81	380	74	19	31	78	0	3
dump2	5 570	147	162	188	28	54	120	1	3
dump3	3 057	5	96	71	0	0	94	1	6
dump4	3 586	71	129	81	1	0	20	2	1
dump5	1 930	67	125	67	54	0	21	0	0
dump6	4 092	16	381	31	0	0	158	0	0

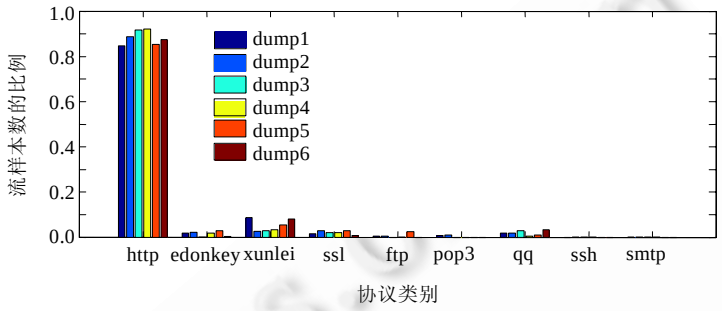


Fig.1 Proportion of different protocol samples in each data set

图 1 每个数据集中各协议样本数所占比例

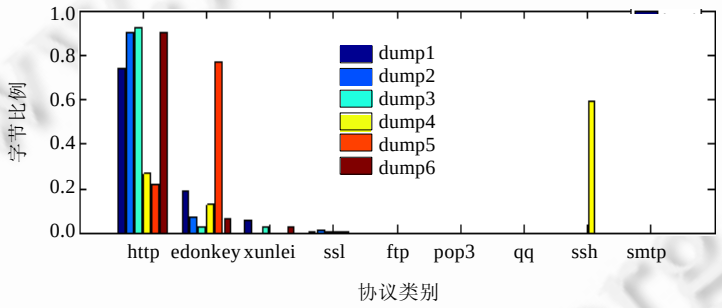


Fig.2 Byte proportion of different protocol samples in each data set

图 2 每个数据集中各协议流所占的字节比例

3.3 选定统计流建立开始的数据包数目

为使基于机器学习的流量分类技术能够在线快速地分类,文献[7,9,15]已经采用提取 TCP 连接的前 5 个数据包的统计特征进行分类.与本文研究的分类算法不同,文献[9,15]使用无监督的 K 均值聚类算法分类流量.我们认为,不同的分类算法由于统计原理不同,所需统计的数据包数目也应该有所不同.文献[7]虽然验证了 C4.5 决策树应该统计流开始的前 5 个数据包,但是他们只是对选定不同数据包时的分类准确性做了简单的比较分析,没有用统计的方法进行证明,这极易产生误差.本节通过单因子方差实验方式,验证 3 种分类算法需要统计流建立开始的数据包数目.

本节针对不同的分类算法,对 TCP 连接开始的前 4~10 个包(包括 TCP 连接 3 次握手的数据包)分别进行测试.通过测试后的准确性,决定选择统计分析合适的数据包数目.本文提取了与流相关的 25 个特征进行分类,这 25 个特征包括包大小和包到达时间间隔的统计信息,以及与 TCP 标志位相关的统计信息等.基于第 3.2 节所描述的 6 个数据集,我们在每个数据集上采用十交叉验证方法,对 SVM,NBK 和 C4.5 决策树这 3 种分类算法进行

测试分析.机器学习算法在选定不同的分类参数时,分类准确性会不同.参数的选择问题已经超出了本文的研究范围.我们采用暴力尝试的方法选择最优参数,对于本文的实验数据集,不同分类器的最优参数见表 4.对于流开始的前 4~10 个数据包,3 种分类算法在十交叉验证后的结果见表 5.

Table 4 Optimal parameters of the three classifiers

表 4 3 种分类器的最优参数

分类器	参数
SVM	*C*=0.8, *gama*=1.0517578125E−5
Naïve Bayes	—
C4.5	*C*=0.25, *M*=2

Table 5 Testing results of each algorithm (%)

表 5 每种算法的测试结果(%)

	包个数	数据集					
		dump1	dump2	dump3	dump4	dump5	dump6
NBK 分类算法测试结果	4 个数据包	80.064 7	91.45	93.992 2	87.863 2	84.444 4	86.479 5
	5 个数据包	67.620 6	61.406	75.285 3	64.096 6	84.986 7	85.839 8
	6 个数据包	83.129 5	85.142 7	88.228 2	87.304	88.053 1	87.580 3
	7 个数据包	83.775 7	86.067 3	93.093 1	87.124 1	85.821 6	88.371 1
	8 个数据包	85.298 9	88.362 8	94.714 7	87.792 3	89.27	82.649 6
	9 个数据包	84.590 5	89.368 8	95.855 9	87.612 4	89.499 1	82.564 1
	10 个数据包	83.875 4	89.687 6	95.375 4	87.406 8	89.191 6	82.542 7
C4.5 决策树分类算法测试结果	4 个数据包	94.063 3	96.506 6	97.086 2	96.940 1	94.222 2	95.671 7
	5 个数据包	95.130 4	96.556 7	96.666 7	96.710 4	93.179 8	95.479 9
	6 个数据包	94.922 7	96.843 6	96.816 8	96.736 1	92.345 1	95.331 9
	7 个数据包	95.176 6	97.226 2	96.816 8	96.838 9	93.286 2	95.386 2
	8 个数据包	95.338 1	97.019	97.057 1	96.247 8	93.271 8	95.320 5
	9 个数据包	95.201 8	96.923 8	97.147 1	96.581 9	93.189 8	94.914 5
	10 个数据包	95.017 3	96.907 9	96.967	96.581 9	92.662 6	94.939 5
SVM 分类算法测试结果	4 个数据包	90.459 7	93.826 8	95.554 2	96.425 8	92.977 8	94.407 5
	5 个数据包	90.63	93.798 8	95.135 1	96.299 2	93.489 8	94.881 58
	6 个数据包	92.291 7	95.153 8	95.015 1	95.887 9	92.787 6	95.010 7
	7 个数据包	91.899 4	95.074 1	95.015 1	95.887 9	93.595 4	95.040 6
	8 个数据包	91.437 8	95.090 1	94.804 8	96.093 5	93.447 7	95.192 3
	9 个数据包	91.049 6	95.154 6	94.714 7	96.093 5	93.277 7	94.935 9
	10 个数据包	90.865 1	95.106 8	94.594 6	96.119 2	93.101 9	94.679 5

