

基于单因素方差分析的 P2P 流特征 向量优化方法

章鹏程

(南京邮电大学 信息网络技术研究所, 江苏 南京 210003)

摘要:文中针对 P2P 流量识别中流量特征向量选择的问题进行了研究,提出了一种基于单因素方差分析的 P2P 流量特征向量优化算法。该算法将统计学中单因素方差分析的方法引入 P2P 流量特征选取中,能够从高维的特征向量中选择出具有显著性作用的低维特征向量,从而实现 P2P 流特征向量的优化。实验结果表明,文中提出的方法可以在显著地提高 P2P 流量识别的效率的同时,将 P2P 流量识别准确率保持在一个可接受的范围内,为 P2P 流量识别的进一步研究提供了铺垫。

关键词:P2P;单因素方差分析;特征向量

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2012)04-0101-03

Optimal Method Based on One-Way Analysis of Variance for P2P Feature Vector

ZHANG Peng-cheng

(Institute of Information and Network Technology, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract:To acquire optimized feature of P2P flow, proposed a feature section method based on one-way analysis of variance. One-way analysis of variance in statistics, which can choose the most obvious factors from the high dimensional feature vector, was introduced to feature section. Experiment results demonstrate that with the use of the method efficiency of the P2P traffic identification is improved obviously and accuracy of P2P traffic identification is kept within acceptable limits. It is a setup for farther research.

Key words:P2P; one-way analysis of variance; feature vector

0 引言

P2P 流量的特征向量选择是基于流量特征进行 P2P 识别的基础,选择精简、有效的特征向量是关键。目前研究人员提出了一些解决此问题的方法,但这些方法在识别效率或识别准确率方面还存在一些不足。文献[1]基于应用层软件状态机不同导致流量特征不同的事实,提出了基于统计指纹的概念的流量特征向量选择方法。该方法解决了流量特征向量冗余、识别效率低下的问题,但选择的流量特征的有效性没有经过验证,实验中并没有获得较高的准确率。文献[2]通过考察 P2P 流量的特征参数,提出了一个 248 维的特征向量。这一方法虽然可以有效识别 P2P 流量,但

由于流量特征向量维数过高导致识别效率低下。文献[3]利用实际观察结果概率和理论分析概率,结合统计学中卡方检验的方法来进行识别,但卡方检验自身的缺陷使其最终的识别准确率只有 80% 左右。旨在得到识别耗时短、识别准确率高的流量特征向量,文中将统计学方法单因素方差分析引入 P2P 流量特征选择,对文献[2]的特征向量进行优化,摒弃原特征向量中统计学上的非显著性因素。

1 单因素方差分析原理

单因素方差分析又称“变异数分析”或“ F 检验”,用于两个及两个以上样本均数差别的显著性检验。一个复杂的事物,往往有许多因素互相制约又互相依存,方差分析的目的是通过数据分析找出对该事物有显著影响的因素、各因素之间的交互作用以及显著因素的最佳水平。该方法将某一因素 W 在分组间的差异分成两个部分,一部分为组内差异,源于本类别内各样本

收稿日期:2011-09-09;修回日期:2011-12-15

基金项目:国家自然科学基金(61003237);江苏省普通高校自然科学研究资助项目(10KJB510018)

作者简介:章鹏程(1985-),男,硕士研究生,研究方向为认知网络监测与管理。

的 W 数值上的离散度;另一部分是组间差异,由分组区别和组内的数据差异两方面原因造成。不难理解,如果组间差异与组内差异的比值 F 接近于 1,说明 W 对分组造成的影响很小,其影响力相对于组内数据的随机性来说可以忽略,则认为 W 不是影响分组的显著性因素。反之, F 越大, W 对分组造成的影响越大。

根据统计学原理,组内差异 MS_w 可用组内方差描述,组间差异可以用组间方差描述,即:

$$MS_w = \frac{SS_w}{df_w} = \frac{\sum_{i=1}^t (n_i - 1) S_i^2}{df_w} \quad (1)$$

$$MS_b = \frac{SS_b}{df_b} = \frac{\sum_{i=1}^t n_i (\bar{X}_i - \bar{X})^2}{df_b} \quad (2)$$

其中, SS_w 和 SS_b 分别表示组内平方和与组间平方和; df_w 和 df_b 分别表示组内自由度和组间自由度; n_i 表示第 i 组样本数; S_i^2 表示第 i 组内数据的方差; \bar{X}_i 表示第 i 组样本的均值; \bar{X} 表示所有样本的均值; t 表示分组数。

