



网络应用流类别不平衡环境下的 SSL 加密应用流识别关键技术

陈雪娇¹, 王攀², 刘世栋³

(1. 南京信息职业技术学院 南京 210023 ;

2. 南京邮电大学 南京 210003; 3. 国网智能电网研究院 南京 210003)

摘要:通过深入研究网络类别不平衡的原因,选择 SMOTE(synthetic minority over-sampling technique)过抽样方法对数据集进行预处理,并充分利用特征匹配高准确性的优点识别和分拣出 SSL 加密流,进而利用基于互信息最大化的聚类方法和 SVM 分类方法进一步识别 SSL 加密应用,这种混合方法有效地结合了静态特征匹配和机器学习方法的优点,达到识别分类方法在准确性和识别速度的均衡。

关键词:流量识别;流量分析;行为特征;行为建模;行为模型

doi: 10.11959/j.issn.1000-0801.2015355

Key Technology of SSL Encrypted Application Identification Under Imbalance of Application Class

Chen Xuejiao¹, Wang Pan², Liu Shidong³

(1. Nanjing College of Information Technology, Nanjing 210023, China;

2. Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

3. State Grid Smart Grid Research Institute, Nanjing 210003, China)

Abstract: Through a in-depth study about the reason of network class imbalance, a method called SMOTE was chosen over the data set sampling preprocess, making full use of the advantages which is high accuracy of traffic model feature matching identification and sorting out the encrypted SSL flow, and then using the clustering method and the SVM based on mutual information classification method to further identify SSL encryption specific application, like HTTPS/POPS etc. The hybrid method effectively combines the advantages of static feature matching and machine learning methods, to achieve the balance of classification method on accuracy and speed.

Key words: traffic identification, traffic analysis, behavior characterization, behavior modeling, behavior pattern

收稿日期:2015-09-01;修回日期:2015-12-15

基金项目:2013 江苏省六大人才高峰计划项目,2013 国家发展和改革委员会信息安全专项资助项目,国家电网公司 2014 年科技项目“电力信息通信网络流量预测和管道智能化关键技术研究及应用”,2015 江苏省产学研前瞻性联合研究项目(No.BY2015011-02)

Foundation Items: 2013 Six Talent Peaks Project in Jiangsu Province, 2013 Information Security Special Funds of the National Development and Reform Commission, State Grid 2014 Science and Technology Project “Research and Application of Network Traffic Prediction and Smart Pipe Key Technologies for Electric Power Information Communication Network”, 2015 Industrialization, Education and Research Joint Prospective Project(No.BY2015011-02)

论文引用格式:陈雪娇,王攀,刘世栋. 网络应用流类别不平衡环境下的 SSL 加密应用流识别关键技术. 电信科学,2015355

Chen X J, Wang P, Liu S D. Key technology of SSL encrypted application identification under imbalance of application class. Telecommunications Science, 2015355



1 引言

网络流量识别对于服务质量保障、网络规划建设、网络安全监管等均有重要意义,是进行流量工程、实施 QoS 保障的基础;此外,网络负载建模、流量整形等问题的解决也依赖于有效的流量识别。近几年来,随着互联网的飞速发展,加密协议应用越来越广泛,尤其随着 SSL 的开放库 OpenSSL 的大量使用,加密业务和应用越来越多,一方面人们通过 SSL 加密来保护通信数据的隐私性;另一方面大量的 SSL 加密用于隐藏自己的非法活动。流量识别主要是基于网络流的静态特征和动态特征进行识别,然而,SSL 在通信过程中的数据报文是加密密文,因此传统的端口、净荷特征匹配等识别方法都失效了。基于动态特征的各种机器学习算法虽然对加密流识别有一定的改善,然而在准确性和在线识别方面都有很多不足。更为重要的是,这些基于机器学习的分类识别研究基础都是基于一种假设:各种网络应用流都是均匀分布在网络中,即网络数据流的应用类别是平衡的;然而,现实网络中加密数据流相比其他数据流稀少得多。网络应用流类别不平衡是指数据集中存在的类别样本数量不均衡,通过训练,这些分类算法可能会忽略少数类别的流样本导致欠拟合,或重视少数类的差别造成过拟合。因此,针对网络应用流类别不平衡环境下的加密应用的少数类识别成为一个既亟待解决又颇具挑战性的课题。

