基于分治排序策略的流量二次特征选择

申 健[1],夏靖波[2],张晓燕[1],赵广辉[3],付 凯[1]
(1. 空军工程大学信息与导航学院,陕西西安 710077;2. 厦门大学嘉庚学院,福建厦门 363105;3. 辽宁科技大学,辽宁鞍山 114000)

摘 要: 网络业务流量的多样化高速化发展给流量识别技术带来了极大挑战,特征选择作为对数据降维处理的有效方法,具有重要的研究意义. 本文描述了流量二次特征选择模型,并以此为基础提出了流量二次特征选择算法. 算法将流量数据分为若干数据子集进行分治处理,对各数据子集提取出的特征进行汇总,以提出的影响度这一指标作为特征评估排序的依据,进行二次特征提取. 实验表明,提出的算法在模型构建上性能更加优越,并且可以选取更少的特征实现对流量更准确的识别.

关键词: 二次特征提取; 分治; 排序; 影响度; 流量识别
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2017)01-0128-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.01.018

Secondary Feature Extraction of Network Traffic Based on Divide-Conquer and Ranking Strategy

SHEN Jian[1], XIA Jing-bo[2], ZHANG Xiao-yan[1], ZHAO Guang-hui[3], FU Kai[1]
(1. *School of Information and Navigation, Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*;
2. *Tan Kah Kee College, Xiamen University, Xiamen, Fujian* 363105, *China*;
3. *University of Science and Technology Liaoning, Anshan, Liaoning* 114000, *China*)

Abstract: The diversified and high-speed development of network traffic presents a great challenge for traffic identification. As an effective method for data dimensionality reduction, the research of feature extraction is of great significance. A secondary traffic feature extraction model is described as the foundation of the secondary feature extraction algorithm of network traffic. The algorithm divides traffic data into several subsets and gathers the features extracted from different subsets. The index of influence is proposed as the reference of feature ranking and extraction. The experiment results show that the secondary traffic feature extraction model has better performance, and the algorithm can identify traffic more accurately with fewer features.

Key words: secondary feature extraction; divide-conquer; ranking; influence; traffic identification

1 引言

网络流量识别是认识、管理、优化各种网络资源的基础和重要依据,对网络的管理、安全分析以及趋势预测都起着非常重要的作用[1]. 网络数据量的爆炸式增长,对网络流量识别的实时性和系统资源利用的合理性提出了更高的要求和挑战. 当前,通过采集流量的外部特征属性并应用机器学习算法进行分类是最常用的流量识别方法[2~4]. 特征选择作为流量识别的预处理过程[5],是对海量数据进行降维的有效预处理方法. 特征选择不仅能够降低流量识别模型的复杂度[6],而且能够节约系统资源,提高对流量的识别准确率. 这是由于冗余特征的存在会降低流量识别算法的效率,而不相关特征的存在会有损算法的性能[7]. 因此,特征选择能够降低识别算法计算代价的同时,也能生成更易理解的结果,构建更紧凑泛化能力更强的模型.

传统特征选择方法主要有过滤特征选择算法和封装特征选择算法[8,9],在此基础上近些年又有了更加广泛和深入的研究. 文献[10]结合流量矩阵和网络结构熵定义了多个参数描述节点间连接行为和数据传输特征,并利用多个周期和时间尺度下的熵指数分析不同流量特征. 文献[11]基于文化基因框架,结合了封装和

收稿日期:2015-12-21;修回日期:2016-04-11;责任编辑:孙瑶
基金项目:陕西省科技计划自然基金重点项目(No. 2012JZ8005);航空科学基金(No. 20141996018)

过滤的特征选择算法,保证了全局最优解的同时加快了寻找最优特征子集的收敛速度. 文献[12,13]中利用信息熵的概念对特征的贡献度进行评价并选择. 文献[14,15]提出了基于预测风险的嵌入式特征选择方法,以特征属性值被平均值代替前后分类精度的差值大小作为特征选择的依据. 文献[16]提出了分治的特征选择策略,能够使现有的并行计算和分布式计算等理念在特征选择过程中得以实现,然而,文中特征选择采用投票机制,忽略了特征之间的相关性,使得随着特征选择数量的增加,流量识别准确率没有明显提高,甚至下降.