本节采用单因子方差分析方法验证统计不同数据包个数是否影响 3 种分类算法的分类准确性.这里的单因子是指数据包个数,它有 7 个不同的取值,即 7 个水平.方差分析有助于检验各个因子水平上的均值是否相等.我们在每种分类算法上分别进行单因子方差分析实验设计,具体过程为:假设统计不同数据包个数时分类算法的准确性并无明显差异,即 H_0:$\mu_1=\mu_2=\mu_3=\ldots=\mu_7$,其备择假设为 H_1:部分均值不等.检验统计量 F_0 服从自由度为 $a-1$ 与 $N-a$ 的 F 分布.a 为因子水平数,这里的 a 值为 7.N 指观察数,N 值为 42.本文将置信水平 α 设置为 0.05.于是,当 $F_0>F_{\alpha,a-1,N-a}$ 时,应该拒绝 H_0,接受 H_1.代入相应的参数值,得到 $F_{\alpha,a-1,N-a}=F_{0.05,6,35}\approx 2.34$.基于表 5 中的数据,我们用 Excel 的数据分析工具得到 F_0 的值.发现 C4.5 决策树和 SVM 分类算法的 F_0 分别为 0.032 和 0.065,都小于 $F_{0.05,6,35}$,故可接受 H_0.即对于 C4.5 决策树和 SVM 分类算法在统计流开始的不同数据包个数时,分类准确性并无显著性差异.随着统计的数据包个数增加,提取相应统计特征的时间开销和空间开销也会有所增加.因此,在本文实验数据集上,对于 C4.5 决策树和 SVM 分类算法而言,选定统计流开始的前 4 个数据包是最合适的.

NBK 分类算法的 F_0 值为 6.006 大于 $F_{0.05,6,35}$,故拒绝 H_0,接受 H_1,即 NBK 分类算法在选定不同数据包个数时,分类准确性有显著性差异.由图 3 可见,NBK 分类算法的平均分类准确性波动幅度较大,尤其在选定前 5 个数据包数目时,分类的平均准确性达到最低.而随着统计分析 6,7,8,9 个数据包,其准确性逐渐升高,在第 9 个数据包时,分类准确性到达最高.但是,统计前 9 个数据包的平均分类准确性,并未比前 4 个数据包的平均分类准确性高

出很多.我们用 t 检验的方法来验证它们之间是否存在显著性差异.

假设统计前 9 个数据包和前 4 个数据包的平均分类准确性并无明显差异,即 H_0:$\mu_4-\mu_9=0$,其备择假设为 H_1:$\mu_4-\mu_9\neq 0$.检验统计量 t_0 服从自由度为 $a-1$ 的 t 分布.这是一个双边假设检验,当 $t_0>t_{\alpha/2}$ 或者 $t_0<-t_{\alpha/2}$ 时,应该拒绝 H_0,接受 H_1.代入相应的参数值,得到 $t_{0.025}$ 为 2.447.我们用 Excel 的数据分析工具得到 t_0 的值为 −0.31,即 $t_0>-2.447$,故接受 H_0.由此得出,NBK 分类算法统计 TCP 连接开始的前 9 个数据包和前 4 个数据包的平均分类准确性并无显著性差异.因此,在本文实验数据集上,对于 NBK 分类算法,选定统计流开始的前 4 个数据包特征是最合适的.

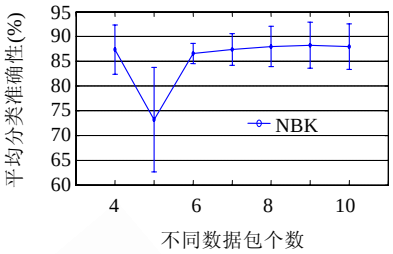


Fig.3 Average accuracy of NBK algorithm

图 3 NBK 分类算法的平均准确性

3.4 特征选择

特征选择是从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的,从而降低计算开销,提高分类效率.本节采用条件熵的方法来度量特征和类别之间的相关性.

设 X 是特征集,Y 是类别集,X 和 Y 之间的相关性计算为

$$H(Y)=-\sum_{y\in R_y}p(y)\log(p(y)),$$

$$H(Y\mid X)=-\sum_{x\in R_x}p(x)\sum_{y\in R_y}p(y\mid x)\log(p(y\mid x)),$$

其中,R_x,R_y 分别是 x 和 y 的取值范围

$$C(Y\mid X)=\frac{H(Y)-H(Y\mid X)}{H(Y)}.$$

$H(Y)$是在数据集上未使用特征集 X 时,类别集 Y 的熵值,而 $H(Y|X)$是在数据集上使用特征集 X 时,类别集 Y 的熵值.$C(Y|X)$是不确定性系数(uncertainty coefficient),用于度量 X 和 Y 之间的相关性.$C(Y|X)$的取值范围在[0,1]之间,当值为 0 时,表示 X 和 Y 没有任何关系;当值为 1 时,表示特征集 X 可以完全确定 Y.每确定一个特征子集 X,就会得到相应的 $C(Y|X)$,本文采用贪婪算法在特征子集空间中进行搜索,选择 $C(Y|X)$值最大的特征子集作为特征选择结果.我们发现,在 dump1~dump6 不同的数据集上,特征选择的结果也不同.这就说明在同一个地点不同的时间内,特征与类别的相关性发生变化.本节取这 6 个测试数据集上特征选择结果的并集作为最后分类的特征.我们从 25 个特征中选择了 8 个最有效的特征,提取相应特征所花费的时间开销和空间开销列于表 6 中,其中,n 为统计的数据包数目.

