轻工学报® 2020 年 7 月 第 35 卷 第 4 期
JOURNAL OF LIGHT INDUSTRY Vol. 35 No. 4 July 2020

引用格式:王宣立,张安琳,黄道颖,等. SDN 环境下不同机器学习算法的网络流量分类分析[J]. 轻工学报,2020,35(4):96 – 102.
中图分类号:TP393 **文献标识码**:A
DOI:10.12187/2020.04.013
文章编号:2096 – 1553(2020)04 – 0096 – 07

SDN 环境下不同机器学习算法的网络流量分类分析

Network traffic classification analysis of different machine learning algorithms in SDN environment

王宣立[1],张安琳[2],黄道颖[1],董帅[1],刘江豪[1]
WANG Xuanli[1],ZHANG Anlin[2],HUANG Daoying[1],DONG Shuai[1],LIU Jianghao[1]

1. 郑州轻工业大学 计算机与通信工程学院,河南 郑州 450001
2. 郑州轻工业大学 工程训练中心,河南 郑州 450001
1. *College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
2. *Engineering Training Center, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*

关键词:
软件定义网络;网络流量分类;机器学习;梯度提升决策树;Moore 数据集

Key words:
softward defined network (SDN); network traffic classification; machine learning; gradient boosting decision tree; Moore dataset

摘要:为对比分析软件定义网络(SDN)环境下不同机器学习算法的网络流量分类效果,对 Moore 数据集进行了平衡处理,在机器学习平台 RapidMiner 上对 K – 近邻(KNN)、随机森林(RF)、支持向量机(SVM)和梯度提升决策树(GBDT)4 种经典机器学习算法选取不同的分类特征进行分类实验. 实验结果表明,较其他 3 种算法,GBDT 算法可以在较短的时间内获得更好的分类效果.

收稿日期:2020 – 03 – 31
基金项目:河南省重点科技攻关项目(132102210418)
作者简介:王宣立(1995—),男,河南省嵩县人,郑州轻工业大学硕士研究生,主要研究方向为软件定义网络.
通信作者:黄道颖(1967—),男,河南省信阳市人,郑州轻工业大学教授,博士,主要研究方向为计算机网络、分布式计算.

Abstract: In order to compare and analyze the network traffic classification effect of different machine learning algorithms in the software defined network (SDN) environment, the Moore dataset was balanced, and four classic machine learning algorithms including KNN, random forest (RF), support vector machine (SVM) and gradient lifting decision tree (GBDT) were supported on the machine learning platform RapidMiner to select different classification features for classification experiments. Experimental results showed that compared with the other three algorithms, the GBDT algorithm could obtain better classification results in a shorter time.

0 引言

随着网络技术的发展、网络规模的扩大、网络协议的增多,以 TCP/IP 协议族为核心的传统网络变得臃肿不堪,管理难度越来越大. 针对传统网络的弊端,软件定义网络 SDN(softward defined network)应运而生,因其开放、灵活的特点而被视为下一代的互联网架构.

SDN 通过解耦传统网络中的控制功能与数据转发功能,形成了集中式的控制平面和由分布式 OpenFlow 交换机组成的数据平面,且两个平面之间依据 OpenFlow 规范进行通信[1]. 相较于传统网络,SDN 的集中式控制平面在对整个网络资源的调度、网络的管理和配置等方面都具有巨大的优势. Google 公司通过 SDN 技术部署的 B4 网络更是展示了 SDN 的巨大潜力.

SDN 注重对网络的可重构能力,致力于整个网络处理逻辑的可编程性和网络设备的白盒化,随着人工智能等技术的发展,学者们对 SDN 的智能化方面进行了大量的尝试. SDN 的架构决定了其可通过在应用层添加功能组件,十分便捷地实现特定需求. 吴艳[2]发现利用 SDN 可以获得网络全局视图、感知链路负载情况的特点,并使用神经网络获得流量分类模型,为不同类型的网络流量提供差异化的 QoS 质量保证;胡孟婷[3]通过在控制器中写入流量态势收集模块,利用 LSTM 算法预测网络流量态势,获得较好的预测效果;刘佳美等[4]提出的 PPME 模型,利用最大熵算法对网络流量变化情况进行预测,优化了分布式 SDN 控制平面 SHLB 的负载问题,在离线数据集上具有优异的表现.