本文系统描述了流量二次特征选择模型,并在此基础上引入了影响度这一评价标准,提出了基于影响度因子的分治排序(Divide-conquer and Ranking,DR)二次特征选择算法,实现了特征子集的优化选取.

2　二次特征选择模型

特征选择的目标是在全部特征集合中,选择部分特征组成特征子集[17],该特征子集需涵盖原始特征集合的全部或者大部分信息,使得通过对特征子集进行较少的计算分析即可得到与使用全部特征集合相近的分析结果. 假设流量数据集 T 的特征集为 F,$P(F)$ 为通过特征集 F 对数据的识别率,特征选择即为通过相应的算法选取特征子集 F',且 $P(F')\approx P(F)$,或者 $P(F')>P(F)$.

分治是求解大规模优化问题的有效方法,本文将分治策略应用于流量的二次特征选择模型构建中,以应对流量数据量大、业务类型多的特点. 分治策略首先将原始数据集 T 划分为若干数据子集 $\{T_1,T_2,\cdots,T_n\}$,然后通过相应的搜索方法和特征选择方法,对不同数据子集中的特征分别进行提取. 不同数据子集中多种类型流量随机分布,各数据子集选择的特征子集会有一定差异. 对多个特征子集进行汇总形成特征汇总集 F^*,由于 F^* 是由若干个数据子集提取汇总而得,特征数量仍旧较多,需对其进行二次选择,从而获取更加精准的特征子集对数据集进行描述. 二次特征选择模型如图 1 所示.

本文选择 BF(Best First)[18] 搜索算法与 CFS(Correlation based Feature Selection)[19] 属性评估算法相结合的方法进行特征初次选择. BF 搜索算法采用带回溯增强的贪婪爬山法,实现了特征子集的搜索. CFS 属性评估算法以类别属性高度相关、同时相互之间相关度低作为特征子集的选择准则,通过考虑单个特征的标识能力,以及特征间的冗余度评估属性子集的价值.

以上描述的是流量二次特征选择模型,本文在此基础上通过定义特征的影响度,提出了流量分治排序二次特征选择算法.

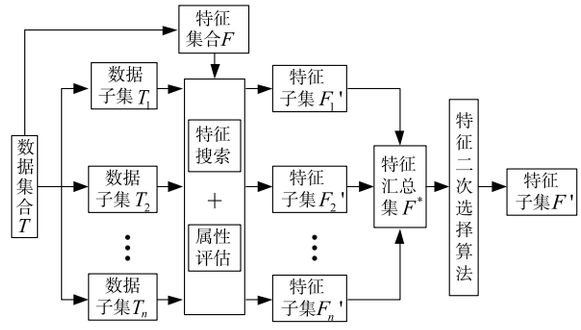


图 1　二次特征选择模型

3　DR 二次特征选择算法

3.1　特征排序

通过分析可知,特征对识别准确度的影响不仅仅和本身信息量有关,而且和特征间的相关度联系紧密. 本文提出了影响度这一评价指标,作为衡量特征对流量识别贡献度的标准,特征 f 的影响度定义为:

$$I(f)=P_N-P_{N-1}(f) \tag{1}$$

其中 N 为特征汇总集的特征数,P_N 为含有 N 个元素的特征汇总集对流量的识别准确率,$P_{N-1}(f)$ 为去除特征 f 后剩余 $N-1$ 个特征对流量的识别准确率. 通过对影响度指标的获取巧妙地回避了需要分别对复杂的特征信息量和冗余度进行计算的问题. 通过分析可知,影响度这一评价指标综合了特征本身的信息量和冗余度,即影响度越高说明该特征包含不重叠于其他特征的信息量越高,对数据的识别能力越强.

在特征排序的过程中,首个被选择的特征 f_1 需满足以下条件

$$\begin{aligned} f_1 &= \arg\max_{f_n}[I(f_n)] \\ &= \arg\max_{f_n}[P_N-P_{N-1}(f_n)] \\ &= \arg\min_{f_n}[P_{N-1}(f_n)];n\in[1,N] \end{aligned} \tag{2}$$

此后选取的特征依旧以特征影响度为评价准则,但需排除已选特征对备选特征影响度的干扰,选择的第 a 个特征需满足以下条件

$$\begin{aligned} f_a &= \arg\max_{f_n}[I(f_n)] \\ &= \arg\max_{f_n}[P_{N-(a-1)}-P_{N-a}(f_n)] \\ &= \arg\min_{f_n}[P_{N-a}(f_n)];n\in[a,N] \end{aligned} \tag{3}$$

$P_{N-(a-1)}$ 为排除前 $a-1$ 个特征后的汇总集的识别准确率,$P_{N-a}(f_n)$ 为在此基础上排除特征 f_n 的识别准确率.

