

基于子空间聚类的网络流量分类方法

李丹丹^{1,2}, 田春伟^{1,2}, 李佰洋^{1,2}, 孙广路^{1,2}, 康健³

(1. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150080; 2. 哈尔滨理工大学
信息安全与智能技术研究中心, 黑龙江 哈尔滨 150080; 3. 北京宇航系统工程研究所, 北京 100076)

摘要: 应用层网络流量分类技术对流量控制与管理等研究具有重要意义. 针对传统的基于有监督机器学习的分类方法对所有应用程序使用相同的特征, 使得某些特征对一种或几种应用类型有区分性, 而对其他应用类型的网络流分类产生干扰等问题, 提出基于子空间聚类方法的网络流分类框架. 利用子空间聚类算法, 在总特征集中为每一种类型应用进行特征选择, 提取与之相对应的关键特征, 自动消除不相关的特征, 使得每种应用类型都产生对应的特征签名集, 并用这些不同的特征签名对未知的网络流进行分类. 实验结果表明: 本文提出的方法能够有效地提出每种应用类型的特征签名, 并且所提出的特征签名具有明显的可区分性, 该方法的分类准确率在 93% 以上, 并且能很好的识别新出现的应用.

关键词: 子空间聚类; 网络流分类; 特征签名

DOI: 10.15938/j.jhust.2015.02.012

中图分类号: TM391.1 **文献标志码:** A **文章编号:** 1007-2683(2015)02-0063-06

Network Traffic Classification Based on Subspace Clustering-based Method

LI Dan-dan^{1,2}, TIAN Chun-wei^{1,2}, LI Bai-yang^{1,2}, SUN Guang-lu^{1,2}, KANG Jian³

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China; 2. Research Center of Information Security & Intelligent Technology, Harbin University of Science and Technology, Harbin 150080, China; 3. Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China)

Abstract: The application-layer network traffic classification is critical to network traffic control and management. In the traditional classification methods with supervised machine learning model, the same features are used to classify different applications. It results in that some features can perform well for some applications, but worse for others. This paper proposes a network traffic classification frame based on Subspace Clustering-based Method (SCM). SCM is utilized in feature selection extracts key features corresponding to each type of application and automatically eliminates irrelevant features. Then each type of application corresponds to its own signatures, which are applied to classify the network traffic in the same overall feature space. The experimental results show that SCM extracts corresponding feature signatures effectively for each application. The classification accuracy of the proposed method is more than 93%. And it identifies the new applications well.

Keywords: subspace clustering; network traffic classification; feature signatures

收稿日期: 2014-09-16

基金项目: 黑龙江省普通高等院校新世纪优秀人才培养计划(1155-ncet-008); 黑龙江省博士后科研启动基金; 黑龙江省教育科学规划课题(GBC1211062).

作者简介: 李丹丹(1989—), 女, 硕士研究生;
田春伟(1990—), 男, 硕士研究生.

通讯作者: 孙广路(1979—), 男, 博士, 教授, E-mail: sunguanglu@hrbust.edu.cn.

0 引言

近年来,具有开放性、共享性等特点的互联网迅速普及并发展壮大,随着网络用户数量的飞速增多,传统的互联网业务已经无法满足人们的需求,越来越多的新型网络应用应运而生.对于网络运营商来说,为了有效利用带宽,并提供更好的服务质量(quality of service, QoS),需要网络能够针对不同应用进行分类.另一方面,互联网的开放性特点允许任何符合其技术标准的设备或软件接入互联网,导致了各种网络恶意攻击事件层出不穷,严重危害网络信息安全与服务,保护网络安全迫在眉睫.网络流量分类技术能够增强网络可控性,帮助研究人员掌握网络上的流量分布情况,帮助网络运营商优化服务质量,预防并阻止各种网络犯罪行为^[1].总而言之,构建网络智能,网络流分类是第一步.

在流量分类研究中,目前广泛采用5元组{源IP、源端口、目的IP、目的端口、传输层协议}(Source IP, Source PORT, Destination IP, Destination PORT, Protocol)定义数据流.迄今为止,网络流量分类方法主要包括基于端口号匹配的分类方法、基于应用层载荷特征的分类方法^[2]、基于主机行为和交互行为的分类方法,以及基于机器学习的分类方法^[3-5].用于分类的特征主要有端口特征、载荷特征、网络流包级和流级统计特征、主机行为和交互行为特征(根据IP地址、端口等衍生出来的关联特征)^[6].

