ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.11, November 2010, pp.2852−2865
doi: 10.3724/SP.J.1001.2010.03747

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

基于卡方统计的应用协议流量行为特征分析方法*

陈 亮[1,2+], 龚 俭[1,2]

[1](东南大学 计算机科学与工程学院,江苏 南京 210096)

[2](江苏省计算机网络技术重点实验室,江苏 南京 210096)

Analyzing the Characteristics of Application Traffic Behavior Based on Chi-Square Statistics

CHEN Liang[1,2+], GONG Jian[1,2]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

[2](Key Laboratory of Computer Network Technology of Jiangsu Province, Nanjing 210096, China)

+ Corresponding author: E-mail: lchen@njnet.edu.cn

Chen L, Gong J. Analyzing the characteristics of application traffic behavior based on chi-square statistics. *Journal of Software*, 2010,21(11):2852−2865. http://www.jos.org.cn/1000-9825/3747.htm

Abstract: Based on the Chi-Square Statistics and Test, this paper proposes a method named ABSA (application behavior significance assessment) to analyze the traffic behavior characteristics of applications. The ABSA method does not focus on any certain applications; in contrast, it aims at providing a quantitative standard for describing the behavior distribution differences among applications, so that the traffic behavior characteristics and their corresponding significances can be determined. The theoretical analysis and experiments results show that 1) ABSA can present the information about characteristics more precisely and copiously to improve the accuracy of application identification; 2) the significance of characteristic is independent of its proportion in sample totals; 3) ABSA can keep the relative significance sequence of behavior characteristics unchanged in a packet sampling environment, which is often used by NetFlow and many other flow information collecting systems to simplify the characteristic re-selecting process when sampling ratio is changed.

Key words: network behavior; application-level protocol; traffic identification; behavior characteristic; chi-square statistics; packet sampling

摘 要: 引入统计理论中的卡方统计检验,提出一种通用的应用协议流量行为特征分析方法——ABSA (application behavior significance assessment).该方法不针对特定的应用协议,旨在提出描述各应用协议间行为测度分布差异情况的统一量化标准,使其可进行比较,从而判断各协议的流量行为特征,并评估相应的显著程度.理论分析及实验结果表明,ABSA 方法不仅可以为协议识别提供更丰富、更准确的特征信息,优化协议识别的结果,而且保证特征显著程度的评估与协议样本在总样本中所占的比例无关,并可用于 NetFlow 等路由器所用的报文抽样环境下,保持以任意比例抽样后的特征相对显著程度顺序评估结果不变,简化了抽样比变化时的特征重选择过程.

关键词: 网络行为;应用协议;流量分类;行为特征;卡方统计;报文抽样

* Supported by the National Basic Research Program of China under Grant No.2009CB320505 (国家重点基础研究发展计划(973)); the National Key Technology R&D Program of China under Grant No.2008BAH37B04 (国家科技支撑计划)

Received 2009-04-16; Revised 2009-07-06; Accepted 2009-10-10

中图法分类号: TP393 文献标识码: A

准确地标识 Internet 流量所使用的应用层协议是网络 QoS、SLA、网络流量和用户行为监控等的前提,且对网络性能管理、网络计费管理和入侵检测等的研究具有指导意义.由于传统的以端口号标识应用协议的低精度[1,2]和以深度报文检测识别应用协议的高时空复杂度的缺点,自 2004 年开始,基于流量行为识别应用协议的方法逐渐成为国内外的研究热点[3−7].然而,目前的研究成果在识别的整体精度、粒度和识别方法的通用性等方面达不到令人满意的效果.其中重要的原因之一是,这类研究只停留在应用各已有的分类算法于流量识别领域,却没有更本质地分析哪些行为测度是应用协议的流量特征,缺乏理论基础.

所谓应用协议流量行为特征是指该应用协议在实际使用时表现出的有别于其他应用协议的测度分布,包括时间维上应用流的行为测度(如流长、流内报文到达间隔)、空间维上应用主机的行为测度(如链接数、上/下行流量比)以及主机间的拓扑测度等.研究测度分布差异及协议行为特征对应用协议识别的意义在于:(1) 对于协议间分布无差异的测度,任何表示方法都不能将其区分,这样的测度对任何协议识别方法都是无用的测度;(2) 只有清楚地认识到某应用协议的哪些行为与网络总体情况相异,可作为行为特征,才有可能更深入地分析这些行为测度变量的分布情况,从而有针对性地选择或提出合适的识别算法,从根本上提高该协议识别的精度;(3) 只有发现了各应用协议的行为特征,才有可能总结出对总体网络流量标识各应用协议的最佳测度集合,为各特征选择方法的效果判定提供依据,为各协议识别方法奠定基础,提高总体流量识别的精度;4) 只有更细致地了解了协议间的哪些行为测度存在分布差异及其显著程度,才有可能发现协议识别的粒度和测度种类/数量之间的关系,从根本上提高协议识别的粒度.

虽然特征选择在机器学习和数据挖掘领域已有很多研究,如基于信息增益的方法、基于相关关系的方法、基于卡方统计的方法等[8],但这些方法在应用于实际环境的协议识别时存在一些弊端:(1) 现有方法均专以样本分类为目的,当分类的目标对象略有变动时就必须重新使用方法选择新的特征,复杂度高且缺乏灵活性;(2) 现有方法只统一地考虑了某特征对识别所有类别的贡献而得出其重要程度,并没有考虑各类别间的区别,因此有时并不能得出最优特征集合;(3) 一些方法如信息增益,其训练样本的类间数量比例对方法的结果有很大影响,因此样本的采集要求高,需完全符合实际情况;(4) 现有方法需要多个测度的统计结果相比较才能得出测度的相对重要性,若只有单一测度的计算值,不能客观判定其是否可作为协议行为特征;(5) 现有的通用网络流信息统计系统(如 NetFlow[9])均采用了报文抽样的方式得出流行为测度值,若能发现报文抽样对协议行为特征选择的影响,则可以极大地提高特征选择方法的适用范围;但现有文献未见对此方面的研究.

另一方面,在网络协议流量行为研究领域,应用层协议流量行为的分析自 Internet 发展起一直都是网络测量管理的研究重点之一[10−14],但迄今为止,这些研究都集中在统计某个或某些应用协议行为测度的均值和分布,尚未有文献对其进行横向比较,提出衡量应用协议间主要行为差异,分析协议行为特征的方法.

针对以上不足,本文提出一种基于数理统计中的卡方统计检验(χ^2-statistics test)分析应用协议行为特征,并对各特征的显著程度进行评估的通用方法 ABSA(application behavior significance assessment).该方法不针对特定的应用协议,旨在提出描述各应用协议间行为测度分布差异情况的统一量化标准,从而判断各协议的行为特征并评估特征的相应显著程度.该方法还可用于 NetFlow 等路由器所使用的报文抽样环境下,保持在不同抽样比环境下特征的相对显著程度顺序评估结果不变,简化了报文抽样比变化时特征的重新选择工作.该方法易与目前通用的流信息统计系统配合实现应用协议的特征分析,为应用协议识别、网络监测与管理提供重要信息.