Table 6 Results of feature selection

表 6 特征选择的结果

特征	描述	空间复杂度	时间复杂度
push_pkts_serv	从服务器到客户端方向,TCP 头中设置 push 位的数据包数	$O(1)$	$O(n)$
init_win_bytes_clnt	从客户端到服务器方向,以初始窗口发送的 TCP 负载大小和	$O(1)$	$O(n)$
init_win_bytes_serv	从服务器到客户端方向,以初始窗口发送的 TCP 负载大小和	$O(1)$	$O(n)$
avg_seg_size_serv	从服务器到客户端方向数据包平均负载大小	$O(1)$	$O(n)$
act_data_pkt_clnt	从客户端到服务器方向,TCP 数据包负载 至少 1 字节的数据包数	$O(1)$	$O(n)$
serv_port	服务器端口	$O(1)$	$O(1)$
second_plength_clnt	三次握手建立后,从客户端到服务器方向第 2 个数据包大小	$O(1)$	$O(1)$
fir_sec_diff_clnt	三次握手建立成功后,从客户端到服务器的第 1 个数据包大小 与三次握手建立时从客户端到服务器方向的 ACK 包大小的差值	$O(1)$	$O(1)$

4 分类算法的比较

我们用 dump1 作为训练数据集建立分类模型,并用剩余的 5 个数据集作为测试.之所以选择 dump1 作为训

练集,原因在于:(1) dump1~dump6 这 6 个数据集是按照时间先后顺序获得的,用 dump1 作为训练集,剩余 5 个作为测试集,为的是分析不同分类器的分类准确性是否随时间变化;(2) dump1 数据集中的样本并没有覆盖所有的类别(ssh 类别中流样本数为 0).用 dump1 作为训练集符合实际流量分类情况,因为网络实际情况中时常出现新的协议流量,而训练集只能够覆盖部分常用的协议类别.

4.1 评价指标

(1) 召回率和精度

基于机器学习的流量识别技术常用召回率(recall)和精度(precision)两个指标来评价识别结果.

漏报(false negative)是指属于类别 C 的流量而被分类成非类别 C,真阳性(true positive)是指属于类别 C 的流量而被分类成类别 C,误报(false positive)是指非类别 C 的流量被分类成为类别 C.假定误报数为 FP,漏报数为 FN,真阳性数为 TP,召回率和精度的计算方法如下:

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN}.$$

(2) 流的准确性和字节准确性

流的准确性指被正确识别的流数占标注数据集中所有流数的百分比,字节的准确性是指被正确识别的数据流承载的字节数占标注数据集总字节数的百分比.

4.2 流准确性的比较

Naïve Bayes 核估计、SVM 和 C4.5 决策树这 3 种分类器的流准确性如图 4 所示.

由图 4 可见,C4.5 决策树分类算法和 NBK 分类算法在不同数据集上,分类的流准确性不稳定,但 SVM 分类算法相对较稳定.由于数据集是按照时间顺序排列的,3 种分类器的流准确性的总体趋势随着时间的推移是逐渐降低的.文献[15]指出,新协议流量的出现是导致流量分类器准确率下降的主要原因.但是,当用 dump2 作为训练集(dump2 数据集已经覆盖了所有类别的流量),而用 dump3,dump4,dump5 和 dump6 作为测试集时,3 种分类器的流准确性也在逐渐下降,如图 5 所示.因此,在本文实验数据集上,新协议流量的出现不是造成分类器流准确性下降的主要原因.从图 1(第 3.2 节所示)可见,不同时间采集的数据集上,流样本分布也不同.流样本分布的动态变化造成了 3 种分类器随着时间的推移整体准确性逐渐下降.由图 4 和图 5 可见,NBK 分类算法的流准确性最不稳定,且随着时间的推移下降趋势最快.

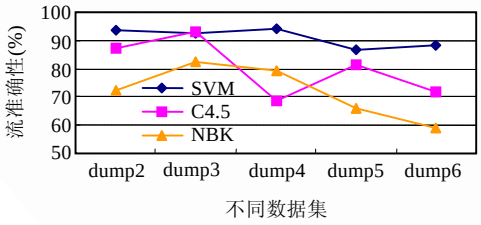


Fig.4 Flow accuracy of the three classifiers

图 4 3 种分类器的流准确性

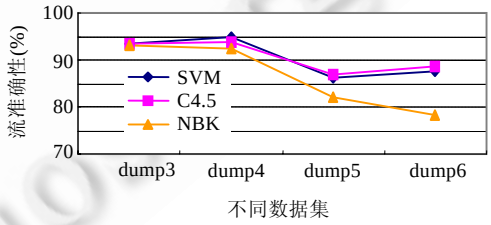


Fig.5 Flow accuracy when dump2 is used as training data set

图 5 dump2 作为训练集时分类的流准确性

当用 dump1 作为训练集时,我们通过对比 3 种分类器识别各种协议类别的召回率(如图 6 所示)和精度(如图 7 所示),发现 3 种分类器识别 http 协议的精度差别不大.对比图 6 和图 4 发现:由于 SVM 对 http 协议的召回率高且稳定,使得 SVM 的整体流准确性是 3 种分类算法中最高且最稳定的;而且在不同数据集上,如果分类算法对 http 协议识别的召回率高,那么该分类算法的整体准确性也高.也就是说,在网络协议流不平衡环境下,分类算法对流样本数比例大的类别识别的召回率影响它分类的整体准确性.

