doi:10.11887/j.cn.202004006 http://journal.nudt.edu.cn

基于深度森林的网络流量分类方法*

戴 瑾[1,2],王天宇[2,3],王少尉[2]
(1. 南京大学金陵学院 信息科学与工程学院,江苏 南京 210089;
2. 南京大学 电子科学与工程学院,江苏 南京 210023;
3. 东南大学 国家移动通信研究实验室,江苏 南京 210096)

摘 要:随着网络应用的迅猛发展,流量分类在网络资源分配、流量调度和网络安全等诸多研究领域受到广泛关注。现有的机器学习流量分类方法对流量数据特征的选取和分布要求苛刻,导致在实际应用中的复杂流量场景下分类精确度和稳定度难以提高。为了解决样本特征属性的复杂性给分类性能带来的不利影响,引入了基于深度森林的流量分类方法。该算法通过级联森林和多粒度扫描机制,能够在样本数量规模和特征属性选取规模有限的情况下,有效地提高流量整体分类性能。通过网络流量公开数据集 Moore 对支持向量机、随机森林和深度森林机器学习算法进行训练和测试,结果表明基于深度森林的网络流量分类器的分类准确率能够达到96.36%,性能优于其他机器学习模型。

关键词:特征选取;多粒度级联森林;机器学习;网络流量分类
中图分类号:TN95 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2020)04-030-05

听语音 聊科研 与作者互动

Network traffic classification method based on deep forest

DAI Jin[1,2], *WANG Tianyu*[2,3], *WANG Shaowei*[2]
(1. School of Information Science and Engineering, Jinling College, Nanjing University, Nanjing 210089, China;
2. School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China;
3. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China)

Abstract: With the rapid development of network applications, the Internet traffic classification has a profound impact on the research fields of network resource allocation, traffic scheduling and network security. The traditional flow analysis method based on machine learning has strict requirements for the feature selection and distribution of network flows, which makes it difficult to accurately and stably classify the complex and changeable flow data in practical application. In order to solve the adverse impact of the complexity of sample features on the traffic classification, a new classification method based on deep forest, which utilizes the cascade forest of decision trees and the multi-grained scanning mechanisms aiming to improve classification performance in the case of limited scale of samples and features, was proposed. The machine learning algorithms including support vector machine, random forest and deep forest were trained and tested by using Moore, which is a flow data set in public domain. The experiment results show that the classification accuracy using deep forest model reaches 96.36%, which outperforms the other machine learning models.

Keywords: feature selection; multi-grained cascade forest; machine learning; network traffic classification

随着 Internet 的发展,各种网络应用不断涌现,导致网络流量的复杂性快速增长[1]。网络流量复杂性的增加,本质上是网络中运行业务的复杂性不断增加而造成的。目前的网络业务主要来源于以下几个方面:一方面,随着生活和无线业务的互联网化,媒体、零售和金融等传统业务通过从线下搬到线上,逐步加入互联网;另一方面,随着人们学习生活需求的变化,不断有像慕课、电子支付和电子导航等新型的网络业务加入互联网;此外,随着5G时代的来临,大量基于增强移动宽带(enhance Mobile BroadBand,eMBB)的3D、超高清视频等大流量移动宽带业务,基于海量机器类通信(massive Machine Type of Communication,mMTC)的智慧城市、环境监测、智能农业等大规模物联网业务,以及基于超高可靠超低时延通信(Ultra-Reliable Low Latency Communication,

* **收稿日期**:2019-12-25
基金项目:国家自然科学基金资助项目(61801208, 61671233, 61931023, U1936202)
作者简介:戴瑾(1973—),女,浙江绍兴人,副教授,硕士,E-mail:030308@jlxy.nju.edu.cn;
王少尉(通信作者),男,教授,博士,博士生导师,E-mail:wangsw@nju.edu.cn