则 F 可以表示成 $F = \frac{MS_b}{MS_w}$, 经证明 F 满足一个 F 分布^[4], 即:

$$F = \frac{MS_b}{MS_w} \sim F(df_b, df_w) \quad (3)$$

因此,单因素方差分析本质上是 F 检验问题,可以通过设置拒绝域和置信度,提出假设检验的方法,处理这类问题。

2 P2P 流量识别过程的特征向量优化方法

文中提出一种基于单因素方差分析的 P2P 流量特征向量优化方法。该方法旨在从文献[2]的 248 维特征向量 \vec{S} 中选择出 P2P 流量的显著性特征组成优化的特征向量 \vec{O} , 以达到降低特征向量维数,提高 P2P 流量识别效率,同时保证 P2P 流量识别准确率的目标。首先将所有样本分为两组:P2P 流和非 P2P 流,即取分组数 $t = 2$,并将特征向量 \vec{S} 包含 248 个属性看作影响分组的 248 个因素。然后,运用方差分析法逐个分析这 248 个因素是否为影响分组的显著性因素,最终将显著性因素组合成优化的特征向量 \vec{O} ,该过程可用图 1 描述。文中取 F 检验的置信度为 0.05,查询 F 检验表知 F 的阈值为 799。

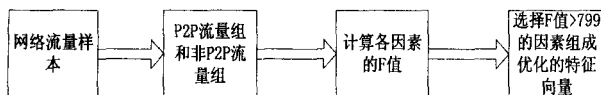


图 1 特征向量优化方法

根据提出的优化方法,将剑桥大学提供的公开数据集 Day3. TCP^[5] 中的数据样本分为 2 组,并导入统计学软件 SPSS^[6] 进行方差分析计算,得出每个属性的 F 值。最终,得到 12 个满足条件的属性,如表 1,从而组成了一个 12 维的优化的特征向量。

3 实验与结果分析

文中分别利用原特征向量 \vec{S} 和优化的 12 维特征向量 \vec{O} 对 P2P 流量进行识别,并从建模时间和识别准确率两方面观察实验结果。由于机器学习算法也是影响 P2P 流量识别的重要因素,所以为了排除机器学习算法对 P2P 流量识别的影响,增强实验结论的说服力,实验中采用了以下 5 种不同的机器学习算法^[7]: 随机森林 (Random Forest, RF)^[8]、朴素贝叶斯算法 (Naive Bayesian, NB)^[9]、贝叶斯网络算法 (Bayesian network, BN)^[10]、径向基函数神经网络 (RBF Network, RBF)^[11] 以及支持向量机算法 (Support Vector Machine, SVM)^[12]。

首先,对数据源进行预处理。将数据集 Day3. TCP 分成 2 部分记为 training1 集和 test1 集。然后,删除 training1 集和 test1 集中除表 1 中 12 个属性以外的所

表 1 方差分析结果

编号	缩写	描述	F 值
1	serv_port	服务器端口	23638.398
2	clnt_port	客户机端口	18636.850
45	Act_data_pkt_clnt	包含大于 1 字节的数据部分的 TCP 数据包总数 (客户机到服务器)	139189.411
59	push_pkts_clnt	TCP 首部设置的所有数据包总数 (客户机到服务器)	151702.206
60	push_pkts_serv	TCP 首部设置的所有数据包总数 (服务器到客户机)	7032.261
83	Min_seg_size_clnt	最小段大小 (客户机到服务器)	55599.070
86	Avg_seg_size_serv	平均段大小 (服务器到客户机)	4618.163
95	init_win_bytes_clnt	被送到初始窗口的总字节数 (客户机到服务器)	36840.091
96	init_win_bytes_serv	被送到初始窗口的总字节数 (服务器到客户机)	4558.660
113	RTT_samples_clnt	RTT 样本的总数 (客户机到服务器)	108020.467
162	IP_bytes_med_clnt	IP 数据包的平均字节 (客户机到服务器)	1207.062
180	data_bytes_var_serv	数据包字节数的方差 (服务器到客户机)	2764.662

有属性,得到两个优化数据集,分别记作 training2 集和 test2 集。显然,training1 集和 test1 集中流量样本的特征向量是原始的 248 维的,而 training2 集和 test2 集中流量样本的特征向量是经优化的 12 维的。

然后,比较特征向量对 P2P 流量识别建模时间的影响。分别利用 5 种机器学习算法对 training1 和 training2 进行建模,得到 10 个分类模型。

观察表 2,无论采用何种机器学习算法,优化数据集 training2 的建模时间上都要少于原数据集 training1 的建模时间。其中,除了 SVM 算法受特征向量维数影响较小外,采用其他 4 种算法时,建模时间均降低了 50% 以上。

表 2 建模时间比较(单位:秒)

	RF	NB	BN	RBF	SVM
training1	216.74	184.02	153.47	108.27	110.84
training2	27.68	3.16	52.43	28.16	80.69