本文第 1 节简单介绍χ^2统计的基本知识和机器学习中基于χ^2统计的特征选择方法,并指出现有χ^2统计特征选择方法的不足.第 2 节分析χ^2统计在协议行为特征分析背景下应用时的问题及解决方法,提出 ABSA 方法.第 3 节使用实际采集的报文 Trace 实验,验证 ABSA 方法的准确性,类间样本数量无关性以及对协议识别精度的改进.第 4 节分析当前路由器所使用的报文抽样方法对行为测度分布差异及 ABSA 的影响,指出 ABSA 方法的适用范围.第 5 节是全文总结.

1 χ^2统计和χ^2特征选择方法

协议行为特征判定的本质,是同一行为测度变量在两对立样本集合中取值分布相似程度的判定.若两分布相同,则样本所指协议在该行为上表现一致,测度不能作为协议特征;否则,行为存在差异,测度可以作为行为特征,且分布的差异程度决定了特征的显著程度.协议行为特征的判定应独立于各分类方法,且不依赖于样本间数量的比例.

χ^2统计[15]的一般提法是:令$(X_1,\ldots,X_n)$是来自可观测随机变量X的一个样本,F是一个已知的分布函数,如果用分布F去拟合样本$(X_1,\ldots,X_n)$,则两分布差异如何?设理论分布F为

$$P(X=a_i)=p_i,i=1,2,\ldots,v,$$

其中,$p_i>0$已知,$\sum_{i=1}^{v} p_i = 1$.以n_i记$X_1,X_2,\ldots,X_n$中等于a_i的个数,即a_i的观察频数,$\sum_{i=1}^{v} n_i = n$.称np_i为a_i的理论频数.做χ^2统计量:

$$\chi^2 = \sum_{i=1}^{v} \frac{(n_i - np_i)^2}{np_i} \tag{1}$$

若样本容量n较大,由上式定义的统计量χ^2的渐近分布为自由度v–1的χ^2-分布.若分布F可以很好地拟合样本,则近似代表观察频数与理论频数之间相对误差平方的χ^2统计量应较小;反之,差异越大,χ^2值越大.

若将F看作另一样本$(X_1,\ldots,X_m)$中的X分布,则拟和检验问题归为同一变量在两样本中的分布相似程度分析问题,即协议行为特征判定的本质问题.因此,假设两样本分布(构成比)如下:

	区间 1	区间 2	…	区间 v
样本 1	n_{11}	n_{12}	…	n_{1v}
样本 2	n_{21}	n_{22}	…	n_{2v}

表中各项为对应样本落在各区间范围内的个数,即观察频数.

将理论频数$n_R p_i = n_R \times \frac{n_{1i}+n_{2i}}{n_{11}+\ldots+n_{1v}+n_{21}+\ldots+n_{2v}} = \frac{\text{行合计}\times\text{列合计}}{\text{总例数}} = \frac{n_R n_C}{n}$代入公式(1),可得计算两样本率之间差别显著性的专用公式:

$$\chi^2 = n\left(\sum_{C=1}^{v}\sum_{R=1}^{2}\frac{n_{RC}^2}{n_R n_C} - 1\right),\ \text{自由度}=v-1 \tag{2}$$

n_{RC}为表中各项数值,n_R为该项所在的行合计,n_C为列合计,n为两样本总例数.将所得样本各项数值代入公式(2),即可得到表示两样本中变量分布差异显著情况的χ^2统计量.

机器学习中的卡方特征选择方法是将公式(2)扩展至多类别样本率的比较[12],即

$$\chi^2 = n\left(\sum_{C=1}^{v}\sum_{R=1}^{u}\frac{n_{RC}^2}{n_R n_C} - 1\right),\ \text{自由度}=(u-1)(v-1) \tag{3}$$

其中,u为样本类别个数.当u确定后,该方法使用公式(3)计算每个测度的χ^2统计值,作为其在类间分布差异程度的度量;而后,对计算所得的χ^2统计量排序,并设定阈值选取相应的测度作为分类的依据特征.

但是,与其他特征选择方法一样,现有卡方特征选择方法缺乏当待识别对象变动时的灵活性,缺乏单特征的判定方法;且其计算方式只考虑了待识别的所有类间的测度分布差异,并不表明其中任意两个或多个分布间存在差异.因此,该方法往往不一定能够得到最优测度集合.另外,卡方统计中的自由度选取、变量的区间划分都会对统计的结果产生极大的影响,需要在应用协议行为差异分析背景下对这些问题作更详细的讨论.

2 基于χ^2统计的行为特征分析方法

针对上节指出的现有特征选择方法应用于协议识别领域的不足,本节提出一种新的特征发现方法——ABSA.该方法仍采用卡方统计作为基本理论,但通过优化的过程准确地分析仅两样本空间中变量分布的差异程度,揭示出某协议流量行为较网络总体流量的特征,或两协议之间的流量行为差异,从而为特征的选择与优化

奠定基础.

2.1 单测度的特征判断

如本文先前所述,现有机器学习中的 ranking 类方法(包括现有的卡方特征选择方法)针对每个特征给出其重要性评估,之后按照评估值排序选用高位的若干特征.但是,这些方法的评估值结果并不实际表明各测度是否真的具有分布差异而可作为特征;对于其中的每单个测度,其重要性评估值是孤立的,并没有任何意义,而只在多测度可比较时才能表现出相对的重要性.换句话说,当仅有若干具有极小评估值的测度可用时,这些方法仍将其排序而选用部分作为特征,虽然这些测度并不能对分类提供任何贡献.

因此,本文在评估特征重要程度之前,首先基于所使用的χ^2 统计,引入χ^2 拟合优度检验[15],以断言两样本空间中的变量是否符合同一分布,从而对每单个测度事先判断其是否可确实作为候选特征.设假设检验式为

H_0:两样本分布无差异.

若假设H_0成立,则根据公式(2)计算所得的χ^2统计量应较小.故可设定置信度α,根据当前的χ^2自由度查χ^2临界分布表,得到是否接受假设 H_0 的断言.对于接受假设断言的测度,不能作为样本协议的流量特征,无须进一步讨论其显著程度.

2.2 变量取值区间划分方法

上文所述的χ^2 统计和拟合优度检验是统计理论和机器学习中的理想情况.在实际的协议行为特征分析中,对变量 X 取值的不同分类会引起χ^2 统计值的改变,并有可能得到不同的假设判定结论.分类的不同表现于两方面:(1) 区间数确定,但区间划分不同;(2) 区间数不同.由公式(3)可知,当待识别的协议数目确定后(u 为常数),区间数和自由度之间存在一一对应关系.因此,分类的不同也可由自由度表示.由于统计检验过程应客观,故对于无明显类界的变量不应在区间划分中加入主观的影响.对此,本文采用基于分布的随机区间划分,并增加不同自由度和统计次数的方法,以减少不同的区间划分方法对结果的影响.具体如下:

对于区间数确定,但区间划分不同的问题,若自由度一定(设为 v),χ^2 检验还必须满足以下两个约束[15],否则计算结果可能产生偏差:

约束 1. 各理论频数 $T=\frac{n_C n_R}{n}\geqslant 1$.

