



其他

基于多层级联算法的网络业务流量识别技术

黄凤,王瑶,黄莉,姚继明,刘世栋

(国网智能电网研究院,江苏 南京 210003)

摘要:当前网络的主要问题是流量需求和网络带宽之间的矛盾,流量管理是有效的解决方案之一,同时能提升网络的服务质量,而网络业务流量识别是进行流量精细化管理的基础。在研究网络业务流量识别和机器学习技术的基础上,提出基于多层级联机器学习算法的网络业务流量识别方法,该方法能适应网络业务流量复杂多变的情况,并且平衡了机器学习算法在时间性能和识别准确率之间的矛盾,进一步地,该方法在级联算法上考虑了代价不均衡问题,更为关注重要业务的识别,有助于提升用户体验。

关键词:流量识别;智能管道;多层级联算法;机器学习

中图分类号: TN915.81

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2015399

Traffic identification of network business based on multi-layer cascade algorithm

HUANG Feng, WANG Yao, HUANG Li, YAO Jiming, LIU Shidong

State Grid Smart Grid Research Institute, Nanjing 210003, China

Abstract: The contradiction between traffic demand and network bandwidth is the main problem of the current network. Traffic management is one of the effective solutions, at the same time it can improve the quality of service of the network, and network traffic identification is the basis of fine traffic management. The traffic identification method of network business based on multi-layer cascade algorithm was proposed, after the research of network traffic identification and machine learning techniques. This method can adapt to the situation of the complex network traffic flow, and can balance the contradiction between the time performance and the accuracy of the machine learning algorithm. Furthermore the cascade algorithm of this method takes the cost imbalance into account, and takes more attention to the identification of important business, so it can help enhance user experience.

Key words: traffic identification, smart pipeline, multi-layer cascade algorithm, machine learning

1 引言

随着互联网应用的不断丰富,网络中的数据种类和流量都呈几何级数膨胀^[1],网络流量需求与网络带宽之间的

矛盾日益增大,网络流量管理的作用显得尤为重要。在尽力而为(best-effort service)的服务模式下,所有网络流量都一视同仁地被尽可能快地传递,网络不提供任何时延或可靠性保障^[2];这将导致 P2P 文件传输等非关键流量消耗着

收稿日期:2015-11-20;修回日期:2015-12-01

基金项目:国家电网公司科技项目“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其应用”

Foundation Item: The Science and Technology Project of State Grid Corporation of China “Research and Application of Key Technologies for Traffic Prediction and Smart Pipeline in Power Communication Network”



大量的带宽,而关键业务流的服务质量却没有得到保证,用户和业务流量的大幅增长给网络服务质量 (quality of service, QoS)^[3,4]带来了巨大的压力,特别是当大量次要业务流量(如 P2P 下载等)抢占有限的网络资源,严重影响一些主要业务的正常开展时,问题更为突出。

提升网络 QoS 保障,简单增加网络带宽是不够的,带宽资源始终是有限的,无法保障不被耗尽,而且提升带宽还会带来巨大的成本。在这种情况下,保障网络的 QoS 是业界的研究热点,例如 IETF 先后提出了综合业务模型 IntServ^[5]和区分服务模型 DiffServ^[6],取得了一定效果,但是可操作性和可扩展性却不尽如人意。与此同时,网络流量管理逐渐受到重视,并成为一种有效的 QoS 解决方案,其核心是流量识别与资源管理,通过对流量的识别,将不同业务的网络流量按照预定的规则、策略和优先级等,采用不同方式进行处理、转发、限速和整形,以实现对网络资源有计划地分配和利用,尽可能减少网络拥塞,限制非关键业务占用过多带宽,从而保障关键业务的服务质量。与此理念相同的是,为优化网络管理以及提高网络资源利用率,业界提出“智能管道”技术,旨在根据用户的需求以及不同类型的应用传输数据,智能区分不同的 QoS 需求,动态调整管理策略,进行精细化的带宽资源管控,达到“用户可识别、业务可区分、流量可调控、网络可管理”的目的。由于网络流量的复杂性,网络业务流量识别相对于用户识别更具挑战性,而网络业务流量识别是进行流量调控和差异化、精细化网络管控的基础。

2 网络业务流量识别

在网络业务流量识别方面,当今的识别方法大体分为以下几种。

- 基于端口的静态流量识别方法:该方法采用 TCP/UDP 端口来识别各种不同的应用程序流量,随着端口动态变化业务的不断出现以及诸多业务采用常规业务端口作为隧道穿越,该方法也变得越来越低效和不准确。
- 基于净荷特征的流量识别方法:该方法采用深度分组检测(deep packet inspection, DPI)技术匹配各种应用特征,从而识别出不同的多媒体业务流量。然而随着私有协议业务、匿名加密或者 VPN 业务的不断出现,这种方法就不适用了。
- 基于流量特征统计的识别方法:该方法不涉及具体