URLLC)的无人驾驶、工业自动化等新兴网络业务将会充斥整个互联网。如何有效地从海量的网络数据中识别出应用类型,如何从流量数据中分析提取有价值的信息,已经成为人们关注的重要技术领域。因此,网络流量分类作为增强网络可控性的关键技术,在网络资源分配、流量调度和网络安全等诸多研究领域受到广泛关注[2]。

网络实际应用中流量数据具有结构复杂、数量庞大和属性随着网络状态动态变化的特性[3],机器学习算法在解决诸如此类规模大、复杂性高的网络流量分类问题中表现出先天的优势。目前研究使用的基于机器学习的流量分类方法主要包括支持向量机(Support Vector Machine,SVM)、卷积神经网络(Convolutional Neural Network,CNN)和随机森林(Random Forest,RF)算法。在文献[4]中,作者采用了基于流的分析方法,使用 SVM 分类算法对动态端口和加密应用程序按流量进行分类,流量分类算法的精度为 88.785%。文献[5]将流量分别处理为时间序列、图片和视频,模型训练前将网络流预处理成为灰度图,把 CNN 对图像识别的能力应用到网络流量分类中,其分类精度可达到 93%。文献[6]提出将 RF 方法应用到网络流量分类问题中,随机森林通过综合多个决策树的预测和样本属性的随机选取,能较好地解决网络流量样本种类数量失衡而产生的过拟合问题。

目前的网络流量分类方法依然存在准确率不高、开销大且应用领域受限等问题。其中单一机器学习分类方法在复杂多变的网络环境下,分类性能往往会迅速下降。例如:SVM 分类器更利于对线性属性的数据样本的处理,处理离散的、大规模的数据分类处理时,分类准确率低;CNN 分类器擅长解决图像样本的分类问题,网络数据流需要经过烦琐的预处理过程生成图像,当网络层数过多时,会出现计算复杂度高、运行速度慢等问题。在文献[7]中比较了 121 个数据集上的数百个分类,结果表明基于 RF 分类器解决网络流量分类问题是最适合的方法,分类精度可达到 94.1%,但 RF 在处理多分类、样本规模小、噪音较大的流量样本时,依然容易出现过拟合的现象。

针对上述问题,本文提出基于多粒度级联森林(multi-grained cascade Forest, gcForest)算法来构建网络流量分类器。gcForest 是一种决策树的深度监督网络算法,该算法采用多粒度扫描样本和级联多种分类器的设计思想,使分类模型不受样本数量规模和特征向量规模的限制[8]。gcForest 在解决商品分类、手部运动识别、情感分类以及垃圾邮件发送者检测等多分类问题中均表现出了优越的性能[9]。

1　数据集

研究中利用了经典的 Moore 网络流量集训练和测试分类模型,它是 Moore 等采集并经过统计处理的剑桥大学的流量数据。Moore 数据集曾被用于众多网络流量识别和分类等领域[10]。

1.1　定义

Moore_set 网络流数据集是通过采集主干网上的数据,经过基于流的分析方法处理获得[11]。数据集中的样本数据包含了大量在会话规模尺度上统计处理网络层和传输层的消息生成的特征属性,包括流量持续时间的分布、流量空闲时间、数据包到达时间、数据包长度等。数据集中的每个样本均由一组统计信息和一个定义网络应用程序的类别来描述。

数据集中每一个样本都是以特征属性描述的一条完整的传输控制协议(Transmission Control Protocol,TCP)流。在 Internet 中,TCP 流可以定义为在两个计算机地址之间传播的一个或多个数据包。TCP 作为一种有状态的协议,对 TCP 流定义了明确的开始标志和结束标志。因此,研究者能够有效地从纷乱的网络数据包中提取出完整的 TCP 流,也称之为会话。进一步对 TCP 流中每个数据包提供的信息进行分类统计,统计结果就是标识流样本的特征属性。本研究中选用的 Moore_set 数据集,就是由剑桥大学的高性能网络监测器在不丢失数据包的情况下,全速率捕获和统计处理数据包而生成的完整 TCP 流的样本数据集合。该数据集中包含的 11 种不同应用类型的分布情况如表 1 所示。其中由于 INTERACTIVE 类型的样本在数据集中的数量太少,仅有 3 条流样本,无法正常参与训练,因此本文所使用的模型训练和测试的数据子集不包含此分类,实验中训练的是 10 种网络流量的分类模型。