约束 2. 对两类别样本,$1\leqslant T<5$ 的数目不应超过 $2(v+1)/5$.

根据以上约束,区间划分需在随机的基础上检查各理论频数,若发现不满足条件的 T,需合并该区间并重新随机拆分 T 较大的区间.迭代此过程,直至满足条件.由此可得出以下随机划分并调整的区间生成启发式算法 *Divide_Interval*.设变量最小值和最大值分别为 min 和 max.

Algorithm *Divide_Interval*(*v*).

```
Generate random numbers a_1,...,a_v in (min,max);
a_0=min; a_{v+1}=max;
flag=TRUE;
while (flag)
    flag=FALSE; count=0;
    for C=1 to v+1 and R=1 to 2
        Compute n_RC, n_R and n_C in every interval (a_{C-1},a_C);
        T_RC=n_R n_C/n;
    for C=1 to v+1 and R=1 to 2
        if (T_RC<1)
            Merge (a_{C-1},a_C) and T_RC with the previous or next interval;
            flag=TRUE; break;
```

```
        else
            if (1≤T_RC<5)
                ++count;
            if (count≥2(v+1)/5)
                Select smallest T_RC;
                Merge (a_{C-1},a_C) and T_RC with the previous or next interval;
                flag=TRUE; break;
    if (flag)
        Selete the largest T_RC;
        Generate random number a in (a_{C-1},a_C);
        Splite (a_{C-1},a_C) into (a_{C-1},a) and (a,a_C);
        Remark a_1,…,a_v;
```

应用算法 *Divide_Interval*,区间的最终划分在满足随机和理论频数要求的基础上,趋向符合测度分布的情况.即测度变量取值分布密集处,区间划分较细;分布离散处,区间跨度较大.从而使得区间划分不仅趋向稳定,且统计量可以更加稳定、准确地表现两分布的差异情况.如图 1 所示为 eDonkey 协议和网络总体流量流内报文数测度分布情况的χ^2统计量图(数据来源见本文第 3 节,自由度为 5).由图 1 可以看出,两统计量均值相同,但使用算法 *Divide_Interval* 后,χ^2统计值更趋向稳定于均值附近,计算值波动比未用 *Divide_Interval* 时要小很多,方差仅为未使用时的 3.62%(3.78E+08/1.05E+10).

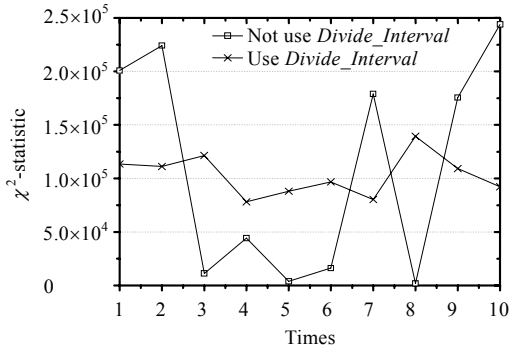


Fig.1 Contrast between χ^2-statistic results before and after using *Divide_Interval*

图 1 使用算法 *Divide_Interval* 前后的χ^2统计效果对比

对于区间数不同的问题,本文取 m 个自由度(如 m=5,v=5,10,20,30,40),对检验判定结果采用投票方式表决.

综合以上两解决方案,可推导出应用协议行为特征判定算法 *Determine_Characteristic*.令 *Vote* 为依赖于假设检验结果的投票函数,

$$Vote(\chi^2,v,\alpha)=\begin{cases}-1, & \chi^2<\chi^2_\alpha(v)\\ 1, & \text{Otherwise}\end{cases},$$

则判定算法 *Determine_Characteristic* 可简述如下:

```
Algorithm Determine_Characteristic().
    for i=1 to m
        for j=1 to t
            Divide_Interval(v_i);
            Compute  χ²_ij  according to formula (2);
```

$\chi_i^2 += \chi_{ij}^2$;

result+=***Vote***($\chi_i^2/t, v_i, \alpha$);

若 *result*<0,两分布吻合;否则,行为测度分布存在差异,测度可作为协议的行为特征.

2.3 行为特征显著性评估方法

由上述工作可判定应用协议的某一行为测度是否与其他协议或网络总体行为相异,从而可作为该协议的行为特征.对于那些可作为特征的测度,本节进一步提出评估这些特征相对显著程度的方法.

行为特征显著程度即该行为测度在两样本空间中的分布差异程度:分布差异越大,该行为作为特征的显著程度越高.由χ^2 统计性质可知,变量分布差异越显著,χ^2 统计量越大.因此,目前机器学习中的卡方特征选择方法直接采用χ^2 统计值作为特征的重要性评估值.但是,该性质的逆命题并不成立,χ^2 统计量的大小还取决于自由度的大小,自由度越大,统计量也越大.因此,只有考虑了自由度的影响,χ^2 统计量才能正确反映两分布的吻合程度.因此,若能对不同自由度下的χ^2 统计量进行均化,即可使用均化结果作为衡量特征显著程度的标准.

文献[16]建议采用χ^2 值与自由度之比作为衡量标准,然而χ^2 统计量的增长往往线性低于自由度的增长[15],如图 2 中方块线与圆圈线所示(方块线为某χ^2 统计量,圆圈线为统计量与自由度比值,统计数据来自于文献[15]).因此,若使用自由度对χ^2统计量进行均化,则随着自由度的增大,统计量所占权重变小,且偏差程度随自由度的取值距离(最大自由度-最小自由度)扩大更加明显,不能达到很好的均化效果.但是,如果取定前文所述的置信度α,使用该置信度下的χ^2 统计临界分位点对统计值进行均化,则其效果要远优于自由度,如图 2 中三角线所示.这是因为分位点不仅随自由度的增长而增长,而且包含更多该自由度下变量的分布信息.