传统的流量识别办法是通过提取流量的特定参数,通过建立流量模型来识别业务和应用。该领域的研究主要集中在分类器的效率。各类研究机构分别在采用不同的专家系统或各种机器学习算法,比如 HMM、朴素贝叶斯模型、AdaBoost^[1-5]或最大熵方法来解决业务识别的问题。然而,对加密业务和应用的识别仍然束手无策。流量加密意味着净荷检测和分析是不可行的,端口识别也不行。

国外的 Haffner 等人比较了 AdaBoost、纯真贝叶斯等多种业务识别模型,结果显示,AdaBoost 的效果最好,对 SSH 能达到 86% 的识别率和零假阳性率,不过,该方法需要考察报文净荷的前 64 byte,由于 SSH 的报文加密在握手报文之后,所以对 SSH 流量的识别效果较好,然而,如果加密算法改变,那么识别效果将受到影响,此外,该方法对其他加密应用如 Skype 或者 VPN 隧道并不通用^[6,7];Baset 等人使用类似的方法,通过握手报文来识别加密的 Skype 流量,该方法对报文顺序十分敏感,报文乱序对准确率影响较大,此

外,识别器需要一定的训练过程,识别耗时较长。

综上所述,这几类方法的共性缺点:仅能对特定加密协议的特定版本有效,识别对象的范围较窄,对未知加密协议不具备识别能力。近年来,Wright 等人探寻通过仅仅使用分组大小、时间戳和连接的方向信息来识别通用的协议或应用^[8-10],采用 kNN 和 HMM 算法来比较性能,尽管这些方法可以分类具体的加密应用,但 SSH 的分类仅有 76% 准确率和 8% 的误报率。

Bernaille 等人首次采用 SSL 协议特征和聚类的方法来识别加密应用^[11];该方法的优点就是仅通过识别少量的前序分组的部分属性(大小、方向)即可识别加密应用类型,然而,一方面,该方法需要依赖于 SSL 协议的特征;另一方面该方法必须等待连接结束才能识别前 4 个分组,而且需使用端口号识别,如此一来,当端口号发生改变,不但性能且稳健性也下降,因为网络环境发生变化就需要重新训练。Williams 等人^[5]对比了包括贝叶斯网络、C4.5、朴素贝叶斯树等用于流量识别的分类算法,发现 C4.5 的性能最好,但他们仅给出每种机器学习算法的整体准确性,而不是每种应用的分类结果;他们发现对于网络应用数据流类别不平衡数据集(SSH 比例为 10%),可能导致大量错分;近期,Alshammari 等人基于 SSH 协议通过明文信令建立连接的事实,选择了 13 个特征字段和 14 个流属性,其中不包括净荷、IP 地址、端口信息,通过基于包括 C4.5/NB/SVM 在内的多种有监督机器学习算法实现对 SSH 协议流量的识别。但是该算法识别的对象并不普遍,局限于 SSH 协议流量,此外,识别过程需要遍历整个数据流,不能够在线识别^[12],结果表明基于 RIPPER 算法的分类器获得 99% 的检测准确率和 0.7% 的误报率;2009 年,Alshammari 等人^[13]又提出了一种机器学习的加密流量识别方法,用于识别 SSH 和 Skype 两种应用,测试了 5 种机器学习方法,包括 AdaBoost、SVM、NB、RIPPER 和 C4.5,仅使用流属性而不使用包括 IP 地址和净荷信息等,结果表明 C4.5 方法更适合 SSH 和 Skype 流量的识别,检测率达到 83.7%,误报率 1.5%。

赵博^[14]等人针对加密流量的在线普适识别问题,提出一种基于加权累积和检验的时延自适应加密流量盲识别算法,利用加密数据的随机性特点,对网络报文逐一实施累积和检验,根据报文长度将结果进行加权综合,无需解密操作,也无需匹配特定内容,实现对加密流量的普适识别。何高峰等人^[15]围绕广泛应用的 Tor 匿名通信系统,深入