由式(2)~(3)可知,每次选择的特征均为当前特征集中影响度最高的特征.

3.2　特征个数确定

由于特征之间关联复杂,会对彼此的识别能力造

成干扰,选择特征数量的增加有可能造成识别准确率的下降. 特征选择的后期,选取的特征包含冗余信息较高而信息量较少,因此随着特征数量的增加,识别准确率大体上会呈现出先上升后下降的趋势. 因此,特征子集元素个数的选取对识别准确率的影响很大. 当特征子集中含有 n 个元素时,对流量识别的准确率记为 P_n,本文默认当随后两次选择的特征均会造成识别准确率依次下降(即 $P_n > P_{n+1} > P_{n+2}$)时,停止特征选择,并确定特征子集元素个数为 n.

3.3 算法流程

步骤一 将数据集 T 均分为若干数据子集.

步骤二 用 BF 搜索与 CFS 属性评估相结合的方法分别对数据子集进行特征选择.

步骤三 将选择出的特征进行合并,生成特征汇总集 F^*.

步骤四 采用 Naive Bayes 分类算法,并利用十折交叉验证方法将数据集分为 10 份,将其中 1 份作为测试集,其余 9 份作为训练集,依次进行 10 次训练和评估. 计算特征汇总集 F^* 中每个特征的影响度 $I(f)$.

步骤五 选择影响度最大的特征为特征子集元素,更新特征子集.

步骤六 在特征汇总集 F^* 中删除已选特征,并更新特征汇总集.

步骤七 计算特征子集识别准确率,如果 $P_n < P_{n-1} < P_{n-2}$,则转至步骤八;否则,返回步骤四.

步骤八 将选择的特征按顺序汇总为特征子集.

4 实验与分析

本文使用数据挖掘软件 Weka-3.7.3 为主要实验工具;实验 PC 机 CPU 为 Inter Pentium(R) Dual-Core 2.60GHz,内存为 DDR-667 2GB;运行 Windows XP 操作系统. 采用 Moore 流量数据集进行实验[20],数据集中的业务流量分为{WWW, MAIL, FTP-CONTROL, FTP-PASV, ATTACK, P2P, DATABASE, FTP-DATA, MULTIMEDIA, SERVICES, INTERACTIVE, GAMES}共 12 个种类,数据集共包含 203355 条数据流以及 248 个特征[21].

实验将数据集分割成 8 个数据子集,采用 BF + CFS 算法对数据子集分别进行全局特征选择,并对生成的特征汇总集再次进行全局特征选择;采用 DV(Divide-conquer and Voting)[16]算法进行分治投票特征选择;采用 DR 算法进行分治排序二次特征选择. 通过以上三种算法生成三个不同特征子集,采用 Naive Bayes 分类算法对业务流量进行识别,比较不同算法选择出的特征子集对流量识别结果的影响.

4.1 数据预处理

将 Moore 流量数据集随机分割为 8 个数据子集,并对分割后的 8 个数据子集分别采用 BF 搜索和 CFS 属性评估相结合的方法进行特征选择,特征选择结果如表 1 所示,表内数字为特征的数字索引.

表 1 分治策略特征选择结果

F_1	4	50	72	91	108	155	202		
F_2	4	71	72	81	119	149	152		
F_3	4	72	78	108					
F_4	4	78	109	137	148				
F_5	4	51	109	137	180				
F_6	4	8	26	62	74	95	139	193	259
F_7	4	29	78	83	137				
F_8	4	66	78	81	137	148			

将提取出的特征进行汇总,特征汇总集为{4,8,26,29,50,51,62,66,71,72,74,78,81,83,91,95,108,109,119,137,139,148,149,152,155,180,193,202,259},通过 DR 算法进行二次特征选择,生成特征子集为{95,4,71,108,119,8},比较不同算法生成的特征子集,结果如表 2 所示.