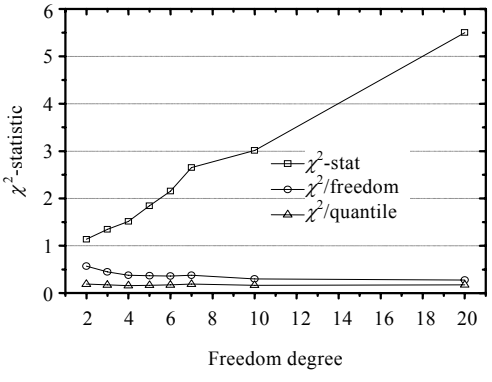


Fig.2 Relationship between χ^2-statistic and freedom degree

图 2 χ^2 统计量与自由度关系曲线

根据以上分析,可使用多自由度下的χ^2 统计量和临界分位点综合定义应用协议行为特征显著程度值 *BS* (behavior significance),即 *BS* 为χ^2 统计量被某置信度α下的临界分位点均化后的平均值:

$$BS=\frac{1}{m}\sum_{freedom=v_1}^{v_m}\frac{\chi^2\text{-}statistic_{freedom}}{quantile_{freedom}},\ significance\text{-}level=\alpha \tag{4}$$

其中,*m* 为第 2.2 节所使用的自由度个数.为进一步消除区间划分对统计结果的影响,上式中的χ^2 统计量是多次应用第 2.2 节算法 *Divide_Interval* 计算χ^2 统计量后取的平均值.据此,对于那些由算法 *Determine_Characteristic* 判断的可作为协议行为特征的测度,若其 *BS* 值越大,表明该行为测度分布差异越明显,测度作为行为特征的显著程度越高.理论上说,只要样本空间所代表的协议行为没有偏差,由第 2.2 节可知,使用算法 *Divide_Interval* 后,χ^2 值较为稳定,故 *BS* 的计算值较恒定,不随样本的选取有太大的抖动,可以作为综合评估协议行为特征显著程度的标准.

综合本节对χ^2 统计检验的所有分析,可将 *BS* 计算融入 *Determine_Characteristic* 中,得出应用行为特征显著

程度评估方法 ABSA:

Algorithm *ABSA*().

for i=1 to m

for j=1 to t

Divide_Interval(v_i);

Compute χ_{ij}^2 according to formula (1);

$\chi_i^2 += \chi_{ij}^2$;

$BS= \chi_i^2 /(t \times quantile_{v_i})$

result+=***Vote***($\chi_i^2 /t, v_i, \alpha$);

BS/=*m*;

若 *result*<0,测度不能作为协议的行为特征;否则,测度可作为特征,*BS* 值即为特征的相对显著程度,*BS* 值越大,行为特征越明显.

3 实验结果与分析

3.1 ABSA准确性分析

本节以目前应用最广的协议之一——eMule/eDonkey(电骡/电驴,以下统称 eDonkey)为例,应用 ABSA 方法分析其部分流量行为较网络总体行为的差异程度,表明 ABSA 方法的合理性与准确性.

实验 Trace 采集自 2008 年 08 月 20 日 15:00~16:00,江苏省教育网边界到 CERNET 国家主干路由之间,全报文长度,称为 Trace 1.由于信道吞吐量巨大,为了在保证 IP 流完整性的基础上方便实验,仅采集了省网外约 1/8 IP 地址范围与省网内所有地址交互的 IP 报文流量.虽然此类似集合抽样的做法可能导致行为分布的改变,但实验的目的仅是依赖当前 Trace 分析 ABSA 方法的正确性,而不是研究实际网络协议的行为分布情况,故不会影响以下分析的准确性.eDonkey 应用流量的识别采用 l7-filter[17]的协议模板 2008-12-18 版本.Trace 1 的总体情况和其中 eDonkey 的流量情况见表 1.

Table 1 General description of Trace 1

表 1 Trace 1 总体描述

TRACE	Number of IPs	Number of flows	Number of *pkts*	Number of *bytes*
ALL	4.15E+5	6.31E+6	1.94E+8	1.15E+11
eDonkey	7.23E+4	1.05E+6	9.69E+7	6.26E+10

为了消除 eDonkey 本身行为对网络总体行为的影响,以下实验所用的对比样本为 Trace 1 的所有流量中删除 eDonkey 流量所余部分,称为 Trace Non_eDonkey.以部分行为测度为例,应用 ABSA 方法,取 m=5,v=5,10,20,30,40,α=0.05,各 *BS* 计算值见表 2.

Table 2 *BS* value of some behavior metrics

表 2 部分行为特征显著性计算值

Metrics	*BS*	Metrics	*BS*	Metrics	*BS*	Metrics	*BS*
TCPflags	2.1E+5	*pkts*	5.3E+4	*pps*	2.9E+3	*pkts_ratio*	1.6E+3
pkt_size	9.0E+4	*duration*	1.1E+4	*pkt_size_ratio*	1.7E+3	*head_size*	5.1E+2
bytes	6.5E+4	*Bps*	3.1E+3	*bytes_ratio*	1.6E+3	—	—

虽然 ABSA 的投票结果表明表 2 中所示的 eDonkey 协议行为较之总体流量行为分布都有差异,但显著程度相差很大.由表 2 可以看出,TCP 标志位、报文长度和流长类测度(如 *bytes*,*pkts*,*duration*)的 *BS* 值最大,流速类(*pps*,*Bps*)和双向吞吐量比例类其次,报文首部长度测度 *head_size*(IP+TCP/UDP 首部)的 *BS* 值较小.对此结果,我们取各类中一个测度,详细分析其分布和差异情况以及形成原因,从而表明 ABSA 方法的判断和 *BS* 值可以准确

地衡量行为测度分布的差异情况.其中,*TCPflags* 的行为差异性已在文献[18]中分析,测度 *head_size*,*Bps*,*bytes* 在两样本空间中的概率分布曲线(probability distributionfunction,简称 PDF)如图 3~图 5 所示.

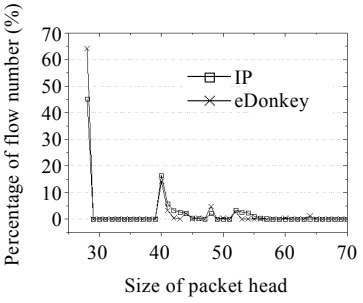


Fig.3 PDF of average packet head size of flow

图 3 平均流内报头长度 PDF

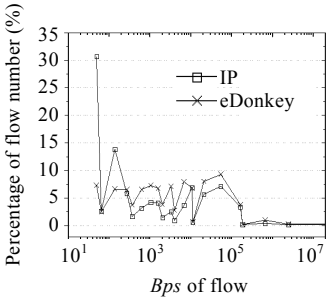


Fig.4 PDF of average *Bps* of flow

图 4 平均流内每秒字节数 PDF

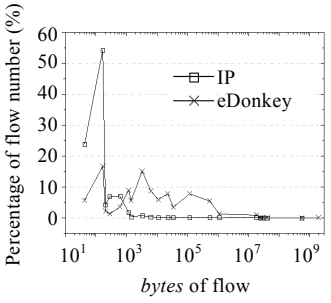


Fig.5 PDF of average *bytes* of flow

图 5 平均流内字节数 PDF

如图 3 所示,eDonkey 和 IP 总体流量的报头长度分布十分相似,差异主要体现于 28 字节和 40~52 字节处.由于 28 字节是 IP+UDP 报头的典型长度,40 字节~52 字节是 IP+TCP 报头的典型长度,我们推测,造成 *head_size* 分布差异的主要原因是 eDonkey 协议较网络总体流量更多地使用了 UDP 协议.因此进一步区分对待不同传输层协议,*head_size* 的 *BS* 值分别为 0.7(UDP)和 1.7(TCP).投票结果均表明,在置信度α=0.05 水平下,该测度在两样本空间中的分布无差异.即若考虑了传输层协议的影响,则 *head_size* 不能作为 eDonkey 协议的行为特征.