在本问题中,流集合被定义为 $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n, \cdots, \boldsymbol{d}_N,\}$,其中每个样本 $\boldsymbol{d}_n$ 包含 248 个特征属性,N 为基于流的数据样本的数目,N = 24 863。定义类别属性集合 $\boldsymbol{C} = \{c_1, c_2, \cdots, c_m, \cdots, c_M\}$,其中 M = 11 表示有 11 个不同的类别。流量分类问题可以定义为构造从流集合 $\boldsymbol{D}$

到类别集合 $\boldsymbol{C}$ 的映射关系：$c_m = f(\boldsymbol{d}_n)$。

表 1 TCP 流数据集的应用类型分布

Tab. 1 Application class distribution of TCP flow data set

分类	流数	占比/%	应用说明
WWW	18 211	73.25	web
MAIL	4146	16.68	SMTP、POP、IMAP
FTP-CONTROL	149	0.60	FTP
FTP-PASV	43	0.17	FTP
ATTACK	122	0.49	Port scans、viruses
P2P	339	1.36	GnuTella、Napster、kazaa
DATABASE	238	0.96	MySQL、Oracle
FTP-DATA	1319	5.30	FTP
SERVICES	206	0.83	DNS、LDAP、NTP
INTERACTIVE	3	0.01	SSH、TELNET
MULTIMEDIA	87	0.35	MediaPlayer、Real、Itunes
合计	24 863	100	

1.2 特征选取

在监督分类的机器学习中，无关和冗余的特征属性往往会给分类的准确性带来负面影响。通过选择出能高度区分不同流类别的特征向量来降低样本维度，对于模型训练的速度以及分类的准确性是至关重要的。因此，本文采用人工选取加算法测试的方法从 Moore_set 的 248 个特征属性中选出 13 个特征属性用于分类模型的训练和测试，样本的特征属性及描述如表 2 所示。

表 2 选取的流特征属性及其描述

Tab. 2 Selected flow features and descriptions

序号	特征属性	描述
1	server_port	服务器端口
2	client_port	客户机端口
3	actual_data_pkts_clientTOserver	至少有一个字节有效负载 TCP 包的数目（客户机→服务器）
4	pushed_data_pkts_clientTOserver	TCP 头部 PUSH 位为 1 的数据包的数目（客户端→服务器）
5	pushed_data_pkts_serverTOclient	TCP 头部 PUSH 位为 1 的数据包的数目（服务器→客户端）
6	min_segm_size_clientTOserver	连接生存期内最小段的大小（客户机→服务器）
7	avg_segm_size_serverTOclient	连接生存期内平均段的大小（客户机→服务器）
8	initial_window_bytes_clientTOserver	初始窗口中发送的字节数（客户机→服务器）
9	initial_window_bytes_serverTOclient	初始窗口中发送的字节数（服务器→客户机）
10	RTT_samples_clientTOserver	有效往返时间大小（客户机→服务器）
11	med_data_ip_clientTOserver	IP 包中总字节的中等位数
12	mean_IAT_clientTOserver	分组到达时间的平均值（客户机→服务器）
13	duration	连接持续时间

2 模型描述

本文提出的网络流量分类器是基于 gcForest 深度机器学习算法，其核心思想是优化的决策森林。算法所采用的多粒度扫描和级联森林两种关键技术在流量分类中发挥了重要的作用。gcForest 分类器组织结构如图 1 所示。

2.1 级联森林

在 gcForest 算法中，森林是由决策树构成的决策树森林，即森林中的每棵树都是由决策树的节点分裂生长而成，每个决策森林中又包含了许多棵决策树。其中决策树的种类和棵数是算法的超参数，可根据解决问题的规模进行适当的调整。在解决本文中的流量分类问题中，可设置为 4 个决策森林：2 个随机森林和 2 个完全随机森林。