分析运行机制,归纳总结流量特征,提出基于 TLS 指纹和基于报文长度分布的 Tor 匿名通信流量识别方法。王伟等人^[16]提出一种基于趋势感知协议指纹的 Skype 加密流量识别算法,通过定义趋势感知加权函数来反映流量特征的变化趋势,利用异常相似度来对 Skype 加密流量进行实时检测。他们也提出了一种基于 M-序列检验的网络数据随机性评估算法 NDRE 来识别加密流量^[17]。该方法根据负载序列长度,自适应训练得到最优参数集,利用最小风险贝叶斯准则,对加密流量进行识别。张宏莉等人^[18]提出在网络协议流不平衡环境下,如何选择合适的机器学习算法以适应网络协议流不平衡的环境。

2 现存问题

目前看来,针对 SSL 加密流的识别算法研究还存在许多问题亟待解决,具体说来,目前存在的主要问题有:

- 由于报文加密,无法直接应用基于静态特征的识别方法,如端口特征、净荷特征等;
- 识别效率比较低,误识别和漏识别的情况较多;
- 很难在识别准确性和识别速度上取得平衡;
- 没有考虑网络应用流类别不均衡问题,导致加密流这类少数类的识别准确率降低;
- 缺乏一个统一的检测评估标准和平台。

总的来说,网络应用流类别不均衡环境下 SSL 加密流的识别技术还不够成熟和完善,面临许多问题与困难,有很大的研究、发展空间,而现存的问题就是今后 SSL 加密流的识别技术的主要研究方向。

3 技术方案

3.1 总体思路

SSL 加密流识别问题近年来已逐渐受到越来越多学

者的重视。同时应该看到,随着流量识别的应用范围越来越广泛,各种网络应用也在不断地发生着变化,促使人们研究更加可靠、准确性更高的识别分类算法。许多经典的分类方法已不能够适应新的加密应用,尤其是在网络类别不平衡情况下,新的技术和思想不断被提出。因此在本课题的研究过程中将跟踪最新的研究动态,积极研究 SSL 加密流的流量特性、半监督机器学习在流量分类识别方面的优势以及 SMOTE (synthetic minority over-sampling technique) 过抽样方法有效解决网络类别不平衡问题的强大优势等思想,在综合比较和分析各种原有分类识别算法的性能优缺点的基础上提出新的算法和方案。在研究手段上将采用过抽样方法、半监督学习、各种机器学习方法 (AdaBoost/C4.5/NB/RIPPER/HMM 等)、互信息最大化聚类方法以及 Weka 等多种工具来分析和建模,在理论分析的基础上,主要通过对各种实际网络流量数据进行测试和验证,以保证所提出算法的有效性和可行性。

3.2 技术方案

本文的研究通过计算机软件仿真 (采用公开数据集、自制数据集) 与实际互联网网络流量数据相结合的方法来实现。首先采集纯净 SSL 加密流量 (可以来源于公开、自制数据集或真实流量), 通过 SMOTE 过抽样方法对数据集进行预处理以解决网络应用流类别不平衡的问题,并构建基于净荷特征匹配和机器学习方法的混合识别模型。在应用机器学习分类方法过程中,首先完成特征参数选择,然后寻找合适的流量模型以描述 SSL 加密流的特性,再选择合适的分类算法完成识别分类。在实现互信息最大化的聚类分类方法和过抽样方法为基础的识别分类方法时,针对多种机器学习方法进行分析与实验,找到其中的共性和特点。最后把各种算法用于实测数据与实际系统中,并评估算法的实时性与可靠性。总体方案如图 1 所示。