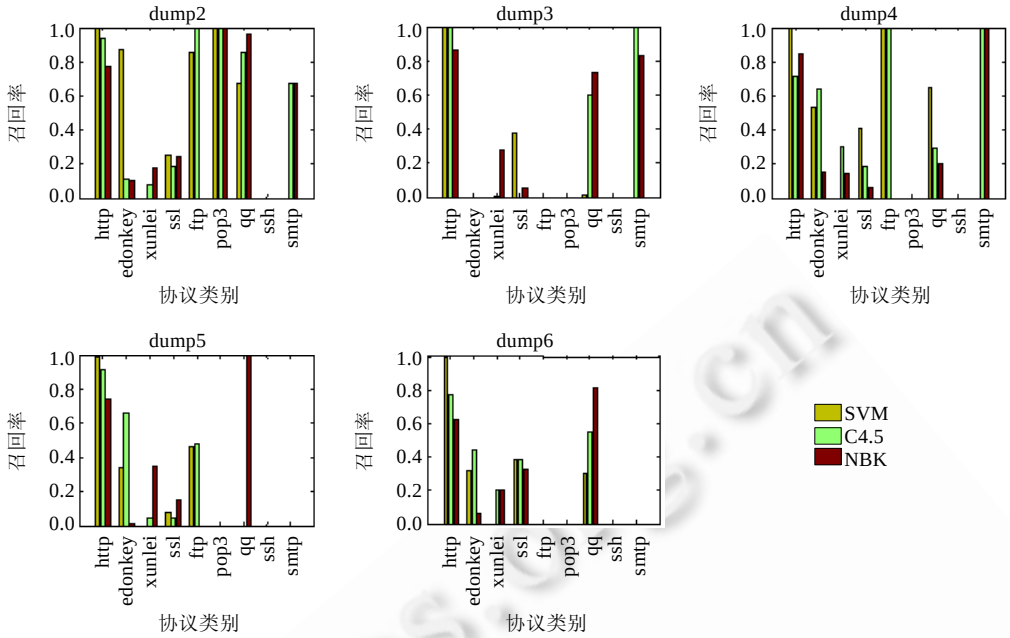


Fig.6 Recall of the three classifiers when classifying each protocol

图 6 3 种分类器分类每种协议的召回率

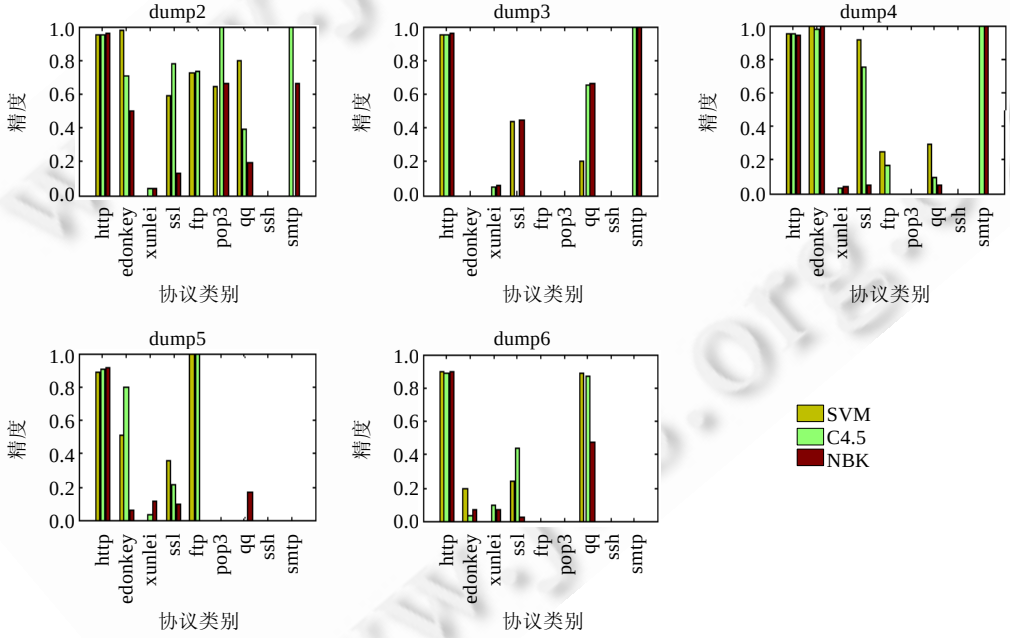


Fig.7 Precision of the three classifiers when classifying each protocol

图 7 3 种分类器分类每种协议的精度

由图 6 和图 7 可见,3 种分类器对于流样本数比例较小的类别、识别的精度和召回率都不稳定.例如,在 dump2 数据集上,识别 qq 协议的召回率和 dump4 数据集上的召回率相差就很悬殊.C4.5 和 NBK 分类算法对于

流样本比例极小的类别,例如 smtp 协议,识别的召回率和精度要高于 SVM.

4.3 字节准确性的比较

3 种分类器总体字节准确性对比如图 8 所示.3 种分类器在 dump4 数据集测试的准确性很低,这是因为训练集 dump1 没有 ssh 的流量样本,而测试集 dump4 有 ssh 大流样本(如图 2 所示),由于 3 种分类器识别不到 ssh 大流,3 种分类器字节准确性下降.这种现象意味着新的协议流量出现,应该重新训练分类器.

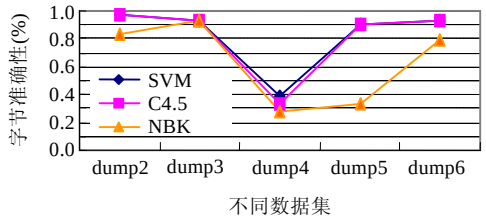


Fig.8 Byte accuracy of the three classifiers

图 8 3 种分类器的字节准确性

SVM 和 C4.5 决策树分类的字节准确性差异不大.在 dump5 测试集上,C4.5 决策树和 SVM 分类算法的字节准确性明显高于 NBK 分类算法.比较 3 种分类器对每种协议识别的字节准确性如图 9 所示,C4.5 决策树和 SVM 分类算法在 dump5 数据集上,对 eDonkey 协议识别的字节准确性都较高;而在这个数据集上,eDonkey 协议出现了大流.所以在 dump5 数据集上,C4.5 决策树和 SVM 分类算法的总体字节准确性要高于 NBK 分类算法.此外,相比于 SVM 分类算法,C4.5 决策树分类算法更易识别样本比例极小的协议类别,例如 smtp 协议流.这意味着,如果样本比例极小的协议产生了大流,C4.5 决策树用作流量分类器是最合适的.