表 2 不同算法提取特征子集比较

算法	特征排序						
	1	2	3	4	5	6	7
BF + CFS	4	51	66	78	119	149	193
DV	4	78	137	72	81	108	109
DR	95	4	71	108	119	8	

采用以上不同算法提取出的特征子集对业务流量数据集进行识别,评估验证采用十折交叉方法对数据进行交叉验证,每一次实验重复 10 次,实验结果取十次重复实验的平均值.

4.2 算法性能比较

4.2.1 算法模型比较分析

实验通过三种不同算法对流量特征进行选择,结果如图 2 所示. DR 算法相比于另外两种算法在特征选择过程中少选取了 1 个特征. DR 算法构建的模型较小,即占用系统内存资源较少. ROC(Receiver Operating Characteristic)曲线是显示分类器真阳性率和假阳性率之间折中的一种图形化方法,Area of ROC 指标是曲线下方的面积,大小越接近 1 说明模型的平均性能越好,由图 2(c)可见 DR 算法的 ROC 面积更接近 1,说明 DR 算法模型平均性能更好. 由图 2(d)可知,DR 算法的案例覆盖度更大,即算法使用分类规则对样本的覆盖程度更高,算法规则更有效. 综合以上性能,DR 算法可以选择更少的特征,并且构建的模型在各指标中均表现出更好的性能.

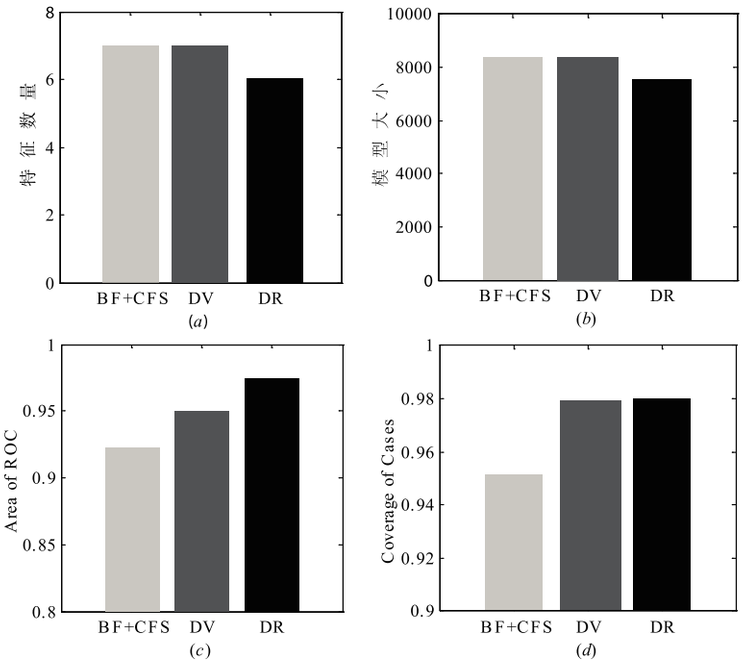


图 2　不同算法模型性能指标比较

4.2.2　算法对流量识别性能影响

算法对流量识别性能的影响(如图 3 所示)通过以下四个参数指标进行度量. 整体准确率 Overall Accuracy 表示分类模型正确预测样本数在预测总数中的比例

$$\text{OverallAccuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

TP(True Positive)和 TN(True Negative)都是正确分类结果,FP(False Positive)和 FN(False Negative)是错误分类结果,具体含义如表 3 所示.

查准率 Precision 是衡量检索系统拒收非相关信息的能力,衡量了对流量正确识别的准确程度. 查全率 Recall 是衡量检索系统检出相关信息的能力,衡量了对流量正确识别的完整程度. F-Measure 是查准率和查全率的调和平均数.

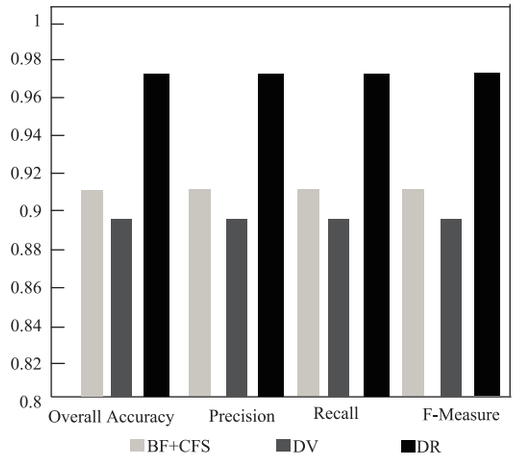


图 3　算法对流量识别性能影响

表 3　分类结果含义

真实类别	预测类别	
	T	F
T	TP	FN
F	FP	TN

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{F-measure} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}.$$