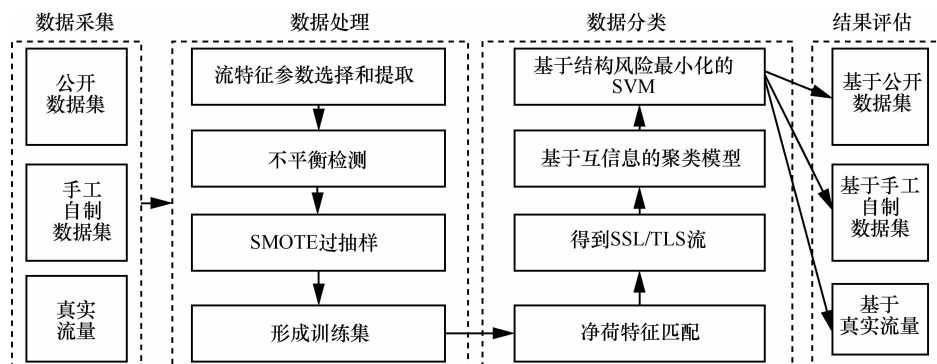


图 1 网络应用流类别不平衡环境下的 SSL 加密流识别技术方案



SSL 加密流量分类系统可分为数据采集、数据处理、数据分类和结果评估,其中数据处理和数据分类两部分是最重要的。数据处理主要是需要解决类别不平衡问题,本课题计划采用 SMOTE 过抽样方法来对训练集的小样本进行数据预处理。流量分类结合 K -means 聚类算法和支持向量机建立分类模型。训练集首先通过聚类学习缩小训练规模,针对聚类形成的簇内应用是不纯净的,簇内进行支持向量机的分类,训练后的性能情况反馈指导分类器的参数选取。SMOTE 过抽样方法可以按照一定的策略规则,采用插值方法合成新样本,有良好的泛化作用,避免随机复制原有样本的盲目和局限。

其中,基于净荷特征匹配和机器学习的混合 SSL 应用流识别模型和框架如图 2 所示。

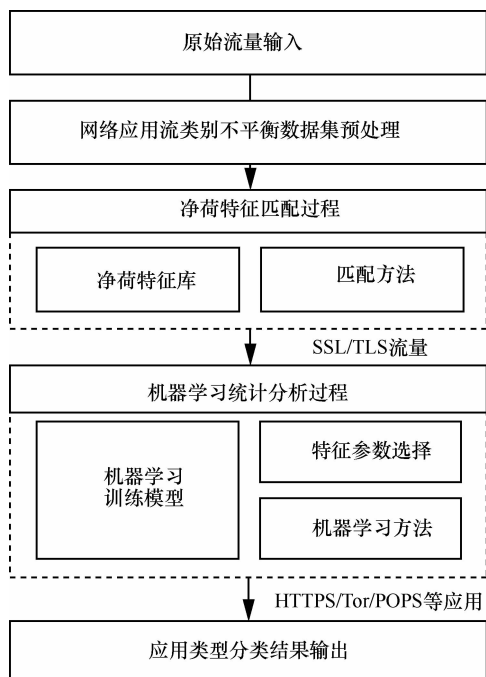


图 2 网络数据流类别不平衡环境下的 SSL 加密流量分类模型框架

图 2 所示的分类识别模型主要包括如下几个步骤。

- 网络数据流类别不平衡的数据预处理。该过程将在第 3.4 节详细介绍。
- 基于净荷特征匹配来检测和识别出 SSL 加密流。

根据 SSL 交互协议,SSL 握手交互过程如图 3 所示:客户端和服务端协商 SSL 版本,选择加密算法;服务端和客户端相互鉴权,协商加密密钥;终止握手,开始在加密信道交换应用数据。

SSL 流识别可以充分利用协议连接建立阶段的信令

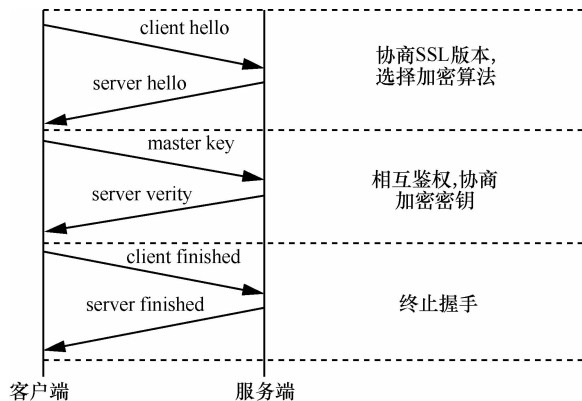


图 3 SSL 交互过程

内容特征来实现 SSL 协议流量的识别,例如,一个连接是否使用了 SSLV2 加密,可以通过连接的第 2 个报文(服务器向客户端发送的第一个报文)来判断,前面 2 bit 内容总为 1 和 0,接下来的 14 bit 则指示了 SSL 的长度,第 3 个字节指示消息类型(1 为客户端的 hello 消息,4 为服务器端的 hello 消息),该方法利用 SSL 协议连接建立阶段的明文特征完成识别,并不适用于其他的加密协议。