由图 4 可以看出,对于流速类测度,eDonkey 协议行为分布形似于 IP 流总体情况,分布差异较不明显.其中,低速流(*Bps*<500)所占比重比总体 IP 流少(23%:48%),流速均值略高于 IP 总流量情况(约为 1.39 倍).

图 5 为 eDonkey 协议和总体 IP 流的流内字节数测度 PDF,可以看出,两分布差异显著.其中,短流(*byte*<1K)所占比重,IP 流样本中约为 90%,而 eDonkey 协议流样本空间中只为 30%,eDonkey 的平均流长要显著大于总体 IP 平均流长(约为 3.50 倍).其主要原因即 eDonkey 作为一种 P2P 文件共享协议,大多被用于传输大媒体文件.这也是 eDonkey 的平均报文长度要显著大于网络总体情况的原因.

注意到,图 3~图 5 的纵轴缩放比例,3 图中测度分布差异情况逐渐显著,与表 2 中各测度的 *BS* 计算值保持一致.以上分析可进一步表明,*BS* 值与测度分布差异情况密切相关,行为测度的分布差异越明显,ABSA 方法的特征显著性评估值 *BS* 就越大.故可得出以下结论:ABSA 方法可以统一地判断某协议行为是否可作为该协议的行为特征;对于可作为特征的行为测度,ABSA 方法使用的 *BS* 值可以准确、合理地评估其作为特征的显著程度.同时,对 eDonkey 协议,报文长度和流长类测度是其区别于总体 IP 流量行为较显著的特征;而报文首部长度的行为差异显著性最小,且可由传输层协议决定.

3.2 类间数量比例对ABSA方法的影响

ABSA 方法的本质是分析行为测度在协议间的分布差异,其间无类似于信息增益等方法所需计算的属于类别 S_v 的样例占原始样例 S 的比例$|S_v|/|S|$.因此理论上来说,只要样本所表示的协议行为没有偏差,ABSA 方法所得出的χ^2 统计和 *BS* 值就应保持稳定,与该协议在总体流量中所占比例无关.

为验证该结论,本节改变 Trace 1 中 eDonkey 协议流数在总流数中的比例,并进一步计算各种情况下的 *BS* 值.具体而言,根据 eDonkey 流数的原始样本比例(16.7%,由表 1 可得),当实验所需样本比例小于原始比例时,随机删除 Trace 1 中的 eDonkey 流;反之,随机删除其中的非 eDonkey 流.不同样本比例环境下的 eDonkey 协议测度 *BS* 值见表 3(限于篇幅,仅以部分测度为例,其余测度情况与此类似).

由表 3 可以看出,当 eDonkey 流数量在总体流量中所占比例在原始样本比例周围变动时(5%~50%),各测度 *BS* 值基本稳定.但当协议流数量变化较大时,*BS* 的计算值出现明显偏差.根据以上分析我们猜测,这样的偏差并不是由于类间数量比例变化造成,而是由于 eDonkey 协议过少的流数量(1%,约 5 万条)不足以体现原先应有的

分布,或较少的非 eDonkey 流(20%,约 25 万条)不足以表现复杂的网络总体流量分布而造成的.

Table 3 *BS* value vs. application sample size ratio (Trace1)

表 3 议样本数量比例对行为特征 *BS* 值的影响(Trace1)

Metrics	Flow percentage					
	1%	5%	10%	25%	50%	80%
pkt_size	1.1E+5	8.7E+4	8.9E+4	8.8E+4	9.2E+4	7.1E+4
bytes	5.7E+4	6.3E+4	6.3E+4	6.3E+4	6.4E+4	5.5E+4
Bps	3.9E+3	3.2E+3	3.4E+3	3.1E+3	3.1E+3	4.7E+3
head_size	4.5E+2	5.2E+2	4.9E+2	5.0E+2	5.1E+2	4.6E+2

对此情况,我们使用与 Trace 1 时间相邻的另一 Trace 对样本数量进行补充(称为 Trace 2,采集时间为 Trace 1 之后的 1 小时 16:00~17:00).扩充后,Trace 的 *BS* 计算结果见表 4.与表 3 相比,虽然 eDonkey 协议流数比例仍为 1%,但所包含的样本数量增加后(约 12 万条),协议行为分布偏差减小,测度 *BS* 值仍保持稳定.同理可知,表 4 中 *BS* 值在协议流数比例为 80%处回复稳定的原因.

Table 4 *BS* value vs. application sample size ratio (Trace 1+Trace2)

表 4 应用协议样本数量比例对行为特征 *BS* 值的影响(Trace 1+Trace2)

Metrics	Flow percentage					
	1%	5%	10%	25%	50%	80%
pkt_size	8.8E+4	8.8E+4	8.9E+4	8.9E+4	8.8E+4	8.7E+4
bytes	6.7E+4	6.7E+4	6.5E+4	6.5E+4	6.4E+4	6.7E+4
Bps	3.3E+3	3.2E+3	3.2E+3	3.1E+3	3.3E+3	2.9E+3
head_size	4.9E+2	5.1E+2	5.0E+2	5.1E+2	5.2E+2	5.3E+2

由此可知,只要样本数量足以表示协议行为的实际分布,不产生偏差,ABSA 方法的分析结果就与该协议流量的类间比例无关.因此,实际网络流量的采集与分析过程可简化于仅满足待分析协议的样本数量要求,而无须完全依照实际环境中的协议类间数量比例.

3.3 ABSA在协议识别的应用

与已有的特征选择方法相比,ABSA 可为协议识别提供更为准确和丰富的信息,优化识别结果.文献[5]中所用的 Naïve Bayes 方法是目前应用协议识别问题中准确率较高的方法.为了体现 ABSA 方法的优势,本节采用与其相同的识别方法、实验 Trace 和测试规则;唯一区别是将其所使用的 FCBF 特征选择算法替换为 ABSA,且仅简单地选取各协议中 *BS* 值最大的 5 个特征测度组成 Bayes 分类方法的输入测度集合.FCBF 的特征重要性排序前 12 位和 ABSA 所选的特征测度集合对比见表 5(测度名具体含义请参见文献[19]).

Table 5 Comparison of metric selection results between FCBF and ABSA

表 5 FCBF 和 ABSA 方法的测度选择结果比较

FCBF	ABSA	FCBF	ABSA
Server port	Server port	*var_data_wire_b a*	*var_data_wire_b a*
pushed_data_pkts_b a	*pushed_data_pkts_b a*	*min_segm_size_a b*	*throughput_b a*
initial_window-bytes_a b	*q1_data_ip*	*RTT_samples_a b*	*RTT_samples_a b*
initial_window-bytes_b a	*idletime_max_b a*	*pushed_data_pkts_a b*	*pushed_data_pkts_a b*
avg_segm_size_b a	*avg_segm_size_b a*	—	*Time_spent_in_bulk*
med_data_ip_a b	*q3_data_ip_a b*	—	*med_IAT_b a*
actual_data_pkts_a b	*actual_data_pkts_a b*	—	—

使用表 5 中 ABSA 的所选测度集合后,文献[5]中所述的应用识别方法准确率见表 6.