通过对图 3 的结果分析可知,DR 算法只有在 Recall 这一度量指标上略低于 BF + CFS 算法,在其余度量指标上性能均好于另外两种算法. 结合不同算法特征选择的数量可以发现,DR 算法在选择更少的特征的情况下,能够通过占用更少的系统资源更好地对流量进行识别,充分体现了该算法的优越性.

4.2.3　特征数量对流量识别性能影响

由于 BF + CFS 算法并不是逐个选择特征,而是通过评估属性子集的价值直接选取特征子集,特征数目固定,所以,此部分只对 DV 算法和 DR 算法进行比较,如图 4 所示.

由图 4 可知,DR 算法在整体准确率、查准率、F-Measure 三个指标中,随着选择特征数量的增加,度量指标性能越来越好,且在选取特征数量为任意值时,DR

算法性能均优于 DV 算法. 在查全率这一指标中,随着特征数量的增加,DR 算法查全率始终保持在 99% 以上,且随着特征数量的增加保持平稳,而 DV 算法在特征选取数量增加时,查全率指标波动较大不稳定,且总体呈下降趋势.

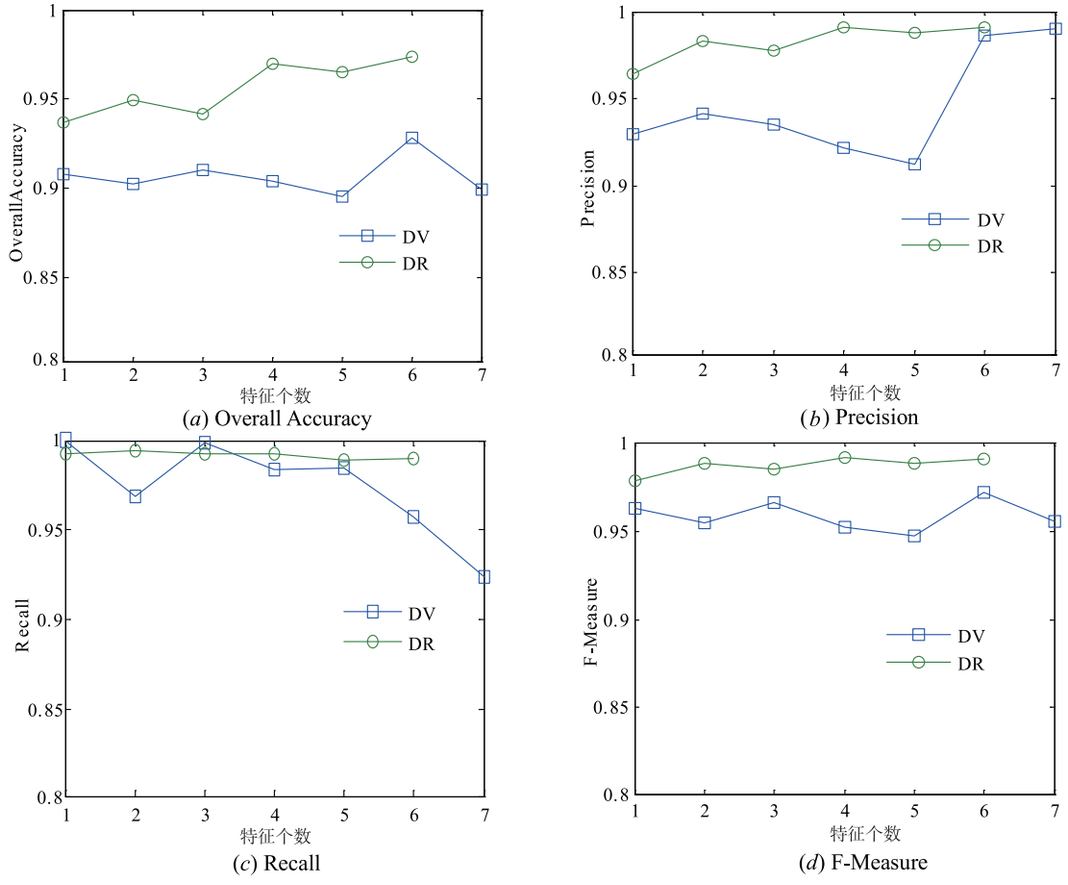


图4 特征选择数量对识别性能影响

从评价算法性能最重要的指标 Overall Accuracy 可以看出,DR 算法特征选择越多流量识别总体准确率越高,而选择的特征越多,对流量识别系统的内存、计算资源需求越大. 因此当系统资源不足时,可以按照 DR 算法的特征选择过程顺序提取出系统能够承载的特征数目,这样不仅节约了系统资源,又能够保证算法较高的识别性能.

5 结束语

在网络流量呈现数据量大、特征变量多的发展现状下,特征选择在对数据降维的预处理过程中被广泛应用. 本文详细描述了流量二次特征选择模型,提出了分治排序的流量特征选择算法;并将二者结合,分阶段合理地对流量特征进行选择;利用影响度这一指标对特征进行评价排序,在优化了特征选择模型的同时,提高了特征选择准确度,最终达到了综合提升流量识别性能的目的. 当系统资源不足时,可以根据系统的具体情况,通过 DR 算法选取少量特征,仍旧能够保证对流量较高的识别准确率. 在选取任意同等特征数量的前提下,DR 算法性能优于其他算法.

本文采用了分治的数据预处理方式,极大地缩减了数据的预处理时间;而作为代价,也成倍的增大了同一时间数据处理量,提高了计算负载,对数据处理设备的分布式并行处理能力提出了更高的要求. 另外,对影响度因子的计算需要对识别准确度多次比较,因此算法计算复杂度相对较高. 在二次特征选择模型的基础上,简化算法的计算复杂度将是下一步的工作重点. 随着网络业务流量的复杂化发展以及云计算技术、数据中心网络的普及建设,本文提出的基于分治排序策略的流量二次特征选择算法将使得云计算技术以及数据中心网络在现有大规模流量识别中发挥重要作用,极大提升网络管理效率,具有较高的理论研究与实际应用价值.

参考文献

[1] 张宾,杨家海,吴建平. Internet 流量模型分析与评述[J]. 软件学报,2011,22(1):115 - 131.