3.3 基于半监督学习方法来识别 SSL 应用流

采用机器学习方法来识别 SSL 加密应用流是基于如下几个考虑:

- SSL 加密并不明显地修改分组数目、大小以及分组到达时间;
- TCP 连接可以被分类,基于流级别信息,比如流持续时间、分组数目、平均到达时间等。

通过机器学习方法来识别 SSL 流量,需要两个阶段来构建分类器:训练和在线分类。训练阶段采用聚类算法应用于 TCP 连接(训练数据集)上。把每个连接用 P 维空间(包括第 P 个分组的大小、方向等)来表示,结果簇包含一个或多个应用,在线分类器使用启发式来把一个 TCP 连接分配给簇中的一个,进一步通过启发式给这个 TCP 连接打标签。

在本方法中,流被定义为两个主机中的一系列分组,通过五元组{源地址、目的地址、源端口、目的端口、传输层协议},流终结通过预设的超时时限或流终结标识来确定。对每个流而言,都需要提取相关参数,比如流持续时间、流大小、平均分组到达时间间隔和平均分组大小。

设定 $X=\{X_1, \dots, X_n\}$ 表示一组流,一个流实例 X_i 由一个属性向量来表示, $X_i=\{X_{ij} | 1 \leq j \leq m\}$, 此处的 m 是属性的数

目, X_{ij} 是第 i 个流的第 j 个属性值。同时, 设定 $Y=\{Y_1, \cdots, Y_q\}$ 代表一组流量类别, 其中 q 代表分类类别个数, Y_q 代表具体的应用类别, 诸如 HTTP, 因此, 本文的应用类型识别就是研究一个 m 维的变量 X 到 Y 的映射。

图 4 展示了一个机器学习的 SSL 加密流量分类的训练过程。第一阶段: 训练数据准备过程主要涉及流属性参数的提取和格式化, 为后期机器学习做好训练数据准备; 第二阶段: 网络类别不平衡处理, 主要解决小样本流量识别在常规机器学习流量分类中的不准确性; 第三阶段, 选择合适的属性参数, 建立标识流的向量空间, 应用机器学习分类算法; 第四阶段, 建立分类识别模型, 并计算识别准确率等性能评价指标。

在这其中特征参数的选择非常关键, 参数过多, 影响识别速度; 参数过少, 影响识别准确性, 因此需要统筹考虑。本课题在研究了国内外同行的特征参数选择方法后, 选择了合适的特征参数以兼顾识别准确性和识别速度, 参数选择见表 1。

3.4 网络数据流类别不平衡的数据预处理方法选择

在分类器进行训练学习之前, 必须采用过抽样技术提高样本的可分性, 拓展分类的决策边界。SMOTE 过抽样方法是 Chawla 等人^[19]在手写识别的应用启发下, 提出通过对特征空间进行处理来合成新的训练数据。这是一种随机过抽样方法, 其主要思想是应用 K -最近邻方法, 在少数类样本之间利用线性插值生成新的样本, 从而增加少数类样本的数量, 使数据集的类别数量相对平衡。

SMOTE 过抽样方法按照一定规则, 利用插值方法合成新样本, 有良好的泛化作用, 避免了随机复制原有样本的盲目和局限。没有引进少数类的冗余信息, 扩展决策区域, 避免过拟合问题, 改善了数据集的可分性。但是, 该算法的简单性也带来隐患, 在少数类与多数类样本空间重叠

表 1 流量特征参数选择

序号	特征参数描述	简称
1	上行分组数目	fpkts
2	下行分组数目	bpkts
3	上行平均分组长度	meanfpktl
4	上行最大分组长度	maxfpktl
5	下行平均分组长度	meanbpktl
6	下行最大分组长度	maxbpktl
7	平均活跃流数目	meanaf
8	最大活跃流数目	maxaf
9	平均静默流数目	meanif
10	最大静默流数目	maxif