其中,传统的基于有监督机器学习的分类方法,主要是利用网络流在网络层和传输层的统计特征来识别流量^[7].这种方法使用的特征和训练得到的分类器对所有应用类型的流都是相同的,没有差异和针对性.某些特征对一种或几种应用类型有区分性,而对其他应用类型的流分类产生干扰,导致整体的分类准确率不高.而且,其对于网络数据中新出现协议的识别效果不能令人满意^[8].

针对以上问题,依据样本在不同的数据簇中常常与某些特定的数据特征子集相对应的^[9]思想,本文提出了一种新的网络流分类方法:基于子空间聚类的网络流分类方法(subspace clustering-based method, SCM).它针对每一种应用流独立建立识别模型,然后将这些识别单一应用的模型合在一起,实现多类别分类,而不是同时区分几种应用.换句话说,SCM方法学习独立的识别类A和识别类B,而不是试着同时区分A和B.本文使用骨干网络

数据集和边界网络数据集对SCM分类方法进行测试,识别精度达到93%,而且对新应用的识别率可以达到90%.由于路由非对称原因导致骨干网中双向流中的某一个方向的流经常丢失,本文中的分类方法使用单向流.

1 子空间聚类算法及网络流分类框架

1.1 子空间聚类算法

子空间聚类问题描述:

设 $B = B_1 \times B_2 \times \cdots \times B_n$ 代表一个 n 维数据空间,其中, n 为正整数. F 是包含 m 个位于 n 维特征空间的数据对象的集合,记为 $F = \{X_i | i \in [1, m]\}$, $x_{ij} = X_i \cdot B_j$,其中,点 $X_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{in})$, X_i 的第 j 个属性值 x_{ij} 为其在 B_j 维上的取值.设 k 维子空间 $S_k \subseteq B$ 其中, $k \leq n$.在 S_k 子空间中的元组集合表示为 $T_{S_k}(x_{i1}, x_{i2}, \cdots, x_{ik}) = \{X_i | X_i \in S_k, x_{ij} = X_i \cdot B_j | j \in [1, k]\}$.

在训练期间,SCM分类框架为每种应用类型产生一组签名.产生签名的过程如下:

- 1) 输入:属于同种应用类型的一组流的集合 F ,每个流 $f \in F$ 用 n 维数值向量表示.
- 2) 输出:流的子集合与特征子空间的对应关系 (F_i, S_i) 即签名,其中 $F_i \subseteq F, S_i \subseteq B$.每一种应用类型可能包含多个签名 $i \in [1, |F|]$.

本文使用文^[10]提出的基于 K 最相似聚类的子空间聚类算法,在得到所有一维空间聚类的基础上,查找包含相同数据对象的相似聚类,继而决定子空间搜索方向,然后合并不同维度中的相似聚类,最终得到子空间聚类.该方法用不同子空间中采用的不同密度阈值替代传统算法中的全局密度阈值^[11],提高了子空间聚类算法聚类结果的精度.

该算法的相关定义如下:

定义1(基本聚类) 将使用DBSCAN^[12]方法产生的所有一维空间的聚类的集合称为基本聚类,记为 C^1 .

定义2(基本聚类相似度) 给定 $c_1, c_2 \in C^1$,其中, c_1 在 B_i 维, c_2 在 B_j 维,且 $i \neq j$,其相似度定义为基本聚类 c_1, c_2 所包含的相同数据对象元素的数目,记为 $\text{sim}(c_1, c_2) = |c_1 \cap c_2|$.

定义3(最相似聚类) 给定聚类 $c \in C^1$, c 的最相似聚类 $MSC(c) \in C^1$ 满足条件:

$$\forall c_p \in C^1, \text{sim}(c, MSC(c)) \geq \text{sim}(c, c_p).$$

定义4(k 最相似聚类) 给定聚类 $c \in C^1$ 和正整

数 k, c 的第 k 个最相似聚类定义为 k 最相似聚类, 记为 $MSC_k(c)$. c 的第 k 个最相似聚类满足: $\forall c_p \in MSC_k(c), \forall c_q \in C^1 - MSC_1(c) - MSC_2(c) - \dots - MSC_k(c), \text{sim}(c, c_p) > \text{sim}(c, c_q)$.