ZHANG Bin, YANG Jia-hai, WU Jian-ping. Survey and analysis on the internet traffic model [J]. Journal of Software, 2011, 22(1): 115 - 131. (in Chinese)

[2] GRIMAUDO L, MELLIA M, BARALIS E, et al. Self-learning classifier for Internet traffic [A]. Proceedings of the IEEE INFOCOM [C]. Turin, Italy: IEEE Press, 2013. 3381 - 3386.

[3] DING L, YU F, PENG S, et al. A classification algorithm for network traffic based on improved support vector machine [J]. Journal of Computers, 2013, 8(4): 1090 - 1096.

[4] 杨建华,谢高岗,张广兴,等. 一种高效的业务流分类算法[J]. 电子学报,2006,34(3):549 - 552.
YANG Jian-hua, XIE Gao-gang, ZHANG Guang-xing, et al. An efficient algorithm for flow classification [J]. Acta Electronica Sinica, 2006, 34(3): 549 - 552. (in Chinese)

[5] 张广兴,张大方,谢高岗,等. Internet 城域出口链路流量测量与特征分析[J]. 电子学报,2007,35(11):2092 - 2097.
ZHANG Guang-xing, ZHANG Da-fang, XIE Gao-gang, et al. Internet traffic measurement and characteristic analysis on output link of metro area network [J]. Acta Electronica Sinica, 2007, 35(11): 2092 - 2097. (in Chinese)

[6] 马祥杰,兰巨龙,毛军鹏,等. 输入排队 Crossbar 架构下的流量模型[J]. 电子学报,2009,37(1):170 - 174.
MA Xiang-jie, LAN Ju-long, MAO Jun-peng, et al. Traffic model for input-queued crossbar fabric[J]. Acta Electronica Sinica, 2009, 37(1): 170 - 174. (in Chinese)

[7] ZHANG H, LU G, QASSRAWI M T, et al. Feature selection for optimizing traffic classification [J]. Computer Communications, 2012, 35(12): 1457 - 1471.