的情况下, 通过插值得到的新样本可能并不属于少数类, 对新样本的类别判定缺乏说服力。另外, SMOTE 过抽样方法没有对少数类样本进行区分, 忽略了噪声及孤立点的影响。真正影响分类的是边界样本, 边界样本分布的稀缺影响少数类的分类性能。但针对 SSL 加密流这种少数类, SMOTE 过抽样方法算是一个不错的选择。

3.5 识别模型和算法的评估

为了测量分类方法的性能, 本文主要采用两个维度: 检测率 DR 和误判率 FPR。其中 DR 反映了被正确分类的加密流数目; 而 FPR 反映了被错误分类的加密流, 很明显, 高检测率和低误判率是本文需要追求的目标。具体计算如下:

$$DR=1-\frac{FN}{\text{加密流总数}}$$

$$FPR=\frac{FQ}{\text{非加密流总数}}$$

其中, FP 为误报数目; FN 为漏报数目。

计划采用的数据集来源有以下几种。

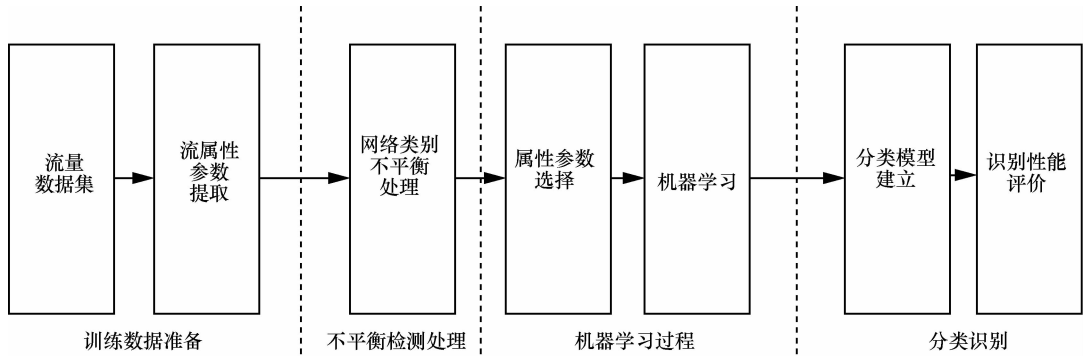


图 4 SSL 加密应用识别的机器学习分类过程



表 2 网络类别不平衡条件下基于半监督的 SSL 加密流识别测试结果

算法	训练集总数:10 000					训练集总数:20 000				
	分 类 准 确率 DR	SSL 分类误 判率 FPR	非 SSL 分类 误判率 FPR	SSL 查 全率	非 SSL 查 全率	分类准确 率 DR	SSL 分类误 判率 FPR	非 SSL 分类 误判率 FPR	SSL 查 全率	非 SSL 查全率
AdaBoost C4.5	95.56%	0.03%	0	1	0.97	89.53%	0.1	0	1	0.9
AdaBoost	94.21%	0.04%	0	1	0.96	93.66%	0.07	0	1	0.93
C4.5	94.02%	0.04%	0	1	0.96	91.33%	0.09	0	1	0.91
Naive Bayes	9.73%	0.92%	0	1	0.08	10.57%	0.9	0	1	0.1
RIPPER	92.21%	0.07%	0	1	0.93	99.51%	0.01	0	1	1

- 公开数据集:意大利都灵理工大学的公开数据集 UNIBS;该流量中包含 HTTPS/SSH/Skype 等加密流量,同时还有常规的 HTTP/SMTP/FTP/P2P 等背景流量。
- 实际流量:南京邮电大学校园网出口流量。
- 手工自制数据集:通过流量发生工具仿真模拟出 SSL 加密流以及背景流量。

4 实验结果

本文从南京邮电大学的校园网出口采集了部分用户上传的流量数据,这其中所有的 SSL 流量均在用户终端侧进行了统计,可以作为识别算法和模型的标准,表 2 进行了测试对比。

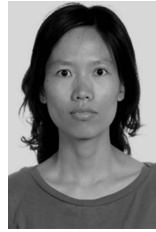
通过实际测试,结果证明在网络类别不平衡的条件下,本文应用了 AdaBoost C4.5、AdaBoost C4.5、NB、RIPPER 等几种机器学习的方法,从 FPR、DR 两个性能指标进行了测试对比,结果表明该方法适用于在网络类别不平衡的条件下对 SSL 流量进行准确识别。