该算法的基本步骤如下:

- 1) 对数据做预处理, 产生每一维上的基本聚类;
- 2) 计算所有基本聚类之间的相似度;
- 3) 查找每个基本聚类的 k 个最相似聚类, 当作该基本聚类的合并候选;
- 4) 使用深度优先方法搜索子空间, 产生子空间聚类;
- 5) 删除剩余数据噪声点.

先找到基本聚类 $c \in C^1$, 形成子空间 S_1 , 计算 C^1 中各个基本聚类间相似度, 通过得到的基本聚类相似度查找每个基本聚类的 k 个最相似聚类, 合并基本聚类相似度大于等于 $\text{para}(S_2)$ 的基本聚类, 形成 S_2 子空间. 然后不断的递归的搜索, 产生子空间聚类.

其中 $\text{para}(S_m)$ 为 S_m 子空间合并阈值, 利用如下公式来获得:

$$\text{para}(S_m) = r \times E [T_{S_m}(c_1, c_2, \dots, c_m)] \quad (1)$$

$E [T_{S_m}(c_1, c_2, \dots, c_m)]$ 就是 S_m 子空间期望, 它是由总的学习样本数 $|F|$ 比上 S_m 子空间投影到每一维产生的不同元组的数目得到的. 算法的详细内容参见 [8].

所有的特征都使用标准的 Z -score 算法进行标准化, 该公式为

$$x_{\text{norm}} = (x - \mu) / \sigma, \quad (2)$$

其中 μ 和 σ 分别是通过训练数据算出来的特征的平均值和标准差.

2.2 SCM 网络流分类框架

基于子空间聚类算法, 本文提出一种新的网络流分类框架 SCM. SCM 执行两个不同的过程: (a) 产生签名 (b) 网络流分类. 具体的分类示意图如图 1.

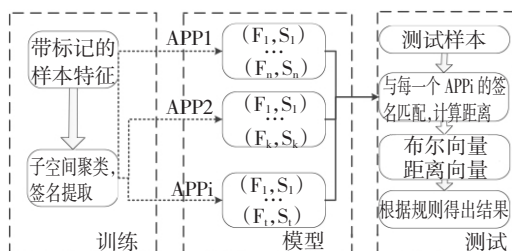


图1 SCM 网络流分类示意图

分类过程: 在分类期间, 一个待分类流需要和所有应用类型各自对应的签名模式做匹配. 实质上, 每一个签名模式都是一个二分类器, 只返回是或否. 下面是分类的详细过程.

匹配一个签名模式: 当一个待分类流匹配某个特定的签名 (F_i, S_i) 时, 首先它投影到 S_i 维空间中 (提取对应特征集), 然后用标准“欧几里得距离”来计算它与 F_i 中的模式流之间的距离 [13]. 本文把待分类流到 F_i 中最近的流的距离作为待分类流到签名的距离 (即距离的最小值). 如果待分类流在该签名的预定半径 r 范围内, 则匹配成功. 实质上, 每一个签名都是一个有 $|S_i|$ 维特征、 $|F_i|$ 个点和固定半径的最近邻分类器. 这个固定的半径保证签名只匹配属于此应用类型的流而不会匹配属于其他应用类型的流.

根据欧几里得公式:

$$d = \sqrt{\sum_{k=1}^{|S_i|} (x_k - y_k)^2}. \quad (3)$$

可以看到, 即使每一维度的距离都很小, 距离 d 还随维度 $|S_i|$ 升高而递增, 而不同子空间的维度可能不相同, 所以对所有的子空间都用同一个固定半径不准确. 本文用基本的缩放因子 $\sqrt{|S_i|}$ 来调节 [14]. 这样, 如果在我们的分类器中使用的一维半径是 r , 那么对于签名 (F_i, S_i) ,

$$r_i = \sqrt{|S_i|} \cdot r. \quad (4)$$

在签名的所有点半径 r_i 的覆盖范围内我们都认为是有效的.

流分类过程: 本文中的 SCM 分类框架包括很多二分类器. 每一个二分类器对应一个应用签名 (F_i, S_i) . 假设现在得到 t 个二分类器:

$$Y = \{y_1, y_2, \dots, y_t\}. \quad (5)$$

每一个进入 SCM 分类框架的流都要被每一个二分类器处理. 每一个二分类器都要返回 true 或 false 和距离 d . 因此, 结果是两个向量:

- 1) 一个 t 维的布尔向量

$$L = \{l_1, l_2, \dots, l_t\}. \quad (6)$$

其中变量表示二分类器 i 的返回值. 比如, 如果 X_i 给流标记为真, $l_i = 1$, 否则, $l_i = 0$.