在网络技术发展的同时,网络规模在新型应用的影响下发生了变化,移动互联网、云计算、大数据等的出现使得网络流量在近几年呈指数式增长,网络流量特征也随之发生了变化. 传统网络对网络流量进行更细粒度流量转发时捉襟见肘,且用户无法根据需要自行定义转发规则,因此对具有不同要求的网络流量无法实现差异化的 QoS 保障. 对网络流量进行识别并提供相应的保障机制,会使网络更高效地运行,对于网络管理、网络安全和网络计费等都具有重要的意义. 利用 SDN 集中控制、易于获取链路状态、流量统计信息等特点,进行网络流量识别研究非常具有可行性,而机器学习在流量识别分类问题上具有更高的准确性.

鉴于此,本文拟选取几种经典机器学习算法在网络流量数据集 Moore[5] 上进行网络流量识别分析,验证在 SDN 环境中不同机器学习算法的识别效果,为在 SDN 网络中选取合适的机器学习算法提供参考.

1 基于机器学习的网络流量识别

网络技术的发展影响着网络流量分类方法的发展. 最初的网络流量分类方法是通过国际互联网代理成员管理局注册的端口映射表,将特定网络应用与端口绑定,通过查询数据包中的端口号信息即可获悉网络流量类型,这种方法被称为基于端口的网络流量分类方法[6]. 该方法实现简单,可以快速识别网络流量类型,因此在高速网络流量环境中具有较广泛的应用.

但随着随机端口技术、P2P 应用等新型网络应用的出现,基于端口的网络流量分类准确率受到了极大的影响,一般只有 50% ~70%[6].

针对基于端口的网络流量分类方法存在因端口伪装技术无法准确识别网络流量的问题,研究者们提出了基于载荷的识别方法[7]. 通过分析数据包中有效负载的应用签名进行分类,可以获得较高的准确率. 但基于载荷的网络流量识别方法在提高准确率的同时,牺牲了识别速度,且提取有效负载的方式会侵犯用户隐私安全,对于加密流量的检测效果并不理想,不适用于高速网络环境的流量分类.

近年来,随着机器学习技术的发展,人们将机器学习应用到网络流量识别中,并取得了很多成果. 文献[8 -9]提取了网络流的子流并统计其流量特征,结合机器学习方法实现了在高速网络环境中的网络流量准确识别. 文献[10]结合模糊集合理论和 K-means 分类方法,改善了 K-means 分类方法初始聚类中心选取困难的问题,并降低了聚类迭代次数,具有更快的收敛速度,在 Moore 数据集上取得了较好的分类效果. 文献[11]通过统计流量特征实现了 SSL 流量分类,根据 SSL 流量的前 3 个应用数据包的大小和传递方向,利用高斯混合模型(GMM)建立分类模型,可实现加密流量较准确的分类. 文献[12]使用 C4.5 决策树对 P2P 流量进行分类,还通过对 P2P 流量特征的分析识别出未知的 P2P 流量,相较于非监督学习流量分类方法,C4.5 决策树具有更高的准确性、更少的训练时间和识别时间. 支持向量机(SVM)学习方法在网络流量分类中具有较好的泛化能力和鲁棒性,但受限于监督学习需要大量标记样,文献[13]通过引入增量学习和半监督学习,对 SVM 进行优化,提高了 SVM 分类的准确度,但并未指出该分类方法的具体实施效果.

机器学习方法以其较高的准确率成为解决 SDN 场景中遇到的问题的新途径. SDN 集中式的控制平面决定了其极易受到 DDOS 攻击,文献[14]通过在控制器中增加流量统计模块,利用 C4.5 算法对 DDOS 攻击流量进行检测,相较于 SVM 和 KNN 算法,C4.5 对攻击流量具有更好的识别效果;文献[15]根据网络视频流量的特点选取分类特征,利用随机森林(RF)算法实现了 SDN 中对视频流量和下载流量的识别;文献[16]在 SDN 环境中采用集成学习算法,在对网络流量进行分类的前提下,利用强化学习对路由进行规划,从而使不同类型的网络流量获得了相应的 QoS 保障.