参考文献

- Bai Y B, Kobayashi H. Intrusion detection systems: technology and development. Proceedings of the 17th International Conference on Advanced Information Networking and Applications, Xi'an, China, 2003: 27~29
- Wright C, Monroe F, Masson G M. HMM profiles for network traffic classification. Proceedings of the ACM DMSEC, Washington DC, USA, 2004: 9~15
- Haffner P, Sen S, Spatscheck O, *et al.* ACAS: automated construction of application signatures. Proceedings of the ACM SIGCOMM, Philadelphia, USA, 2005: 197~202
- Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. Proceedings of the ACM SIGMETRICS, Banff, Alberta, Canada, 2005: 50~60
- Moore A, Papagiannaki K. Toward the accurate identification of network applications. Proceedings of the Passive & Active Measurement Workshop, Boston, USA, 2005
- Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow comparison. ACM SIGCOMM Computer Communication Review, 2006, 36(5): 5~16
- Zhang Y, Paxson V. Detecting back doors. Proceedings of the 9th USENIX Security Symposium, Denver, USA, 2000: 157~170
- Dreger H, Feldmann A, Mai M, *et al.* Dynamic application layer protocol analysis for network intrusion detection. Proceedings of the 15th USENIX Security Symposium, Vancouver, Canada, 2006: 257~272
- Early J, Brodley C, Rosenberg C. Behavioral authentication of server flows. Proceedings of the 19th Annual Computer Security Applications Conference, Las Vegas, USA, 2003: 46~55
- Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark. Proceedings of Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, USA, 2005: 229~240
- Wright C V, Monroe F, Masson G M. On inferring application protocol behaviors in encrypted network traffic. Journal of Machine Learning Research, 2006(7): 2745~2769
- Bernaille L, Teixeira R. Early recognition of encrypted applications. Proceedings of Passive and Active Measurement Conference (PAM), Louvain-la-neuve, Belgium, 2007
- Alshammari R, Nur Z H A. A flow based approach for SSH traffic Detection. Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Montreal, Canada, 2007: 296~301
- 赵博,郭虹,刘勤让等. 基于加权累积和检验的加密流量盲识别算法. 软件学报, 2013, 24(6): 1334~1345
- Zhao B, Guo H, Liu Q R, *et al.* Protocol independent identification of encrypted traffic based on weighted cumulative sum test. Journal of Software, 2013, 24(6): 1334~1345

- 15 何高峰,杨明,罗军舟等. Tor 匿名通信流量在线识别方法, 软件学报, 2013, 24(3): 540~556
He G F, Yang M, Luo J Z, *et al.* Online identification of Tor anonymous communication traffic. Journal of Software, 2013, 34(3): 540~556
- 16 王炜,程东年,马海龙. 基于趋势感知协议指纹的 Skype 加密流量识别算法. 计算机应用研究, 2014(8): 64~71
Wang W, Cheng D N, Ma H L. Skype encrypted traffic identification based on trend-aware protocol fingerprints. Application Research of Computers, 2014(8): 64~71
- 17 王炜,程东年. 基于 M-序列检验的加密流量识别. 计算机工程与设计, 2014, 35(11): 3712~3716
Wang W, Cheng D N. M-serial test based encrypted traffic identification. Computer Engineering and Design, 2014, 35(11): 3712~3716
- 18 张宏莉,鲁刚. 分类不平衡协议流的机器学习算法评估与比较. 软件学报, 2012, 23(6): 1500~1516
Zhang H L, Lu G. Machine learning algorithms classifying the imbalanced protocol flows: evaluation and comparison. Journal of Software, 2012, 23(6): 1500~1516
- 19 Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321~357

[作者简介]



陈雪娇,女,南京信息职业技术学院讲师,主要研究方向为信息网络、移动互联网、信息安全。



王攀,男,博士,南京邮电大学副研究员,主要研究方向为信息网络、业务感知、DPI、大数据分析。



刘世栋,男,博士,国网智能电网研究院信息通信研究所主任工程师,主要研究方向为电力通信网。