- 2) 一个 n 维的数值向量:

$$D = \{d_1, d_2, \dots, d_t\}. \quad (7)$$

其中 d_i 表示该待分类流到签名 (F_i, S_i) 之间的距离. 由于属于一个应用类型的待分类流可能和多个签名有关, 它可能被映射到多个二分类器中. 为了跟

踪二分类器和应用类型的映射关系(在训练中得到),我们引进了一个新的向量 M ,如果二分类器 i 来自应用 App ,则 $M(i) = App$,可能会有多个 i 对应一个应用 App .

最后,测试流是否被标记为 App 或 $Unknown$ 由下面的算法决定,这个算法用到向量 L 、 D 和 M :

添加所有 $L(i) = \text{true}$ 的 i 值到结果集 R 中;

如果 $|R| = 0$,则将该待分类流标记为 $\text{label} = \text{Unknown}$,因为没有分类器标记这个流;

如果 $|R| = 1, R = \{k\}$ 则将此流标记为 $M(k)$,因为只有一个二分类器标记了这个流如果 $|R| > 1$,且 $\forall i \in R, |S_i| < |S_k|$,则将此流标记为 $M(k)$. 即如果多个分类器同时标记此流,则选择签名维度更高的分类器做标记,因为维度越高,特性越明显.

2 实验结果和分析

2.1 数据集

本文使用3份真实流量对SCM子空间聚类方法进行实验验证.这3份数据一份是骨干网络数据和两份是不同的校园网数据.分别收集于2010年与2012年.

本文使用基于载荷的DPI分类技术^[15-18]对网络流进行标记.使用的数据集中包括如下几种应用:Web,Streaming,P2P,Chat,Dns,Games,Mail和其他类型,其中未标记的流约占20%.实验时将这些未识别的流去除掉.图2描述了3份网络数据集中应用层应用类型的分布情况.

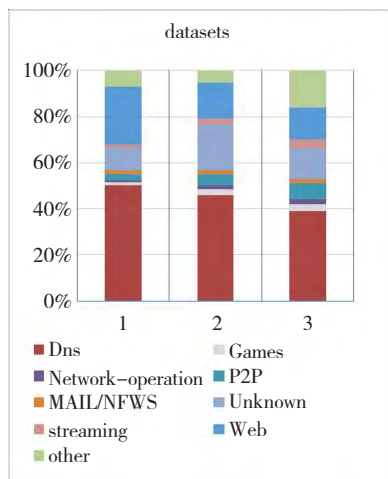


图2 应用类型分布图

2.2 特征

本文使用了网络流包级、流级的主要特征,包括

流的前6个包的大小、内部时间间隔(IAT)、总字节数、总包数和IAT的最大值,最小值,平均值、标准差^[7]等等.

表1 网络流特征

特征编号	特征描述	特征编号	特征描述
1~6	前6个数据包大小	13	平均包到达时间间隔
7	最大数据包大小	14	数据流字节总数
8	最小数据包大小	15	数据流的持续时间
9	数据包大小平均值	16	PSH个数
10	数据包大小方差	17	URG个数
11	最大包到达时间间隔	18	RESET个数
12	最小包到达时间间隔	19~35	前16个载荷字节对应值

还包含了每个流的前16字节载荷.每个载荷字节都用 $\{0,1,\dots,255\}$ 之间的数值表示.据研究,应用层协议头在流中会重现,而载荷的前16字节会包括应用层协议头.所以载荷的前16字节也是重要的特征^[18].每个特征用一个流的特征向量的一维来表示.使用的详细特征见表1.

2.3 实验理论和评估指标

为了评估SCM方法的性能,本文把每一个数据集分成两部分.用其中一部分做训练,用另一部分做测试.训练分类器时,本文只考虑实际流数超过1000个的应用类型.这个数对于收集训练数据不难,而且足够用于抽取好的签名.本文只对单向流产生签名,这样可以分别从客户端到服务器端的交互和从服务器端到客户端的交互中抽取不同的签名.当用对比基准与SCM分类预测结果对照时,需要同时看流的两个方向上所给的标签.如果一个方向标记为未知流,反方向标记为某种已知应用类型,则用反方向标签标记该流.如果两个方向标记的标签不同,则将此流标记为未知.文中所有的实验结果均是对多次实验结果取得的平均值.