本文采用 K - 近邻(KNN)、SVM、RF 和梯度提升决策树(GBDT)这 4 种经典的机器学习算法进行对比实验. KNN 算法实现简单,在多分类问题上具有较好的分类效果;SVM 泛化能力较好,适合用于小数量的数据集;RF 算法是一种集成学习算法,分类准确率更高;GBDT 算法是对真实分布拟合得最好的算法之一,具有较高的分类速度和较好的鲁棒性.

2　数据集预处理

Moore 数据集是网络流量分类领域的经典数据集,涵盖网络流量多个方面详细的特征信息,为了使这些特征可以进行量化比较,A. W. Moore等[5]对其进行了傅里叶变换,这使得 Moore 数据集在网络流量分类实验中得以大量使用. 经分析,Moore 数据集中的大部分特征均为在 SDN 环境中通过功能模块或根据计数器的统计信息获得,故本文采用 Moore 数据集进行实验具有可行性.

Moore 数据集是在某骨干网络采集的10 个时间段的网络流量信息. 数据集中每个样本包含 248 个流量特征和 1 个流量类别标签,且对样本的每个特征进行详细描述. Moore 数据集共由 10 个子集组成,每个子集均包含对应的应

用,Moore 数据集各类型的流量样本统计如表 1 所示.

表 1 Moore 数据集各类型的流量样本统计

Table 1 Traffic sample statistics of various types in Moore dataset

流量类型	应用	数量/条	占比/%
WWW	www	328 091	86.906
MAIL	Imap,pop2/3,SMTP	28 567	7.567
BULK	FTP	11 539	3.056
DATABASE	Postgres,sqlnet oracle,ingress	2648	0.701
SERVICES	X11,dns,ident,ldap,ntp	2099	0.556
P2P	KazaA,bittorrent,gnutella	2094	0.555
ATTACK	Internet worm and virus attacks	1793	0.475
MUITIMEDIA	Windows media player,real	576	0.153
INTERACTIVE	Ssh,klogin,rlogin,telnet	110	0.029
GAME	Half-life	8	0.002
总计		377 526	100.000

由表 1 可以看出,在 Moore 数据集中,不同类型的应用样本比例并不均衡,如被标记为 WWW 和 MAIL 的两类数据样本共占整个数据集的 94.473%,而 GAME 类的样本比例仅为 0.002%,这样的数据集被称为不平衡数据集. 对于不平衡数据集,若直接对其进行分类实验,会因为数量较多的样本能正确分类而获得较高的准确率,但忽略了小数量样本的准确率,实际的分类效果并不理想,因此需要对该数据集进行预处理.

通常采用欠采样、过采样、改变分类算法、生成合成数据 4 种方法对不平衡数据集进行预处理. 欠采样是从多数样本类中随机抽取部分样本进行训练;过采样是在少数类样本中随机采样,增加少数类样本的数量以达到平衡数据集的效果;改变分类算法是在样本中引入代价函数,增大少数类样本的权重,对多数类样本的权重赋予较小的值,从而避免忽略少数类样本;生成合成数据是以从少数类中创建合成样本,从而增加少数类样本基数来平衡数据集的.

欠采样和过采样使用的随机采样方法实现简单,效果也比较理想,但这种通过改变数据集中样本数量的方式也改变了原有数据集的样本分布,可能会增大误差,从而影响分类效果;而改变分类算法引入的代价函数则会产生权重值难以确定的问题,故本文采用 Borderline-SMOTE 算法[17] 合成数据以平衡数据集. SMOTE 算法利用 KNN 算法从少数类样本选择出 k 个近邻后,从中随机选取 n 个样本进行特征值的随机线性插值,从而构造新样本,实现数据集的平衡. 但这种随机插值的方式增大了类间重复的可能性,因此需对 SMOTE 算法进行优化. Borderline-SMOTE 算法通过对少数类样本的边界样本进行随机插值处理,使合成的数据样本更有效. Borderline-SMOTE 算法的流程如图 1 所示.