由表 6 可以看出,依赖 ABSA 方法的协议识别准确率高于原 FCBF 方法所得的结果,特别是对于流数较少的应用类别(如 P2P,Interactive),准确率提高更为明显.这表明,ABSA 方法所选择的协议特征测度可以更优地体现协议间的流量行为差异,并比现有方法更多地关注于小样本类别,从而为应用协议识别方法提供更多分类依据信息,提高协议识别的准确率.

Table 6 Comparison of accuracy after using FCBF and ABSA

表 6 应用 FCBF 和 ABSA 算法后的识别准确率比较

Category	Accuracy (%)		Category	Accuracy (%)		Category	Accuracy (%)	
	FCBF	ABSA		FCBF	ABSA		FACF	ABSA
BULK	82.25	88.38	ATTACK	13.46	34.01	INTERACTIVE	0	38.18
WWW	99.27	99.20	DATABASE	86.91	97.65	GAMES	0	50.00
MAIL	94.78	94.80	MULTIMEDIA	80.75	96.97	—	—	—
P2P	36.45	64.17	SERVICES	63.68	84.79	Total	96.29	98.15

本节实验中,ABSA 所选测度集合的确定只是简单地组合各协议行为特征测度的前 5 位.显然,这不是最好的特征测度选择方法,而且有可能存在多个协议较网络总体流量显著特征一致的情况.因此,如何根据各协议行为特征的情况组合出最优的特征测度集合,如何根据识别的准确率结果动态地调整测度集合以及如何使用 ABSA 方法分析协议间的测度分布差异,从而更细致地精化识别结果,是下一步继续研究的重点.

4 抽样对 ABSA 方法的影响

本文的前述工作是建立在采集到所有报文,对行为分布进行无偏分析的基础上.然而为了减少资源消耗,目前各高端路由器中的流信息统计系统都采用了报文抽样策略[9,20].若能发现报文抽样对 χ^2 统计结果和 ABSA 方法的影响,则可望将 ABSA 方法直接运用于 NetFlow 等从路由器采集的流记录,极大地提高该方法的实用性.本节以此为目的,分析目前路由器常用的报文系统抽样策略(文献[10,21]表明,在统计意义下,系统抽样等同于随机抽样,但实现代价更小)对协议行为分布和差异的影响,以及进而对 χ^2 统计和 ABSA 方法的影响.

报文抽样对协议行为分布的影响主要包括两个方面:报文抽样对时间维上流传输行为分布的影响和报文抽样所间接造成的流抽样对空间维上主机行为分布的影响.

前者典型如流长(流内报文数)分布.对其抽样前后的分布差异情况,有如下定理:

定理 1. 若流长在两样本空间中的原始分布无差异,则以任意抽样比抽样后的分布仍无差异.

证明:令流长抽样前的原始分布为 $P_b(Len)$,抽样比为 p,则抽样后的流长分布 $P_a(Len)$为

$$P_a(Len=l)=\sum_{i=l}^{\infty}P_b(Len=i)C_i^l p^l(1-p)^{i-l}.$$

已知两原始分布无差异,即对任意的 i,两样本空间中的 $P_b(Len=i)$均相同;又 C_i^l,p^l,$(1-p)^{i-l}$ 均为常数.故对任意的 l,两样本空间中 $P_a(Len=l)$相同,即抽样后分布仍无差异. □

定理 1 表明,若在未抽样时 ABSA 方法判流长不能作为特征,则以任意抽样比抽样后,χ^2 统计量和 BS 值仍应维持在接近 0 的水平,ABSA 方法仍然可以准确判断其不可作为特征.

定理 2. 若流长在两样本空间中的原始分布存在差异,则抽样后的分布差异较原始分布差异显著程度变小,χ^2 统计量及 BS 值变小.

证明:对每条单独的流,其被抽中的概率为

$$P_{flow}(Len=i)=1-(1-p)^i \tag{5}$$

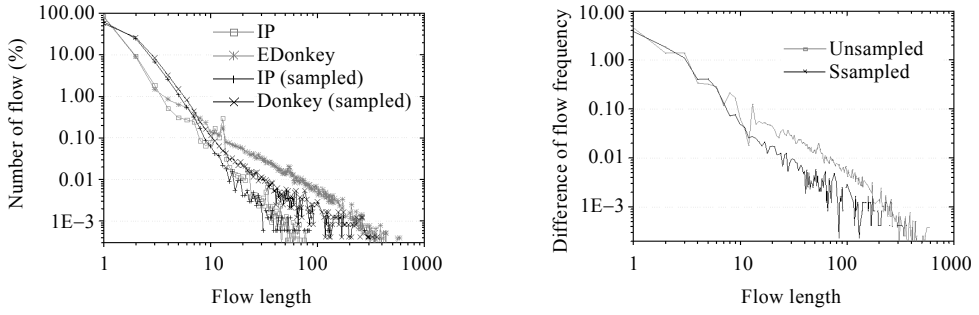
虽然流长不同流被抽中的概率不同,但当 i 取定时,此处流抽样比为定值,因此:

① 若两原始分布在 $Len=i$ 处有差异,则抽样过程中流数比例较多的样本在此处会被抽去占其总数比例较多的流;

② 若两原始分布在 $Len=i$ 处无差异,则抽样过程中两样本在此处被抽去同样比例的流.

综合①、②可知,抽样后分布差异变小,即 χ^2 值变小;且原分布差异越大,变化情况越明显. □

图 6 所示为抽样比 1/10 条件下,*pkts* 测度在抽样前后 TRACE eDonkey 和 TRACE Non_eDonkey 中的分布及差异情况.图 6(a)为在抽样前后测度实际分布情况,图 6(b)为抽样前后各自分布频率之差的绝对值曲线.由图 6 可以看出,除了极少数点以外,抽样后的分布差异均小于抽样前,总体分布差异情况变小,进一步表明定理 2 的正确性.



(a) Distribution of *pkts* before and after packet sampling
(a) 报文抽样前后 *pkts* 分布

(b) Difference distribution of *pkts* before and after packet sampling
(b) 报文抽样前后 *pkts* 差异分布

Fig.6 Comparison between distribution difference of metric *pkts* before and after packet samping
图 6 报文抽样前后 *pkts* 测度分布差异对比

推论 1. 若流长在两样本空间中的原始分布存在差异,则抽样比越大,抽样后分布差异显著程度越小,χ^2 统计量及 *BS* 值越小.

证明:令有两抽样比 $p_1,p_2,1>p_1>p_2>0$,即 p_2 所代表的抽样比大于 p_1 所代表的抽样比,则$\exists p=p_2/p_1,0<p<1$,使得 $p_2=p_1\times p$.