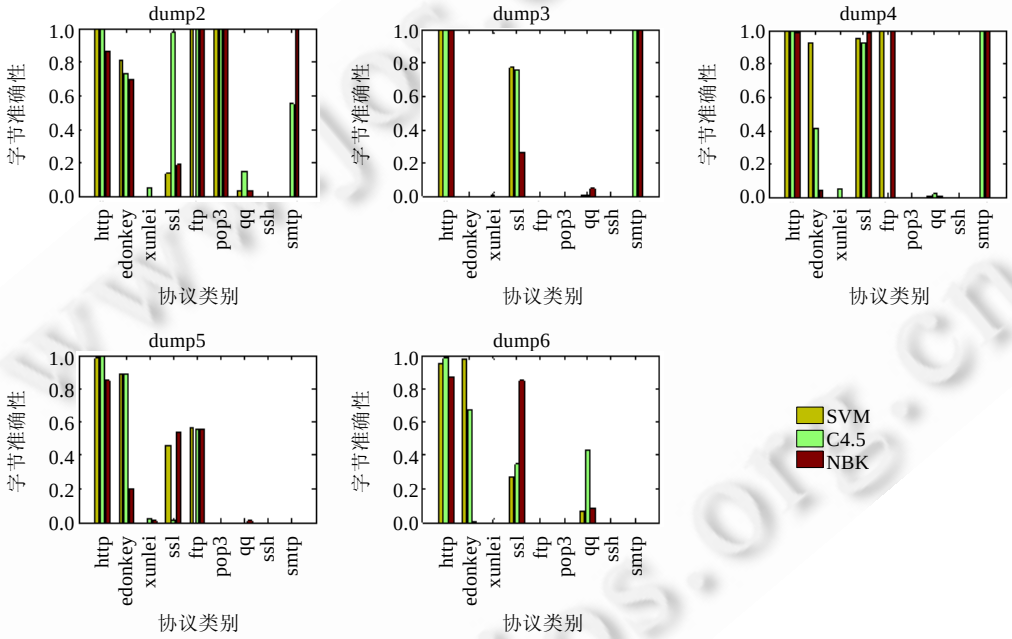


Fig.9 Byte accuracy of the three classifiers when classifying each protocol

图 9 3 种分类器分类每种协议的字节准确性

4.4 3种分类算法的建模和测试时间

本节实验的测试环境为:测试主机运行 Windows XP 系统,主存为 3GB,处理器是 AMD 9650 2.3GHZ.SVM, C4.5 决策树和 NBK 这 3 种分类算法在 dump1 数据集上的建模时间对比见表 7.

Table 7 Modeling time of the three classifiers (s)

表 7 3 种分类器的建模时间 (秒)

算法	SVM	C4.5	NBK
建模时间	6.38	0.31	0.03

由表 7 可见,SVM 分类算法的建模时间明显比 C4.5 决策树分类算法和 NBK 分类算法的建模时间长.进一步比较 3 种分类算法在 5 个测试数据集上的测试时间,见表 8.我们发现,SVM 和 NBK 分类算法的测试时间明显高于 C4.5 算法的测试时间,即 C4.5 决策树分类算法更适合于在线分类.

Table 8 Testing time of the three classifiers (s)

表 8 3 种分类器的测试时间 (秒)

测试数据集	算法		
	SVM	C4.5	NBK
dump2	3.828	0.078	3.516
dump3	1.985	0.032	1.687
dump4	2.312	0.063	2.188
dump5	1.344	0.016	1.25
dump6	2.766	0.032	2.672

5 基于 Bagging 集成学习的流量分类器

目前,机器学习领域中常采用两种方案解决类别不平衡问题:

- 重新采集样本技术(resampling techniques),即样本比例较小的类别多采集一些样本,或者样本比例较大的类别删除一些样本.这种方案主要的缺点在于破坏了原有各类别样本分布的状况;
- 构建集成学习分类器.集成学习分类器由一系列单独训练的分量分类器(component classifier)构成,其目的是利用单个模型之间的差异改善模型的泛化性能.集成学习通过选取不同的数据集来获取个体模型间的差异性.一般来说,差异性越大,泛化能力越强,分类器就越稳定.

本节选用 Bagging 算法建立集成分类器,因为 Bagging 算法通过调节一个参数就可以确定分量分类器间的差异性.

5.1 Bagging算法描述

Bagging 算法从大小为 N 的原始训练集 X 中,依次有放回地独立随机抽取 $N_1(N_1 \leqslant N)$个样本形成自助训练集,并将这个过程独立重复进行多次,直到产生多个独立的自助数据集.然后,在每个自助数据集上独立地训练一个分量分类器,最终的分类判决将根据这些分量分类器各自判决结果的投票决定.

由于 C4.5 决策树分类算法能够确保样本比例极小的类别识别率,且测试时间和建模时间都较短,因此本文选择 C4.5 决策树分类算法建立分量分类器,试图通过 Bagging 算法解决原始 C4.5 决策树分类算法的不稳定性问题.

鉴于文献[18]的研究方法,我们认为,当 N 很大时,N_1 近似服从 Poisson(λ)分布,通过调节参数λ值来确定集成学习分类器的差异性.文献[18]认为,当λ值越小,差异性越大,分类器越稳定.但是本文发现,λ值越小将导致集成分类器对样本比例小的类别召回率降低.

图 10 比较了λ=0.05,λ=0.1 和λ=0.2 时集成学习分类器的召回率情况.例如,eDonkey 协议属于样本比例小的类别,从图 10 中还可以看出,λ值越小,对 eDonkey 协议识别的召回率就越小.我们在分类稳定性与召回率间做了折中,选择λ=0.1 来建立集成学习分类器.