本文使用覆盖率(coverage)、覆盖集识别准确率(accuracy)以及新应用检测率(newApp)等评价方法^[8]来评价SCM分类器的分类性能.

覆盖率:实际做出预测的样本的数量(没被标记为未知)除以样本总数.

覆盖集的准确率是被正确分类的样本的数量除以预测样本总数.即:

$$\text{Accuracy} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)} \quad (8)$$

其中,TP 表示真正(true positives),FP 表示正错误(false positives).具体定义如下:

1) 真正: 指待测方法将 P2P 流识别为 P2P 流的个数;

2) 正错误: 指待测方法将非 P2P 的流错误识别为 P2P 流的个数;

新应用的检测率:

$$newApp_i = \frac{Unknown_i}{Total_i} \quad (9)$$

当 SCM 分类框架对一个训练数据集中不存在的应用类型的流分类时,理想的结果是这个流被标记为未知.例如,如果训练得到的分类器中没有 Games 类型的签名,那么分类器应该将所有测试数据中的 Games 流标记为未知.由于训练的签名中不存在 Games 类型,Games 类型就代表了一种新的应用.本文暂时删除训练数据中的一种应用类型的流,然后观察分类器如何标记这些“新”应用类型的流.

2.4 实验结果及分析

本文在不同配置参数下测试了 SCM 方法的性能.先用一个数据来测试不同的配置,确定参数供实验中其他数据使用.本文使用 DBSCAN 算法来产生基簇,DBSCAN 有两个参数,流间允许的最大距离 ε 和最小点数 minPts.一维特征空间的有效区域的半径 r 也设为与 ε 相同的值 $r = \varepsilon$.通过删除低维签名可以调控分类器的准确率.经过实验,我们得到 μ 为 3,minClu 为 4,MSCG(c) 集合大小为 4 时子空间聚类算法的聚类效果最好.

为了研究参数敏感性,本文大范围的变换 ε , minPts 的值来观察实验结果.实验结果表明,参数的变化对我们分类器的分类效果影响很大.本文中使用的默认参数是 minPts = 10.这些参数值是测试清华骨干网数据得到的,然后对其余 2 个数据集用与这个分类器同样的配置.图 3 展示了 SCM 分类器在 3 个数据集上的分类表现.虽然使用通过 edge1 数据得到的分类器配置,但是在 3 个数据集上的分类效果都很好. Accuracy 可达到 93%,与朴素贝叶斯分类器^[9]的分类结果对比如图 3.在调整好参数的情况下,SCM 的准确率优于朴素贝叶斯分类器.

SCM 在识别新应用方面表现也很好,本文对每种应用类型都做一次新应用测试,然后用所有应用类型各自得到的 newApp 求平均值. newApp 的平均值可达到 90%,coverage 可达到 75%,如图 4 所示.

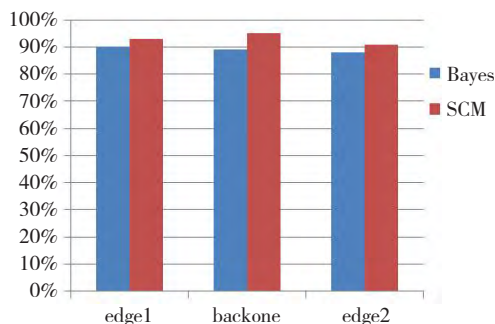


图3 SCM与Bayes分类准确率比较

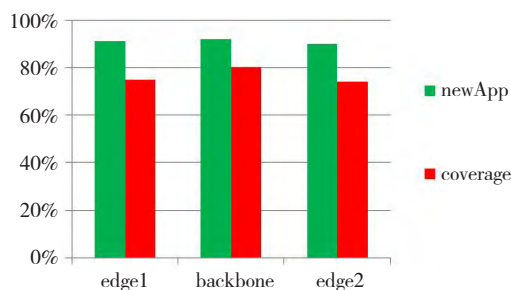


图4 SCM对新应用的识别率

3 结 语

本文将子空间聚类方法应用于应用层流量分类中,提出 SCM 网络流分类框架.在训练过程中,为每种应用类型产生多个流与子特征空间的签名对即二分类器,用于识别单一应用类型的应用.又将多个二分类器组合在一起实现多类别分类. SCM 能够有效的对每个应用类型抽取出相应的特征签名,并且所提出的特征签名具有明显的可区分性.该方法的分类准确率能达到 93% 以上,超过了朴素贝叶斯分类器,并且能很好的对新出现的应用进行识别.