使用 p_2 进行抽样的结果等同于分别使用 p_1 和 p 进行连续抽样后的结果,即使用 p_2 的抽样结果等同于在 p_1 的抽样结果基础上再使用 p 进行抽样.故由定理 2 可知, $BS_{p_2}=BS_{p_1\times p}<BS_{p_1}<BS_{unsampled}$. □

推论 2. 存在抽样比阈值 p_0,当实际抽样比 $p<p_0$ 时,抽样后流长分布统计无差异.

证明:由公式(5)可知,当 $p=0$ 时,流抽样比 $P_{flow}=0$,此时测度分布差异为 0,$\chi^2=0$.

结合推论 1 可知,当 $p\to 0^+$时,分布差异$\to 0^+$,$\chi^2\to 0$.

故由极限存在定义可知,对$\forall\delta>0,\exists p_0>0$,当 $0<p<p_0$ 时,恒有$(\chi^2-0)<\delta$.

现取$\delta=\chi^2_\alpha$,则当 $p<p_0$ 时,有$\chi^2<\chi^2_\alpha$,即分布统计无差异. □

其他时间维行为测度的分析与上述 *pkts* 测度类似.需要特别说明的是平均报文长度和各比例类型的测度(如流内双向报文数比):文献[10]表明,报文抽样不改变网络总体报文大小分布特征.然而该研究是针对于网络所有报文的分布,而不是每流的平均报文长度.由于网络总报文数很大,概率抽样不会产生分布偏差;但每流内报文数较少,流内各长度的报文是否被抽中存在一定的偶然性,造成流内平均报文长度的改变.同理,虽然报文抽样独立于报文方向,但由于流内报文数较少,结果分布仍会改变,并且较原始分布差异显著程度变小.对于空间维测度,若将主机之间协议交互的流量规约为广义上的流,则吞吐量类(主机上某协议传输的报文数、字节数)、比例类的行为测度差异分布及χ^2 统计情况类似于相应时间维流测度差异分析.同时,若将链接数看作流内报文数,流抽样比看作相应的报文抽样比,则对主机协议链接数分布的分析也类似于上文中的流长分析.定理 1 和定理 2 的结论仍成立.

定理 2 及其推论表明,抽样比存在一个阈值,当实际抽样比大于该阈值时,原先在某协议间存在分布差异的测度变成统计无差异,该测度就不能再为该协议识别提供任何有用信息,不能再作为该协议的特征,为任何协议识别方法所用.随着抽样比的增大,将不断有各协议间的测度进入分布无差异的集合中,协议特征不断减少,应用识别可用的测度不断减少.这样的阈值随协议和测度的不同而不同.但在任意的抽样比环境下,仍可保证测度分布差异的显著程度相对顺序不会发生改变,见定理 3.

定理 3. 对任意抽样比 $p>0$,各行为测度使用同一 p 抽样前后的分布差异相对显著程度顺序均不变,χ^2 值和 *BS* 值相对大小顺序不变.

在证明定理 3 之前,先提出以下引理:

引理 1. 存在$\varepsilon>0$,使得当抽样比$p\in(1-\varepsilon,1)$时,测度M抽样后的总体分布差异较原始差异在统计意义下最多减少 1 个样本单位.

证明:因为当$p\to1^-$,分布差异变化量$\to0^+$,

故由左极限存在定义,可知对$\forall\delta>0,\exists\varepsilon>0$,使当$0<1-p<\varepsilon$时,恒有(分布差异变化量−0)$<\delta$.

现取δ=1/样本总数,则引理 1 成立. □

由此可得定理 3 的证明.

证明:令有测度 M_1,M_2,不妨设未抽样时 M_1 分布差异显著程度大于 M_2 分布差异显著程度,即 $BS(M_1)>BS(M_2)$.抽样比为 p,$0<p\leqslant1$.

由引理 1 知,$\exists\varepsilon_1>0$,使当抽样比落入区间$(1-\varepsilon_1,1)$时,M_1 分布差异变化至多减少 1 个单位;且$\exists\varepsilon_2>0$,使当抽样比落入区间$(1-\varepsilon_2,1)$时,M_2 分布差异变化最多减少 1 个单位.取$\varepsilon=\min(\varepsilon_1,\varepsilon_2)$.

若$p\in(1-\varepsilon,1)$,则显然抽样后 M_1 的分布差异显著程度仍大于等于 M_2 的分布差异显著程度,$BS(M_1)\geqslant BS(M_2)$,定理成立.

若$p\leqslant1-\varepsilon$,假设采用 p 进行某次抽样后,分布差异相对显著程度发生改变,即 $BS(M_1)<BS(M_2)$.

分解$p=p_{11}\times p_{12}$($p<p_{11}<1$,$p<p_{12}<1$),则使用 p 进行抽样的结果等同于分别使用 p_{11} 和 p_{12} 进行连续抽样的结果,同时,有且仅有 1 次抽样前后 BS 值大小顺序发生改变,不妨设为使用 p_{11} 抽样时.同理,若 $p_{11}\in(1-\varepsilon,1]$,由引理 1,应有抽样前后 $BS(M_1)>BS(M_2)$,BS 值大小顺序不变,假设不成立,差异相对显著程度应不变.

若$p_{11}\leqslant1-\varepsilon$,继续分解$p_{11}$,令使抽样前后$BS$值大小顺序发生改变的抽样比为$p_{i1}$,由于$p<p_{11}<p_{21}<...<p_{i1}<...<1$,有 $p_{i1}\to1^-$,则必存在某个 $p_{n1}\in(1-\varepsilon,1]$,使 $BS(M_1)>BS(M_2)$,与假设矛盾,故抽样前后差异相对显著程度应不变.

由于 p 可在(0,1)中取任意值,故在任意抽样比例下,上述证明均成立. □

定理 3 表明,测度分布差异的显著程度顺序与抽样比无关,ABSA 方法在某抽样比下分析所得的分布差异显著程度顺序同样适用于其他抽样比;其保证了在不同抽样比环境下特征的选择顺序唯一.同时结合定理 2,定理 3 还保证了当抽样比变化时,测度进入或离开分布无差异集合的顺序唯一,即按照 BS 值的大小顺序.因此,当抽样比变化时,无须完全重新检查每一个测度的现分布差异情况并重新选择协议特征,仅需在当前特征测度集合基础上进行按 BS 值序的扩充或删减,极大地简化了抽样比变化时协议特征的重新选择过程.

表 7 为报文抽样比为 1/10,1/100 和 1/1000 时,TRACE eDonkey 和 TRACE Non_eDonkey 的各测度分布差异 BS 值.对比表 2 和表 7 中各项数值可验证各定理和推论的正确性,以及对抽样后测度分布差异和 BS 值变化情况的分析过程.从而可得出以下结论:

(1) 若某行为测度分布抽样前无差异,则抽样后仍无差异,ABSA 方法仍判断其不可作为特征.

(2) 若某测度分布抽样前有差异,则抽样后差异程度变小,BS 值变小;抽样比越大,BS 值越小.且存在抽样比阈值,当实际抽样比大于该阈值时,分布统计无差异,测度不可再作为该协议的特征.阈值随测度和协议的不同而不同.