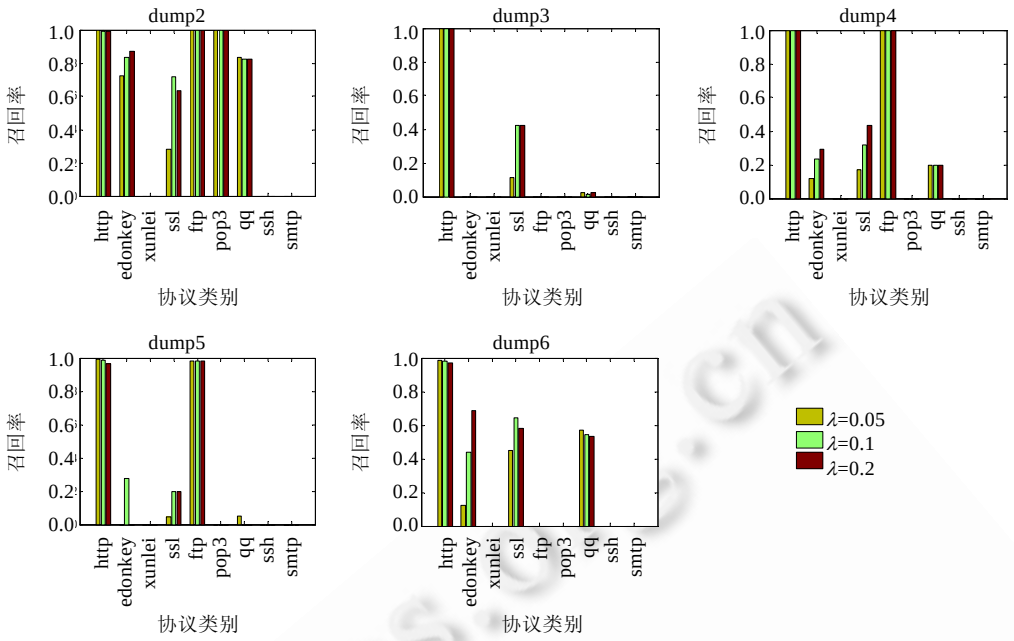


Fig.10 Recall of Bagging with different λ

图 10 选择不同λ时 Bagging 算法的召回率

5.2 流准确性

本节对比分析 C4.5 决策树、SVM、NBK 分类算法以及 Bagging 集成学习算法分类的流准确性.

由图 11 可见,基于 Bagging 算法的流量分类器,其稳定性和准确性相似于 SVM 分类算法,明显好于 C4.5 决策树和 NBK 分类算法.但是分类的流准确性也在随着时间逐渐下降(横轴是按照时间顺序排列的数据集),下降趋势要比 C4.5 决策树和 NBK 分类算法更加缓慢.由召回率和精度对比(如图 12、图 13 所示)可以看出,基于 Bagging 算法的流量分类器对样本数极少的类别分类能力不如单独使用 C4.5 决策树和 NBK 分类算法.例如,Bagging 集成学习分类器没有识别出 smtp 协议流.