Borderline-SMOTE 处理不平衡数据集的过程可分为两步:1)利用 KNN 算法对每个少数类样本随机选取 n 个近邻,若这 n 个近邻均为多数类样本,则说明可能是异常数据,将其标记为噪声类;若这 n 个近邻均为少数类样本,则将其标记为安全类,噪声类和安全类均不做处理;若这 n 个近邻中一半以上为多数类样本,则将其标记为边界类样本. 2)对于边界类样本,使用随机插值算法生成新的数据.

因本文中使用的 Moore 数据集中 GAME 类型与其他类型样本量差距太大,即使使用 Borderline-SMOTE 算法也无法避免样本重叠,所以在进行实验时,剔除了标记为 GAME 类型的样本,然后选取 WWW 类型部分样本和其他类型的全部样本组成新的训练数据集,采用 Borderline-SMOTE 算法对其进行平衡化处理. 平衡后的数据集各类型及其样本数量如表 2 所示. 因平衡后的数据集样本数量不足百万,属于小数量数据集,所以将数据集的 80% 作为训练数据集,剩余的 20% 作为测试数据集.

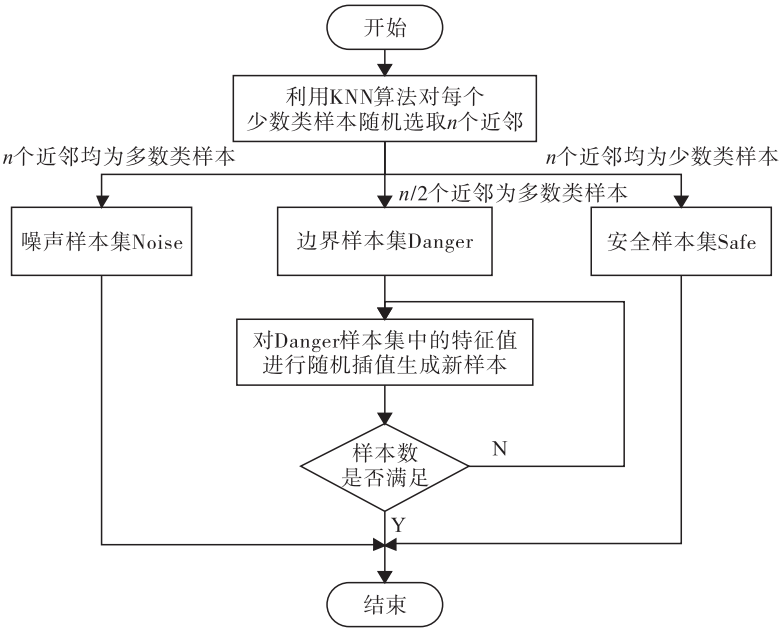

图 1　Borderline-SMOTE 算法流程

Fig. 1　Borderline-SMOTE algorithm flow

表 2　平衡后的数据集类型及其样本数量

Table 2　The number of samples of each type in the balanced dataset

流量类型	数量/条
WWW	65 000
MAIL	65 000
BULK	30 000
DATABASE	30 000
SERVICES	30 000
P2P	30 000
ATTACK	20 000
MUITIMEDIA	10 000
INTERACTIVE	10 000
总计	270 000

3　分类特征的选取

虽然 Moore 数据集的每个样本都有 248 个特征，但其中有 100 多个特征是通过傅里叶变换得到的，再用于实时的网络流量分类，网络设备中待分类数目可能达到数十万条，若对所有的数据流进行傅里叶变换，对于硬件设备会是一个巨大的挑战. 文献[18]指出，对网络流量进行统计后也可获得可靠性较好的特征，将端口号、有效负载、流量统计信息等特征结合可以获得更好的分类效果. 为选取分类效果更好的机器学习算法，本文从 Moore 数据集中选取了 5 组分类特征，分别为端口号、数据包数量、数据包大小、时间相关特征和数据包信息标志位，具体内容如表 3 所示.