级联森林拥有多个网络层，每个森林将产生一个类向量，分别代表检测的网络应用的分布概率，如图 1 所示。在这个网络流量分类问题中，每个森林会产生一个 10 维的类向量，每层输出的结果和原始输入属性相连接构成了级联森林下一层的输入。训练过程中，层级将会不断深入，直到某层触发终止条件，停止层级的增长。终止条件为达到所需精度或达到最大层数。在最后一层上，gcForest 算法计算求得四类向量的平均值，并从中选出概率最大的类作为最终的分类结果。由此可见，这种级联机制可以实现样本特征向量的跨层级处理，且层数根据计算规模自动调整，有效减少了算法的复杂度。

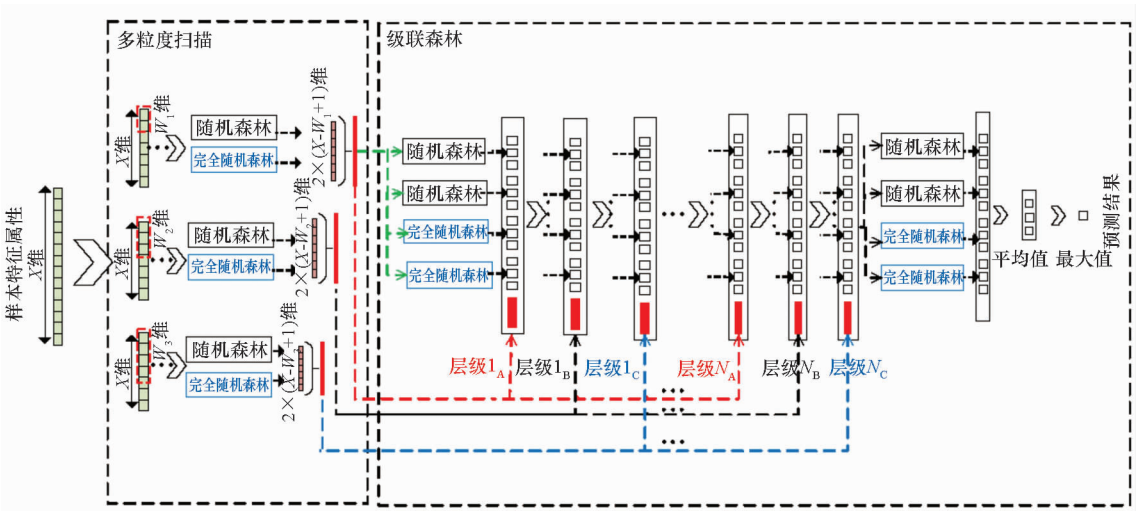


图 1　gcForest 结构

Fig. 1　Constructure of gcForest

2.2　多粒度扫描

在深度学习模型的启发下,gcForest 还利用多粒度扫描技术对级联森林进行增强。多粒度扫描是在模型中引入了类似卷积神经网络的多个滑动窗口。如图 1 所示,在一个具有多分类能力的深度森林分类器中,输入一个完整的 X 维样本,首先通过长度为 W 的滑动窗口进行采样,得到 $X-W+1$ 个 W 维特征向量的子样本集合;随后子样本集合分别送入完全随机森林和普通随机森林进行训练,每个子样本从每个森林中分别获得一个长度为分类个数的概率向量;然后把每层所得所有森林的结果拼接在一起,得到该层输出。gcForest 通过多粒度扫描,实现了对同一个样本不同规模的局部特征的选取,有效地增强了样本的维度和样本特征属性之间的关联。