净荷内容,而是通过对业务流的各种统计信息的分析来区分不同业务流量,如流内分组大小分布、分组间到达间隔、平均流存货周期、连接特性等,流量统计特征识别法最大的缺点是无法做到高准确率的识别。

- 跨层业务识别方法:该方法综合净荷特征识别和流量特征统计识别方法,在一定程度上弥补了流量特征统计方法准确性上的不足,然而对于新增业务或者业务特征变化的情况,如匿名加密 VoIP、动态变化且特征不规则的 P2P 多媒体信息流等,仍无法适用。
- 基于智能计算的流量识别方法:该方法通过自学习的方式适应业务的动态变化特征,特别是针对“流变种”业务和新增业务的情况,一般采取基于机器学习和模式识别的方式,针对流量特征识别业务类型。该方法具有较强的适应能力,对于有监督学习模式下,识别准确率受学习样本以及算法参数设置的影响,需要结合流量特征选择合适的算法类型和参数配置,目前,该类流量识别方法仍属于前沿类研究热点。

以流量控制为目标的流量识别,需要对不同业务流量进行区分管理。实时网络流量的分布具有随机性,业务类型并不均衡,此外,重要业务和次要业务的识别错误代价也不同。基于以上需求,本文设计并实现了一种基于多层级联机器学习算法的网络业务流量识别方法,该方法以网络流量的统计特征作为输入,充分利用流量类别的先验知识,采用有监督学习的方式,学习得到性能和效率较高的流量识别器。该方法采用多层级联的方式,优先识别较容易被识别的流量类型,识别代价高的流量则被推送到下一级别的识别器中,这种方式可以保障识别器的时间性能,不至于因为整体流量类型的复杂性导致整个识别器的时间性能的下降,识别时间是实时流量识别和流量管理至关重要的一点;此外,层级之间的级联算法充分考虑了重要业务被错分类的代价高于次要业务也被错分类的情况,有助于保障重要业务的 QoS 特性。

3 多层级联机器学习算法框架

3.1 有监督机器学习的流量识别模式

基于有监督机器学习方式进行流量识别的一般过程为,首先采集样本数据进行学习,得到识别器,而后基于测试数据对识别器的性能进行测试,如果算法的准确率和召

回率等性能参数符合预期目标,则放弃学习;如果不符合预期目标,则需要修改样本数据或学习算法继续学习,直到得到性能参数优异的识别器,然后基于识别器进行网络流量业务识别,图 1 为通用的机器学习识别过程。

3.2 多层级联学习的流量识别模式

基于多层级联机器学习算法 TrafCasd 的网络业务流量识别过程与图 1 类似,其核心在于该方法构建了一组多层级联式学习器,不同层级的学习器通过学习某种流量特征的内在模式,从而实现某些类型业务流量的快速识别。设在第 i 级的学习器为 h_i ,当流量特征数据 x 到达第 i 级后,若 h_i 能够以高置信度确定其为某一类型业务流量,则将 x 标记为该业务类型,若不能以高置信度进行识别,则送交下一级学习器进行判断,直到最终获得识别结果为止。基于级联的设计,使得不同层级的学习器可以关注不同的模式,从而使容易被识别的业务流量在前一两个层级就能被正确识别,而级联结构中偏后的学习器则负责关注如何识别更加难以识别的流量模式,而不会受到前几个层级识别出的较为容易识别的流量特征的影响,如图 2 所示。

为了进一步考虑业务流量的错分代价,本文提出的多层级联流量识别算法可以在每一层级中根据错误识别代价,调整学习器的输出和级联之间的权值,从而体

现出重要业务流量的代价敏感性。本文采用的级联结构是 Asymmetric AdaBoost^[7,8],它采用了考虑代价敏感的级联关系:

$$ALoss(i)=\begin{cases} \sqrt{k} & , y_i=1 \ \& \ C(x_i)=-1 \\ \frac{1}{\sqrt{k}} & , y_i=-1 \ \& \ C(x_i)=1 \\ 0 & , \text{其他} \end{cases} \quad (1)$$

其中, $C(x_i)$ 是学习器给流量特征数据的识别标记,重要业务被错分为次要业务的代价是次要业务被错分为重要业务的代价的 k 倍,将式(1)所示的错误代价嵌入传统 AdaBoost 优化目标中:

$$\prod_i Z_i = \sum_i \exp(-y_i \sum_i h_i(x_i)) \quad (2)$$

推导得出非对称级联的优化目标:

$$\prod_i Z_i = \sum_i (\exp(-y_i \sum_i h_i(x_i)) \exp(y_i \lg \sqrt{k})) \quad (3)$$