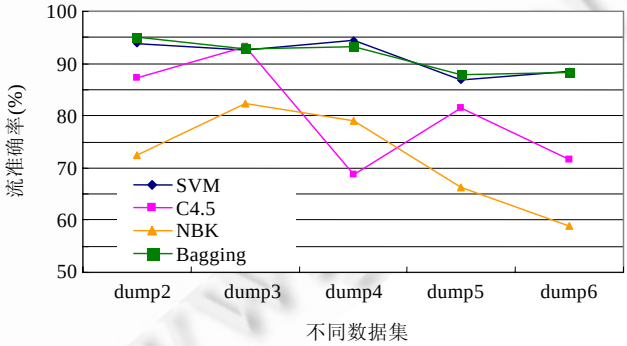


Fig.11 Flow accuracy of four classification algorithms

图 11 4 种分类算法的流准确性

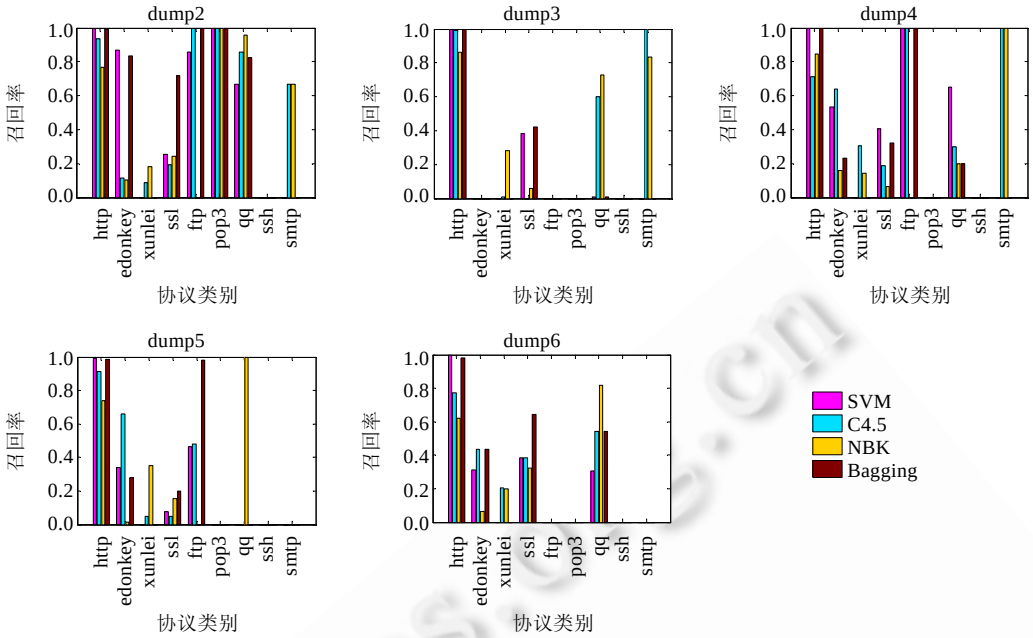


Fig.12 Recall of four classification algorithms when classifying each protocol

图 12 4 种分类算法分类每种协议的召回率

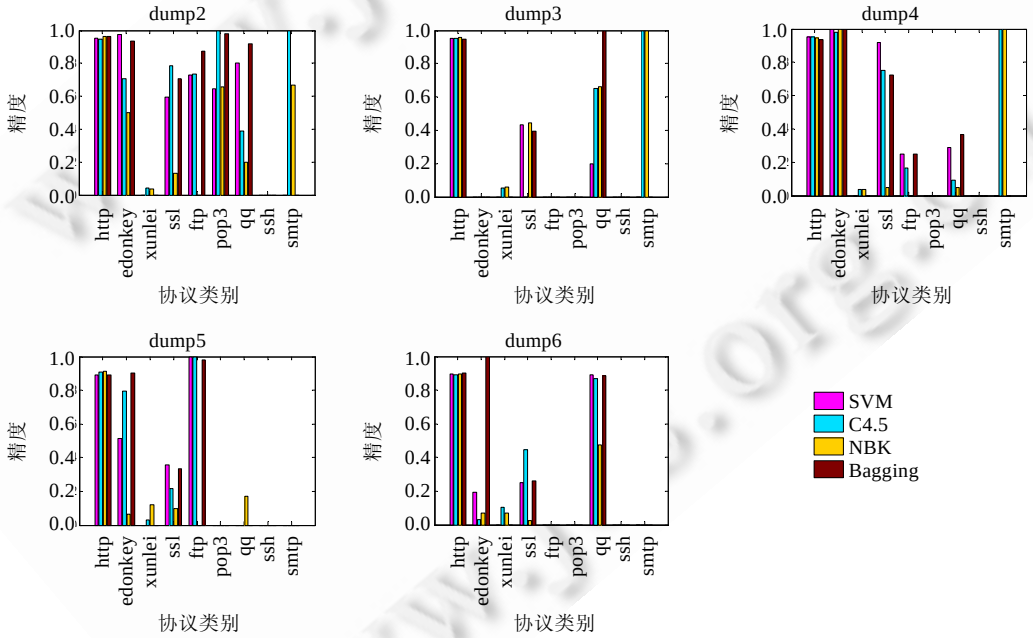


Fig.13 Precision of four classifiers when classifying each protocol

图 13 4 种分类算法分类每种协议的精度

5.3 字节准确性

就字节准确性而言,由图 14 可见,由于未识别出测试集 dump4 中的 ssh 大流样本,4 种分类器的字节准确性

陡然下降.这就意味着需要重新训练分类器以识别新协议流.

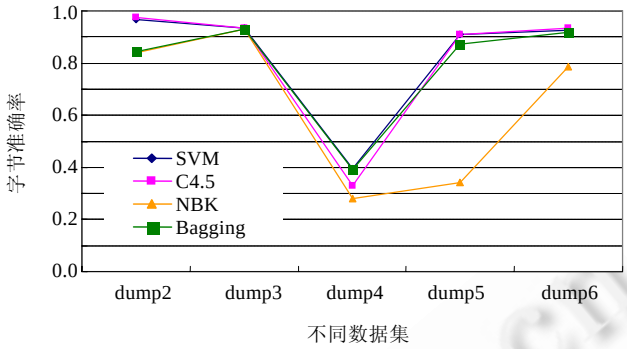


Fig.14 Byte accuracy of four classifiers

图 14 4 种分类器的字节准确性

由图 15 可以看出:Bagging 集成学习流量分类器对于流样本数比例较大的类别,分类字节准确性较高而且稳定;而对于流样本数比例极小的类别,例如 smtp 协议,其识别的字节准确性不如 C4.5 和 NBK 算法.

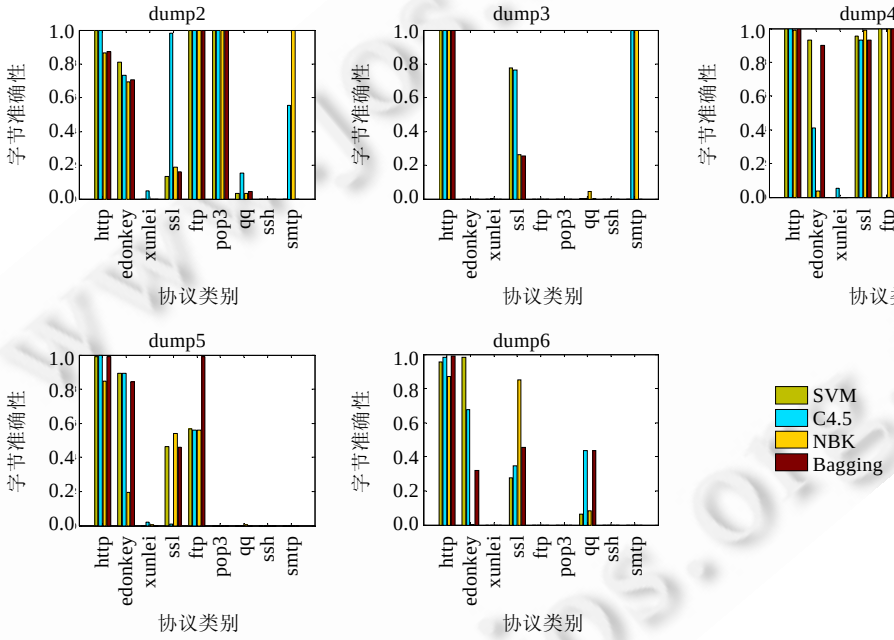


Fig.15 Byte accuracy of four classifiers when classifying each protocol

图 15 4 种分类算法分类每种协议的字节准确性

5.4 Bagging分类算法的建模时间和测试时间

我们在 dump1 数据集上建立 Bagging 分类模型的时间为 0.39s,其建模时间略高于 C4.5 决策树分类算法,而明显低于 SVM 分类算法的建模时间(见表 7).Bagging 集成学习分类器的测试时间见表 9.

对比表 8 和表 9 我们发现,Bagging 算法的测试时间略高于 C4.5 决策树算法的测试时间,高出的时间范围仅在[0.04,0.1]之间.而 Bagging 算法的测试时间明显小于 SVM 和 NBK 分类算法的测试时间.

Table 9 Testing time of Bagging (s)

表 9 Bagging 分类器的测试时间 (秒)

数据集	dump2	dump3	dump4	dump5	dump6
测试时间	0.187	0.078	0.125	0.062	0.14

综上分析,基于 Bagging 集成学习算法的流量分类器的分类准确性较高且稳定,其测试时间和建模时间较短,因此更适合于在线分类网络流量.