参 考 文 献:

- [1] 柏骏,夏靖波,吴吉祥,等. 实时网络流量分类研究综述 [J]. 计算机科学, 2013, 40(9): 8-15.
- [2] FINSTERBUSCH M, RICHTER C, ROCHA E, et al. A Survey of Payload-Based Traffic Classification Approaches [J]. Communications Surveys & Tutorials, IEEE, 2014, 16(2): 1135-1156.
- [3] 熊刚,孟姣,曹自刚,等. 网络流量分类研究进展与展望 [J]. 集成技术, 2012, 1(1): 31-41.
- [4] Dainotti A, Pescapé A, Claffy K C. Issues and Future Directions in Traffic Classification [J]. Network, IEEE, 2012, 26(1): 35-40.
- [5] JIN Y, DUFFIELD N, ERMAN J, et al. A Modular Machine Learning System for Flow-level Traffic Classification in Large Networks [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(1): 4.

- [6] ROUGHAN M, SEN S, SPATSCHECK O, et al. Class-of-service Mapping for QoS: a Statistical Signature-based Approach to IP Traffic Classification [C] //Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, ACM, 2004: 135 – 148.
- [7] LI B, SPRINGER J, BEBIS G, et al. A Survey of Network Flow Applications [J]. Journal of Network and Computer Applications, 2013, 36(2): 567 – 581.
- [8] KIM H, CLAFFY K C, FOMENKOV M, et al. Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices [C] //Proceedings of the 2008 ACM CoNEXT Conference, ACM, 2008: 11.
- [9] 牛琨, 张舒博, 陈俊亮. 采用属性聚类的高维子空间聚类算法 [J]. 北京邮电大学学报, 2007, 30(3): 1 – 5.
- [10] 朱林, 雷景生, 毕忠勤, 等. 一种基于数据流的软子空间聚类算法 [J]. 软件学报, 2013, 24(11): 2610 – 2627.
- [11] ESTER M, KRIEGEL H P, SANDER J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C] //Kdd. 1996, 96: 226 – 231.
- [12] 单世民, 闫妍, 张宪超. 基于 k 最相似聚类的子空间聚类算法 [J]. 计算机工程, 2009, 35(14): 4 – 6.
- [13] XIE G, ILIOFOTOU M, KERALAPURA R, et al. Subflow: Towards Practical Flow-level Traffic Classification [C] //INFOCOM, 2012 Proceedings IEEE, 2012: 2541 – 2545.
- [14] KRIEGEL H P, KROGER P, RENZ M, et al. A Generic Framework for Efficient Subspace Clustering of High-dimensional Data [C] //Data Mining, Fifth IEEE International Conference on IEEE, 2005: 1 – 8.
- [15] BUJLOW T, CARELA-ESPANOL V, BARLET-ROS P. Comparison of Deep Packet Inspection (DPI) Tools for Traffic Classification [R]. Universitat Politècnica de Catalunya, 2013.
- [16] MELO W, LOPES P, ANTONELLO R, et al. On the Performance of DPI Signature Matching with Dynamic Priority [C] //Computers and Communication (ISCC), 2014 IEEE Symposium on. IEEE, 2014: 1 – 6.
- [17] MOORE A W, PAPAGIANNAKI K. Toward the Accurate Identification of Network Applications [M] //Passive and Active Network Measurement. Springer Berlin Heidelberg, 2005: 41 – 54.
- [18] ILIOFOTOU M, KIM H, FALOUTSOS M, et al. Graph-based P2P Traffic Classification at the Internet Backbone [C] //INFOCOM Workshops 2009 IEEE, 2009: 1 – 6.
- [19] BARNAGHI P M, SAHZABI V A, BAKAR A A. A Comparative Study for Various Zethods of Classification [C] //Int. Conference on Information and Computer Networks (ICICN 2012), IPCSIT. 2012, 27.

(编辑: 温泽宇)