(3) 不同抽样比环境下,ABSA 方法保证各行为测度的相对特征显著程度顺序不变,协议行为特征的选择顺序与抽样比无关.

Table 7 *BS* value of some behavior metrics after packet sampling

表 7 部分行为特征在报文抽样环境下的 *BS* 值

Metrics	Sampling rate			Metrics	Sampling rate		
	1/10	1/100	1/1000		1/10	1/100	1/1000
TCPflags	4.3E+4	9.3E+3	1.8E+3	*pps*	7.7E+2	3.2E+2	31.4
pkt_size	1.2E+4	1.9E+3	2.3E+2	*pkt_size_ratio*	7.6E+2	2.4E+2	30.2
bytes	8.7E+3	1.3E+3	1.8E+2	*bytes_ratio*	7.6E+2	2.1E+2	26.1
pkts	6.3E+3	6.3E+2	1.3E+2	*pkts_ratio*	4.4E+2	14.3	12.7
duration	7.9E+2	5.4E+2	44.8	*head_size*	1.5E+2	5.0	1.1
Bps	7.7E+2	3.5E+2	38.5	—	—	—	—

5 结论和下一步的工作

目前,基于行为的网络流量应用层协议识别方法在整体精度或粒度方面达不到令人满意的效果,且缺乏通用性,其主要原因是缺乏对各应用协议行为特征的认识.因此,本文提出了一种判断某行为测度是否可作为网络应用协议的行为特征,并对该行为特征的显著程度进行评估的方法——ABSA.该方法以数理统计理论中的χ^2统计检验为基础,使用趋向变量分布的区间随机划分启发式算法,保证χ^2统计量的稳定性;使用多自由度间投票判断方式,消除自由度的不确定性;使用某置信度水平下各自由度的临界分位点对χ^2统计量进行均化,消除统计量随自由度的增长对衡量差异造成的不利影响,保证各自由度下χ^2统计量所占权重一致,从而可作为准确评估行为测度分布差异程度的标准.文章还同时分析了 ABSA 方法在报文抽样环境下的适用性.

理论分析及实验结果表明,若某行为测度在协议间存在分布差异,ABSA 方法可准确判其作为协议的行为特征,并根据分布的差异情况合理地评估该行为特征的显著程度,从而为分类方法提供了更丰富、更准确的信息,优化分类结果;当协议样本可准确表达协议行为时,ABSA 方法可保持特征显著程度的稳定性,与协议样本在总样本中所占的比例无关;在报文抽样环境下,各测度分布实际差异程度变小,但方法仍保持特征相对显著程度顺序评估结果不变,保证了抽样比变化时特征选择顺序的唯一性,简化了特征重选择过程.ABSA 方法可与 NetFlow 等通用流信息统计系统配合,或用于各类以流记录为输入的网络监测系统中,发现各应用协议的行为差异和特征,为应用协议识别、网络测量管理、保障 QoS 的实施提供基础.

下一步的工作主要是借助 ABSA 方法分析各主要应用协议所具有的行为特征及相对的重要程度,研究如何根据各协议较网络总流量的特征和其间的行为差异进行合理的特征组合及动态调整,进一步优化目前识别算法的精度和粒度.

References:

[1] Moore AW, Papagiannaki K. Toward the accurate identification of network applications. In: Proc. of the PAM 2005. 2005. 41−54.

[2] Kim MS, Won YJ, Hong JWK. Application-Level traffic monitoring and an analysis on IP networks. ETRI Journal, 2005,27(11): 22−42. [doi: 10.4218/etrij.05.0104.0040]

[3] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark. In: Proc. of the ACM SIGCOMM 2005. 2005. 229−240.

[4] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms. In: Proc. of the ACM SIGCOMM 2006. 2006. 281−286.

[5] Moore AW, Zuev D. Internet traffic classification using Bayesian analysis techniques. In: Proc. of the ACM SIGMETRICS 2005. 2005. 50−60.

[6] Auld T, Moore AW, Gull SF. Bayesian neural networks for Internet traffic classification. IEEE Trans. on Neural Networks, 2007, 18(1):223−239. [doi: 10.1109/TNN.2006.883010]

[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] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed., San Francisco: Morgan Kaufmann Publishers, 2005.

[9] Cisco. Cisco IOS NetFlow introduction. 2006. http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html

[10] Claffy KC. Internet traffic characterization [Ph.D. Thesis]. San Diego: University of California, 1994.

[11] Pitkow J. Summary of WWW characterizations. World Wide Web, 1999,2(2):3−13. [doi: 10.1023/A:1019284202914]

[12] Dewes C, Wichmann A, Feldmann A. An analysis of Internet chat systems. In: Proc. of the ACM SIGCOMM IMC 2003. 2003. 51−64.

[13] Plissonneau L, Costeux JL, Brown P. Analysis of peer-to-peer traffic on ADSL. In: Proc. of the 6th Annual Passive and Active Measurements Workshop (PAM 2005). 2005. 69−82.

[14] Schneider F, Agarwal S, Alpcan T, Feldmann A. The new Web: Characterizing AJAX traffic. In: Proc. of the PAM 2008. 2008. 31−40.

[15] Cao ZH, Zhao P, Hu YQ. The Theory of Probability and Mathematical Statistic. Nanjing: Southeast University Press, 2003 (in Chinese).

[16] Guangdong University of Business Studies. Structural equation model. 2008 (in Chinese). http://jljjx.jpkc.gdcc.edu.cn/show.aspx?id=506&cid=24

[17] Quadong, Sommere. SourceForge.net: Linux layer 7 packet classifier. 2009. http://sourceforge.net/projects/l7-filter

[18] Zhang W, Hou LD. P2P traffic identification based on transport layer flags. Science & Technology Information, 2007,1:169−170 (in Chinese with English abstract).

[19] Moore AW, Zuev D. Discriminators for Use in Flow-Based Classification. London: Intel Research, University of Cambridge, 2005.

[20] Huawei Technologies Co., Ltd. NetStream Technology White Paper. 2007 (in Chinese). http://www.huawei.com/cn/products/datacomm/pdf/view.do?f=269

[21] Duffield N, Lund C, Thorup M. Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans. on Networking, 2005, 13(5):933−946. [doi: 10.1109/TNET.2005.852874]

附中文参考文献:

[15] 曹振华,赵平,胡跃清.概率论与数理统计.南京:东南大学出版社,2003.

[16] 广东商学院.结构方程式模型.2008. http://jljjx.jpkc.gdcc.edu.cn/show.aspx?id=506&cid=24

[18] 张文,侯立东.基于传输层标志位的 P2P 流量识别技术.科技资讯,2007,1:169−170.

[20] 华为技术有限公司.NetStream 技术白皮书.2007. http://www.huawei.com/cn/products/datacomm/pdf/view.do?f=269

陈亮(1981－),男,江苏南京人,博士生,主要研究领域为网络行为学.

龚俭(1957－),男,博士,教授,博士生导师,主要研究领域为网络行为学,网络安全,网络管理.