6 总 结

基于机器学习算法的流量分类器应用于在线分类,是目前流量分类研究的一个趋势.网络环境的变化会影响机器学习流量分类器在线分类的性能.本文主要分析协议流不平衡性对流量分类器的影响,协议流不平衡是网络环境中常见现象.本文在网络协议流不平衡的数据集上进行实验分析,得出以下几点结论:

(1) 流量分类器在线分类时,要能够确保在连接开始时尽早地分类流.与以往研究结论不同的是,我们发现,在统计流开始的前 4~10 个数据包时,SVM 和 C4.5 决策树分类算法的准确性并无显著性差异.根据尽早分类流的原则,我们选择统计流开始的前 4 个数据包.尽管统计不同的数据包数目对于 NBK 分类算法有显著性差异,但是对于 NBK 分类算法统计前 4 个数据包也是最合适的;

(2) 在网络协议流不平衡的环境下,SVM 分类算法要比 C4.5 决策树和 NBK 分类算法稳定.SVM 分类算法对样本比例大的类别具有最好的召回率,但是 C4.5 决策树和 NBK 分类算法更能够确保对样本比例小的类别的召回率和精度.3 种分类算法相比之下,C4.5 分类算法的测试时间最短,NBK 的建模时间最短,而 SVM 的测试时间和建模时间最长;

(3) 在网络协议流不平衡的环境下,分类器对流样本比例大的类别识别的召回率将影响分类器的总体准确性.流样本分布的动态变化,是造成机器学习流量分类器整体准确性随着时间的推移逐渐下降的原因之一;

(4) 在网络协议流不平衡的环境下,基于 Bagging 集成学习的流量分类器较为稳定,其准确性较高,且测试时间较短,适合于在线分类网络流.此外,本文发现,虽然提高 Bagging 集成学习流量分类器的差异性会使分类器更加稳定,但是这种差异性的提高也会使分类器对样本比例较小的类别识别准确性下降.相比之下,C4.5 决策树和 NBK 分类算法更能确保对样本比例较小的类别识别的流准确性和字节准确性.

我们未来的研究工作在于进一步改进 Bagging 算法,使其对流样本比例极小的类别能够有较高的识别率.此外我们发现,SVM、C4.5 决策树、NBK 和 Bagging 这 4 种分类算法对迅雷协议流量识别的召回率很低(如图 12 所示),这是因为迅雷流量默认使用 80 端口进行通信,其统计特征与 http 协议流量很相似.如何将 http 协议流量与迅雷协议流量准确地区分开,可以作为我们进一步的研究方向.

References:

[1] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification. In: Ceballos S, ed. Proc. of the 14th IEEE Int'l Symp. on Modeling, Analysis, and Simulation. Washington: IEEE Computer Society Press, 2006. 179−188. [doi: 10.1109/MASCOTS.2006.6]

[2] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark. In: Guerin R, Govindan R, Minshall G, eds. Proc. of the 2005 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM 2005). New York: ACM Press, 2005. [doi: 10.1145/1080091.1080119]

[3] Kim H, Claffy KC, Fomenkov M, Barman D, Faloutsos M, Lee K. Internet traffic classification demystified: Myths, caveats, and the best practices. In: Azcorra A, Veciana G, eds. Proc. of the 2008 ACM CoNEXT Conf. New York: ACM Press, 2008. 1−12. [doi: 10.1145/1544012.1544023]

[4] Tstat. TCP statistic and analysis tool. 2011. http://tstat.tlc.polito.it

[5] Bonfiglio D, Mellia M, Meo M, Rossi D, Tofanelli P. Revealing skype traffic: When randomness plays with you. ACM SIGCOMM Computer Communication Review, 2007,37(4):37−48. [doi: 10.1145/1282380.1282386]

[6] Moore AW, Zuev D. Internet traffic classification using bayesian analysis techniques. ACM SIGMETRICS Performance Evaluation Review, 2005,33(1):50−60. [doi: 10.1145/1064212.1064220]

[7] Li W, Canini M, Moore AW, Bolla R. Efficient application identification and the temporal and spatial stability of classification schema. Computer Networks, 2009,53(6):790−809. [doi: 10.1016/j.comnet.2008.11.016]

[8] Este A, Gringoli F, Salgarelli L. On the stability of the information carried by traffic flow features at the packet level. ACM SIGCOMM Computer Communication Review, 2009,39(3):13−18. [doi: 10.1145/1568613.1568616]

[9] Bernaille L, Teixeira R, Akodkenou I. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2):23−26. [doi: 10.1145/1129582.1129589]

[10] Soysal M, Schmidt EG. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Performance Evaluation, 2010,67(6):451−467. [doi: 10.1016/j.peva.2010.01.001]

[11] Pietrzyk M, Costeux JL, Urvoy-Keller G, En-Najjary T. Challenging statistical classification for operational usage: The ADSL case. In: Feldmann A, Mathy L, eds. Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement Conf. (IMC 2009). New York: ACM Press, 2009. 122−135. [doi: 10.1145/1644893.1644908]

[12] Xu P, Lin S. Internet traffic classification using C4.5 decision tree. Journal of Software, 2009,20(10):2692−2704 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

[13] Erman J, Mahanti A, Arlitt M. Byte me: A case for byte accuracy in traffic classification. In: Sen S, Sahu S, eds. Proc. of the 3rd Annual ACM Workshop on Mining Network Data (MineNet 2007). New York: ACM Press, 2007. 35−37. [doi: 10.1145/1269880.1269890]

[14] Xu P, Liu Q, Lin S. Internet traffic classification using support vector machine. Journal of Computer Research and Development, 2009,46(3):407−414 (in Chinese with English abstract).

[15] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Diot C, Ammar M, eds. Proc. of the 2006 ACM CoNEXT Conf. New York: ACM Press, 2006. 1−12. [doi: 10.1145/1368436.1368445]

[16] L7-filter. Application layer packet classifier for Linux. 2009. http://l7-filter.sourceforge.net

[17] L7-filter supported protocols. 2009. http://l7-filter.sourceforge.net/protocols

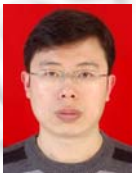
[18] Minku LL, White AP, Yao X. The impact of diversity on online ensemble learning in the presence of concept drift. IEEE Trans. on Knowledge and Data Engineering, 2010,22(5):730−742. [doi: 10.1109/TKDE.2009.156]

附中文参考文献:

[12] 徐鹏,林森.基于 C4.5 决策树的流量分类方法.软件学报,2009,20(10):2692−2704. http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

[14] 徐鹏,刘琼,林森.基于支持向量机的 Internet 流量分类研究.计算机研究与发展,2009,46(3):407−414.

张宏莉(1973－),女,吉林榆树人,博士,教授,博士生导师,主要研究领域为网络信息安全,网络测量,并行处理.



鲁刚(1982－),男,博士生,主要研究领域为流量分类,P2P 技术.