多层级联流量识别算法的每一层级的学习方法和采用的特征可以根据具体的应用场景需求进行针对性的调整,最初的几个层次可以采用识别速度快的方法,如线性支持向量机 (SVM),后面的层级中则采用相对复杂的模

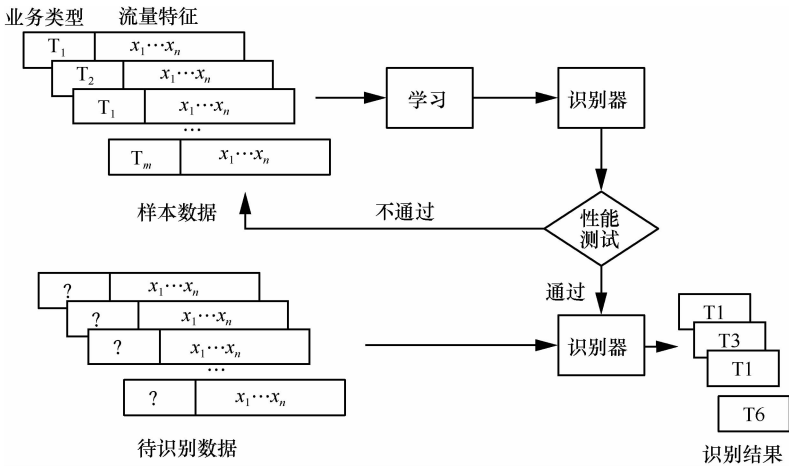


图 1 基于机器学习的流量识别过程

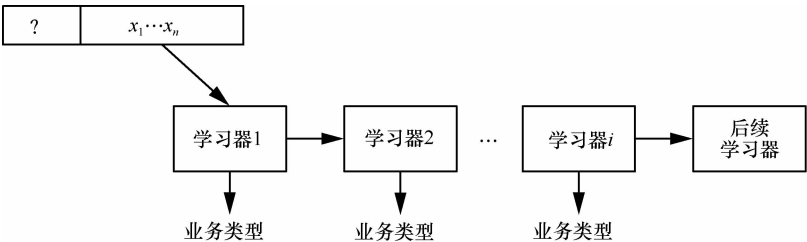


图 2 多层级联算法的流量识别过程



型,如神经网络或多学习器集成方法。每一层级的流量特征的选取也可根据需要进行调整,以优化算法性能。当网络中出现“流变种”业务和新增业务时,多层级联流量识别方法可以通过重新获样本数据进行学习,以达到正确识别“流变种”业务和新增业务的目的。

4 实验分析

本文采用 Java 语言构建流量识别实验环境,基于多层级联流量识别算法 TrafCasd 开发实现了多种机器学习算法的级联,包括支持向量机、Rtree、Boost、 k 近邻以及神经网络,在学习阶段需要设置算法的层级个数、每一层级的算法类型以及流量特征。选取包括网络类型、APN、浏览工具在内的 38 项流量特征进行业务识别,部分流量特征见表 1。

为测试算法性能,选取支持向量机、Rtree、Boost 以及神经网络单算法识别作为对比。多层级联流量识别算法 TrafCasd 设置为 3 层结构,3 层算法逐层复杂,分别为 SVM、Rtree 以及神经网络算法。为获得更为客观的实验结果,采用多倍交叉验证的方法获取算法性能,即将流量特征数据源等分成若干份,取其中一份作为测试数据,其余作为训练数据,调整不同的组合,再进行训练和测试,求取平均值获得最终结果,本文采用 5 倍交叉验证法。本文采用流量识别的预测准确率(precision)、召回率(recall)以及综合准确率和召回率的 F1 作为评价指标,实验结果如图 3 所示。

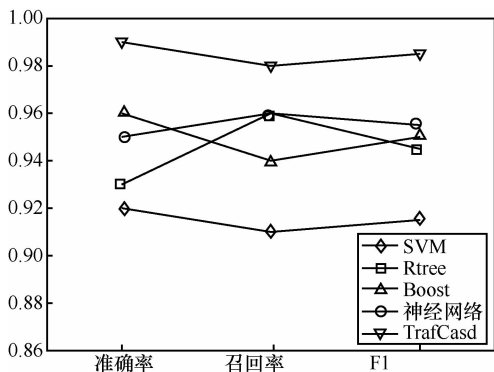


图 3 实验结果对比

多层级联流量识别算法可用性强,可根据具体的网络情况挑选合适的算法组合,经过多次训练和测试,可以获得性能优异的流量识别器。图 3 表明,基于多种算法级联的流量识别方法在性能上优于单个机器学习算法,多层级