4　实验结果与分析

本实验的硬件环境为 AMD r7 2700X 的 CPU，AMD rx580 显卡，所用的数据分析平台为 RapidMiner. 采用 KNN、SVM、RF 和 GBDT 这 4 种经典的机器学习算法，选取不同的特征进行分类实验，以分析 SDN 环境下不同算法的分类效果，以及不同特征对分类效果的影响.

不同算法对各种特征的分类准确率如图 2 所示. 由图 2 可以看出，在对网络流量进行分类时，不同的特征对识别准确率的影响很大，基于端口号的识别方法、基于时间信息的识别方法和基于信息标志位的识别方法无论在何种机器学习方法下识别率均不高，而将数据包数量、数

据包大小等网络流量统计信息作为分类特征,则会有较好的分类效果;在 4 种分类算法中,相

表 3 选取的分类特征

Table 3 Selected classification features

特征分组	特征缩写	说明
端口号	server-port	服务器端口号
	client-port	客户端端口号
数据包数量	total_pkts	数据包总数
	actual_data-pkts	携带负载的数据包总数
	rexmt_data-pkts	重传的数据包总数
	URGENT_data-pkts	首部含有 URG 标志的数据包总数
	zwnd_probe-pkts	窗口探测数据包总数
	outoforder_pkts	未按序到达的数据包总数
数据包大小	min/med/mean/max_data_wire	以太网数据包长度的最小、中位数、平均值和最大值
	min/med/mean/max_data_ip	IP 数据包负载长度的最小、中位数、平均值和最大值
	min/med/mean/max_data_ctrl	TCP 控制报文首部长度的最小、中位数、平均值和最大值
	init_wnd_bytes	在 TCP 初始窗口发送的字节总数
时间相关特征	duration	连接持续时间
	time_spent_in_bulk	单向传输数据的总时间
	time_spent_in_idle	无数据传输的总空闲时间
	min/max/avg/sdv_retr_time	在所有重传的数据包中,任意两个数据包间隔时间的最小值、最大值、平均值和均方差
数据包信息标志位	ack/SYN/FIN_pkts_sent	TCP 首部含有 SYN(FIN)标志位的数据包总数

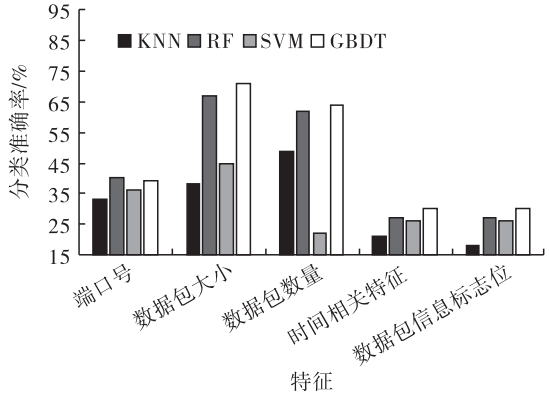

图 2 不同算法对各种特征的分类准确率

Fig. 2 Accuracy of various eigenvalues of different algorithms

较于 KNN 算法和 SVM 算法,RF 和 GBDT 两种算法的分类效果更为理想,这是因为这两种算法均采用了集成学习的方式构造分类器.

不同算法对组合特征的分类准确率如图 3 所示. 由图 3 可以看出,在获得足够的样本特征时,4 种算法都有较好的分类效果,相差不大. 结合不同算法的训练时间(见图 4),RF 算法和 GBDT 算法均在较少的时间内获得了 97% 以上的分类准确率,且 GBDT 算法可以在最少的训练时间内获得最好的分类效果.