最后,利用得到的分类模型对 test1 和 test2 中的样本进行分类,并计算每次识别的真阳性率(TPR)和假阳性率(FPR)。TPR 和 FPR 是评估算法准确度的常用参数,TPR 越高,FPR 越低,则算法的准确率就越高。

观察图 2,无论采取何种机器学习算法,采用优化数据集时,P2P 流量识别准确率都有所下降,这是因为文中的优化方法本质上降低了特征向量的维数,不可避免地忽略了流量的一些特征,从而影响了 P2P 流量识别准确率。进一步观察可以发现每种算法在运用优化的特征向量后,识别准确率仅有略微下降,真阳性率(TPR)的下降比例均在 5% 以内,假阳性率(FPR)的上升也都控制在了 3% 以内。所以,运用优化的特征向量后,虽然识别准确率都有了一定的下降,但在一般情况下,这种损失是可接受的。

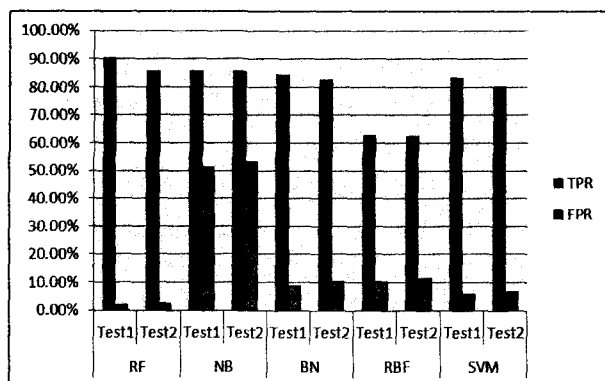


图 2 P2P 识别准确率

综上,对高维流量特征向量采取文中的优化方法虽然会损失一定的 P2P 识别准确率,但考虑到识别效率上的巨大收益,在一般情况下,识别准确率上的微小损失是可以接受的。

4 结束语

文中针对 P2P 流量特征选择问题进行了研究,提出了一种基于单因素方差分析的 P2P 流量特征向量优化算法。该方法将统计学中单因素方差分析的方法引入 P2P 流量特征选取中,能够从高维的特征向量中选择具有显著性的因素,从而实现对特征向量的优化。实验结果表明,该方法可以显著提高 P2P 流量识别的效率,同时将 P2P 流量识别准确率保持在可接受范围内。由于该方法会在一定程度上降低 P2P 流量识别的准确率,所以在某些对识别准确率敏感的环境中并不适用,因此,下一步的研究将关注在提高识别效率的同时防止识别准确率的下降。

参考文献:

- [1] Silvestre G, Fernandes S, Kamiński C, et al. Most Wanted Internet Applications: A Framework for P2P Identification[C]// Proceedings of the 2010 8th Annual Communication Networks and Services Research Conference. [s. l.]: [s. n.], 2010: 341-347.
- [2] Li Wei, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks, 2009(10): 790-809.
- [3] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques[C]//Proc. of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer System. [s. l.]: [s. n.], 2005: 50-60.
- [4] Howell J R. A computer technique for handling analysis of variance[C]//Communications of the ACM. [s. l.]: [s. n.], 1962: 433-434.
- [5] Internet Traffic Classification Using Bayesian Analysis Techniques[EB/OL]. 2011-07. <http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/index.html>.
- [6] 宇传华. SPSS 与统计分析[M]. 北京: 电子工业出版社, 2007.
- [7] Alpaydm E. Introduction to Machine Learning[R]. [s. l.]: MIT Press, 2010.
- [8] Xiao Jiamin. Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure[J]. BMC Bioinformatics, 2011(12): 165-166.
- [9] Dong Tao, Shang Wenqian, Zhu Haibin. Naive Bayesian Classifier Based on the Improved Feature Weighting Algorithm[J]. Communications in Computer and Information Science, 2011(6): 142-147.
- [10] Jr G C. Evaluating In-Clique and Topological Parallelism Strategies for Junction Tree-Based Bayesian Network Inference Algorithm on the Cray XMT[C]// IEEE International

(下转第 107 页)

的开关指令,并且只能处于关闭、开启(但未开始加热)和加热三种状态之一。设想控温器只向加热炉发送加热指令,而不再对加热炉是否有回应进行验证的情况。如果温控器向加热炉发送一个加热指令而此时加热炉处于关闭状态,那它就会漏掉该条指令。若它再被开启,它也只能按照默认开启状态运行而不会执行之前的加热指令除非收到下一条加热指令,而温控器并不会传送另一条指令因为之前发过,它会等待温度上升至设定点后再发送一条停止加热指令。