联的方式均衡了不同类型机器学习算法的优缺点,如 SVM 简单快速但是识别效率低,可作为第一层级算法,而神经网络算法经优化后识别率高,但是在时间性能上不如 SVM,可作为后面层级的算法,使得容易识别的业务流量被 SVM 快速识别,而难以识别的业务流量被后续的复杂模型识别,以此获得在时间性能和识别率上的平衡,为实时流量管理提供技术保障;此外,由于在级联算法上考虑了代价敏感性,TrafCasd 算法更为重视重要业务的识别准确率,因此在重要业务的 QoS 保障方面具有突出优势。

表 1 部分流量特征

流量特征	名称
NetType	网络类型
APN	APN
Phone	用户号码
BrowsingT	浏览工具
ProtocolT	传输层协议
LAC	LAC
UserIP	用户端 IP 地址
SerIP	服务端 IP 地址
SendPKTS	上行分组数
CI	CI
SGSNIP	SGSNIP
SendRate	上行速率
RecvRate	下行速率
URL	URL
HostWeb	host
SendByte	上行流量
...	...

5 结束语

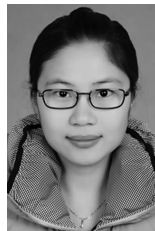
本文针对网络业务流量复杂多变、识别算法难以应对网络业务发展的局面,提出一种多层级联的流量识别算法,该算法采用非均衡代价敏感级联设计,一方面弥补了不同类型机器学习算法在时间特征和识别准确率方面此长彼消的矛盾,获得优异的识别性能;另一方面算法考虑不同等级业务的错分代价,级联算法充分考虑了重要业务被错误识别为次要业务的代价是次要业务被错误识别为重要业务的多倍关系,使得算法更为重视重要业务的识别。更为重要的是,在网络业务流量出现变化识别性能下降时,可通过重新采样学习和训练以提高算法的识别性能,具有很强的适应能力。在网络业务发展迅速的情况下,基于业务流量识别的流量管理是提高整个网络性能的关键。

键,本文提出的多层次联流量识别算法是基于智能计算流量识别方法中的一种,可为智能化的网络管理及“智能管道”提供技术保障。

参考文献:

- [1] 李明. Linux 的带宽管理技术研究与应用[D]. 成都:电子科技大学, 2009.
LI M. The research and application of bandwidth management for Linux[D]. Chengdu: University of Electronic Science and Technology of China, 2009.
- [2] SHELDON T. Encyclopedia of networking & telecommunications[M]. New York: McGraw-Hill, 2001.
- [3] MARCHESE M. QoS over heterogeneous networks [M]. New Jersey: Wiley, 2007.
- [4] EVANS J, FILSFILS C. Deploying IP and MPLS QoS for multiservice networks: theory and practice [M]. California: Morgan Kaufmann, 2007.
- [5] An architecture for differentiated services: IETF RFC 2475[S/OL]. (1994-06-11) [2015-10-20]. http://wenku.baidu.com/link?url=S4zppGzxZTfp_GmDsiCfoHuhPiES6Ew_qmlziWCAGIt7eY9xRdSkjd5ys1_fh6_Nqz4mwGu9c8i9LF5eh_5oabKMkr3VCuoOpxp7-CURWyUq&qq-pf-to=pcqq.c2c.
- [6] Integrated services in the internet architecture: an overview: IETF RFC 1633 [S/OL]. (1998-12-20) [2015-10-20]. http://wenku.baidu.com/link?url=kteAi0a0LiupeCGVaXrIDo_C2MdJW_GJg3T3NjlpV7d7EsDK-cQU_7aiuZZ_NbrznkMzmOfKTnlhBw69haQIGqSUrmFTBipTgB9i8nIna_9Gl3&qq-pf-to=pcqq.c2c
- [7] MASNADI-SHIRAZI H, VASCONCELOS N. Asymmetric boosting [C]// The 24th International Conference on Machine Learning, June 20-24, 2007, Corvallis, Oregon, USA. New York: ACM Press, 2007: 609-619.
- [8] WEN J B, XIONG Y S. Asymmetric constraint optimization based adaptive boosting for cascade face detector [C]// The 7th International Conference on Advanced Intelligent Computing Theories and Applications: with Aspects of Artificial Intelligence, August 11-14, 2011, Zhengzhou, China. Berlin: Springer-Verlag Berlin Heidelberg, 2012: 226-234.

[作者简介]



黄凤(1984-),女,国网智能电网研究院工程师,主要研究方向为电力信息及信息通信技术。

王瑶(1986-),女,国网智能电网研究院工程师,主要研究方向为电力通信及信息通信技术。

黄莉(1986-),女,国网智能电网研究院工程师,主要研究方向为电力信息及信息通信技术。

姚继明(1986-),男,国网智能电网研究院工程师,主要研究方向为电力通信技术。

刘世栋(1971-),男,国网智能电网研究院高级工程师,主要研究方向为电力通信技术。