[8] AMIRI F, REZAEI Y M, LUCAS C, et al. Mutual information-based feature selection for intrusion detection systems [J]. Journal of Network and Computer Applications, 2011, 34(4): 1184 - 1199.

[9] YANG J, MA J, CHENG G, et al. An empirical investigation of filter attribute selection techniques for high-speed network traffic flow classification [J]. Wireless Personal Communications, 2012, 66(3): 541 - 558.

[10] 叶春明,王珍,陈思,等. 基于节点行为特征分析的网络流量分类方法[J]. 电子与信息学报,2014,36(9):2158 - 2165.
YE Chun-ming, WANG Zhen, CHEN Si, et al. Internet traffic classification based on hosts behavior analysis [J]. Journal of Electronics & Information Technology, 2014, 36(9): 2158 - 2165. (in Chinese)

[11] 苗长胜,原常青,王兴伟,等. 基于互信息和文化基因算法的网络流量特征选择[J]. 东北大学学报,2014,35(11):1530 - 1534.
MIAO Chang-sheng, YUAN Chang-qing, WANG Xing-wei, et al. A hybrid feature selection algorithm based on mutual information and memetic framework to optimize traffic classification [J]. Journal of Northeastern University(Natural Science), 2014, 35(11): 1530 - 1534. (in Chinese)

[12] 黄志艳. 一种基于信息增益的特征选择方法[J]. 山东农业大学学报(自然科学版),2013,44(2):252 - 256.
HUANG Zhi-yan. Based on the information gain text feature selection method [J]. Journal of Shandong Agricultural University(Natural Science Edition), 2013, 44(2): 252 - 256. (in Chinese)

[13] 张振海,李士宁,李志刚,等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展,2013,50(6):1177 - 1184.
ZHANG Zhen-hai, LI Shining, LI Zhi-gang, et al. Multi-label feature selection algorithm based on information entropy [J]. Journal of Computer Research and Development, 2013, 50(6): 1177 - 1184. (in Chinese)

[14] LI G Z, YOU M, GE L, et al. Feature selection for semi-supervised multi-label learning with application to gene function analysis [A]. Proceedings of the 2010 ACM International Confernece on Bioinformatics and Computational Biology [C]. New York: Association for Computing Machinery, 2010. 354 - 357.

[15] YOU M Y, LIU J M, LI G Z, et al. Embedded feature selection for multi-label classification of music emotion [J]. International Journal of Computational Intelligence Systems, 2012, 5(4): 668 - 678.

[16] 高文,钱亚冠,吴春明. 网络流量特征选择方法中的分治投票策略研究[J]. 电子学报,2015,43(4):795 - 799.
GAO Wen, QIAN Ya-guan, WU Chun-ming. The divide-conquer and voting strategy for traffic feature selection [J]. Acta Electronica Sinica, 2015, 43(4): 795 - 799. (in Chinese)

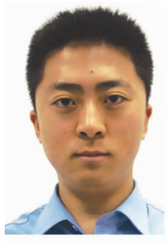
[17] 宁卓,孙知信,龚俭,等. 利用流量特征的 GIDS 报文分类优化算法[J]. 电子学报,2012,40(3):530 - 537.
NING Zhuo, SUN Zhi-xin, GONG Jian, et al. An improved GIDS packet classification algorithm using the characteristic of the traffic [J]. 2012, 40(3): 530 - 537. (in Chinese)

[18] RICH E, KNIGHT K. Artificial Intelligence [M]. New York, US: McGraw-Hill, 1991.

[19] HALL M A. Correlation-Based Feature Selection for Ma-

chine Learning [D]. Waikato, New Zealand: The University of Waikato, 1999.
[20] MOORE A W. Dataset [OL]. http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmet-rics/index.html, 2013 - 8 - 1.
[21] MOORE A, ZUEV D, CROGAN M. Discriminators for Use in Flow-Based Classification [R]. London: Queen Mary and Westfield College, Department of Computer Science, 2005.

作者简介



申 健 男,1988 年生,辽宁鞍山人. 2014 年毕业于空军工程大学信息与导航学院,获工学硕士学位. 现为空军工程大学博士研究生. 主要研究方向为流量分类识别、网络管理.
E-mail: shenjian2018@126.com

夏靖波 男,1963 年生,河北秦皇岛人,博士. 现为厦门大学嘉庚学院教授、博士生导师,主要研究方向为信息网络管理与安全.