因此,为使系统能够捕捉这种错误并同时能给出一个补救提示,先单独对 FSP 做基础的分析。通过提出类似活跃度检查的方法来将问题推进。如果温控器发出一条加热指令,房间的温度最终会不会上升?在 LTSA 中很容易对活跃度进行分析,通过这种分析不仅能捕获上述错误,而且还能通过什么地方活跃度不达要求从而得出某种可能引起加热失败的原因。从而便可立刻就能发现这个错误,即原来是因为控温器未知加热炉的开关状态引起的,并很可能产生一个使加热炉向温控器汇报当前的状态的纠正措施。修正后的模型如图 2 所示,图中新增了一个加热炉与温控器之间的关闭命令连接器。

然而这种方案中如果加热炉是一个仅仅能接收指令而不能发送认证信号的哑装置就难以实现了。不过,还有另一个方法,这就是让温控器在某段时间(超时)间隙后检查房间温度是否上升到预期范围。如果温度没上升,就说明加热指令需要重发。因此,这种方案的成功依赖于对超时周期的选择,而这个周期可由加热和降温时的屋内温度变化率来确定。但时间和温度变化率都不能用纯离散的 FSP 模型表示。因此,可采用 LHA, LHA 中这二者都可以表示,用 LHA 中的 PHAVer 分析完后,通过加热或降温时温度变化率的大小范围就可以得到合适的超时周期的范围。

5 结束语

文中对 CPS 进行了分析并介绍了相关建模工具。通过这种建模方式,得出新的方案从而可以优化对系统行为特征建模。目前较多的研究集中在以不同视角对 CPS 进行建模,每种视角都对应某种数学抽象方法且不同视角的建模方式也会带来不同的系统性能。系

统建模方法和对应系统性能又产生了网络设计最优化问题,因为它们之间是通过公共变量和协议进行交互的。因此,只有研究出方法规范和解决这类互联设计问题,才能综合地对物理系统进行分析和设计。

参考文献:

- [1] 杨敬中,张广泉.面向方面的软件体系结构描述语言 AOP-ADL[J].计算机工程,2008,34(10):80-82.
- [2] 李长云,李赣生,何频捷.一种形式化的动态体系结构描述语言 D2ADL[J].软件学报,2006,17(6):1349-1359.
- [3] Garlan D, Monroe R T, Wile D. Acme: Architectural Description of Component-Based Systems[M]//Foundations of Component-Based Systems. Cambridge: Cambridge University Press, 2000: 47-68.
- [4] 刘长林,张广泉,黄静.一种基于 ACME 的面向方面软件体系结构设计方法[J].苏州大学学报(工科版),2011(2):6-13.
- [5] 郑凤梅,常会友.面向方面的建模方法的设计与应用[D].广州:中山大学,2005.
- [6] 李建中,高宏,于博.信息物理融合系统(cps)的概念、特点、挑战和研究进展[D].哈尔滨:哈尔滨工业大学,2010.
- [7] Dvorak D, Rasmussen R, Reeves G, et al. Software Architecture Themes in JPL's Mission Data System[C]//AIAA Space Technology Conference and Expo. Albuquerque, NM: [s. n.], 1999.
- [8] Rajhans A, Cheng Shangwen. An Architectural Approach to the Design and Analysis of Cyber-Physical Systems[D]. Pittsburgh: Carnegie Mellon University, 2009.
- [9] Magee J, Kramer J. Concurrency: State Models and Java Programming[M]. 2nd ed. [s. l.]: Wiley, 2006.
- [10] 顾庆,陈道蓄,谢立,等.基于有限状态进程的事件约束定义[J].软件学报,2002,13(11):2162-2167.
- [11] Henzinger T A. The theory of hybrid automata[C]//Proc of 11th Annual IEEE Symposium on Logic in Computer Science, LICS'96. New Brunswick, New Jersey: IEEE Computer Society Press, 1996.
- [12] Frehse G. PHAVer: Algorithmic Verification of Hybrid Systems Past HyTech [C]//Proceedings of the Fifth International Workshop on Hybrid Systems: Computation and Control (HSCC), Lecture Notes in Computer Science 3414. [s. l.]: Springer-Verlag, 2005.

(上接第 103 页)

- Symposium on Parallel and Distributed Processing Workshops and PhD Forum. [s. l.]: [s. n.], 2011.
- [11] Zhang Yan, Zhang Li, Xing Guolin, et al. Predictive Control of Nonlinear System Based on MPSO-RBF Neural Network[J]. Communications in Computer and Information Science, 2011

(2): 567-573.

- [12] Moguerza J M, Muñoz A, Psarakis S. Monitoring Nonlinear Profiles Using Support Vector Machines[J]. Lecture Notes in Computer Science, 2007(6): 574-583.