5 结语

本文在机器学习平台 RapidMiner 上验证了 SDN 环境下 KNN、SVM、RF 和 GBDT 这 4 种算法的网络流量分类效果. 实验结果表明,GBDT 算法利用每次学习的残差进行迭代训练,相较于其他 3 种算法具有更好的分类效果和较短的训练时间. SDN 的网络可重构能力为将 GBDT 算法作为功能模块融入到整个网络的

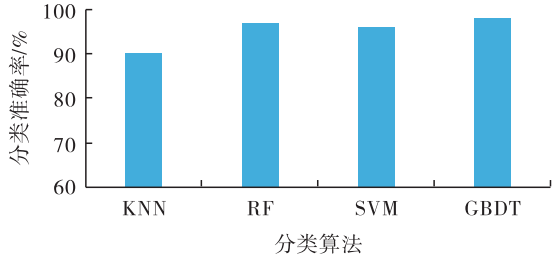

图 3 不同算法对组合特征的分类准确率

Fig. 3 Accuracy of combined feature classification of different algorithms

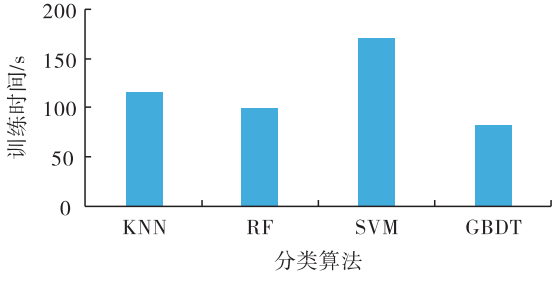

图 4 不同算法的训练时间

Fig. 4 Training time of different algorithms

管理策略中提供了可行性,两者的进一步融合是后续研究的重点.

参考文献:

[1] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow: Enabling innovation in campus networks[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69.

[2] 吴艳. 基于流量分类的智能 SDN 路由优化技术研究[D]. 杭州:浙江工商大学,2019.

[3] 胡孟婷. SDN 网络流量态势评估及预测技术研究[D]. 成都:电子科技大学,2019.

[4] 刘佳美,徐巧枝. 基于机器学习的 SDN 网络流量预测与部署策略[J/OL]. 计算机工程: 1-7[2020-03-16]. https://doi.org/10.19678/j.issn.1000-3428.0056436.

[5] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network applications[C]// Proceedings of International Workshop on Passive and Active Network Measurement. Heidelberg: Springer, 2005: 41.

[6] 彭芸,刘琼. Internet 流分类方法的比较研究[J]. 计算机科学,2007(8):58.

[7] SEN S, SPATSCHECK O, WANG D. Accurate, scalable in-network identification of p2p traffic using application signatures[C]// Proceedings of the Web Conference. Manhattan: [s. n.], 2004: 512.

[8] NGUYEN T T T, ARMITAGE G. Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks[C]// Proceedings of 2006 31st IEEE Conference on Local Computer Networks. Piscataway: IEEE, 2006: 369.

[9] BERNAILLE L, TEIXEIRA R. Early recognition of encrypted applications[C]// Proceedings of International Conference on Passive and Active Network Measurement. Heidelberg: Springer, 2007: 165.

[10] 吴辉. 基于模糊 K-Means 的网络流分类系统研究与实现[D]. 广州:广东工业大学,2016.

[11] LI J, ZHANG S, LU Y, et al. Real-time P2P traffic identification[C]// Proceedings of 2008 IEEE Global Telecommunications Conference. Piscataway: IEEE, 2008: 1.

[12] QUINLAN J R. C4.5: Programs for machine learning[M]. Amsterdam: Elsevier, 2014.

[13] 李平红,王勇,陶晓玲. 支持向量机的半监督网络流量分类方法[J]. 计算机应用,2013,33(6):33.

[14] 何建涛. SDN 中基于机器学习的 DDoS 攻击检测与防御方法研究[D]. 合肥:安徽大学,2019.

[15] 李兆斌,韩禹,魏占祯,等. SDN 中基于机器学习的网络流量分类方法研究[J]. 计算机应用与软件,2019,36(5):75.

[16] 王赋翼. 机器学习在流量分类中的应用[D]. 成都:电子科技大学,2019.

[17] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]// Proceedings of International Conference on Intelligent Computing. Heidelberg: Springer, 2005: 878.

[18] 张龙璨,柳斌,李芝棠. 机器学习分类下网络流量的特征选取[J]. 广西大学学报(自然科学版),2011,36(S1):6.