3　实验结果和分析

3.1　gcForest 分类模型

在本实验中,采用了人工选取特征结合模型测试的数据预处理方法。根据常用 TCP 流量统计特征预选出 17 个特征属性,逐一剔除特征属性构建训练数据集,输入 gcForest 算法训练模型并测试分类准确率。当某项特征属性去除后,准确率没有变化或有微弱的上升,表明它属于冗余、相关性极弱的特征,可以剔除。数据集经过预处理后,剔除了 4 个特征属性——pure_acks_sent_serverTOclient、pure_acks_sent_clientTOserver、mss_requested_serverTOclient 和 mss_requested_clientTOserver,最终确定在样本中引入了 server_port、client_port、actual_data_pkts_clientTOserver、duration 等 13 种特征属性,详细内容见表 2。

实验训练数据利用 Moore_set 的 24 863 个样本,从中选出 25% 的样本共 6226 条网络流用于 gcForest 算法分类模型训练和测试数据子集,其中 4980 条网络流用于模型训练,占数据子集的 80%;剩余的 20%(包含样本 1246 条网络流)作为测试集,用于模型的精度测试。样本集中包含了 WWW、MAIL、DATABASE、FTP-DATA 等 10 种网络应用类别。

使用多分类查全率和查准率对分类器的性能进行评估。第 i 分类的查全率定义为 $R_i=\frac{TP_i}{TP_i+FN_i}$,第 i 分类的查准率定义为 $P_i=\frac{TP_i}{TP_i+FP_i}$。其中,$TP_i$ 是第 i 分类的样本中分类模型预测正确的样本数量;FN_i 是第 i 分类样本中被错误预测为其他分类的样本数量;FP_i 是非 i 分类的样本中被错误预测为类型 i 的样本数量。如表 3 所示,gcForest 网络流量分类器在 WWW、MAIL、DATABASE、FTP-CONTROL、FTP-DATA、SERVICES 分类上查全率和查准率均高于 95%,分类的准确性高。ATTACK 分类上查全率和查准率均较低于 0.1,原因在于 ATTACK 分类的流样本数量较少,且和 WWW 分类统计流的特征属性相似度极高,ATTACK 分类的大多数样本被误判为 WWW 分类。

表3 各类样本的查全率和查准率

Tab. 3 Recall and precision of all kinds of samples

分类	查全率	查准率
WWW	1.000 0	0.958 5
MAIL	0.998 2	0.996 4
DATABASE	0.967 2	0.983 3
FTP-CONTROL	1.000 0	1.000 0
FTP-PASV	0.916 7	0.733 3
ATTACK	0.083 3	0.062 5
P2P	0.240 7	1.000 0
FTP-DATA	0.993 0	1.000 0
SERVICES	0.951 6	1.000 0
MULTIMEDIA	0.692 3	0.720 0

3.2 模型性能

在相同测试集上,将 gcForest 分类器同传统的 SVM 和 RF 分类器在分类正确率和测试时间两个方面进行了比较,如表4所示。实验设备采用主频 1.6 GHz 的 Intel Core i5 – 8250U 处理器,内存容量 4.0 GB 的单机。实验数据显示,基于 RBF 核函数的 SVM 的准确率为 49.14%;基于线性核函数的 SVM 的准确率为 88.17%;RF 的准确率是 96.15%;gcForest 准确率最高,为 96.36%。然而,gcForest 分类模型在线预测 1000 条流量数据的时间为 0.902 s,计算速度低于基于线性核函数的 SVM 分类器和 RF 分类器,但略高于基于 RBF 核函数的 SVM 分类器。

实验结果表明,尽管在测试时间的效率上 gcForest 分类器没有表现出绝对的优势,但就模型分类准确率而言,gcForest 流量分类器远高于传统的 SVM 单分类算法,甚至比目前公认最适宜解决流量分类问题的 RF 算法的性能更好。

表4 不同分类模型的性能比较

Tab. 4 Performance comparison of different classification models

模型	准确率/%	每1000条流量数据的计算时间/s
基于 RBF 核函数的 SVM	49.14	1.000
基于线性核函数的 SVM	88.17	0.023
RF	96.15	0.182
gcForest	96.36	0.902

4 结论

对网络流量进行准确的识别是网络资源分配和网络安全保障的重要依据,也是提高网络应用服务质量的重要手段。为了进一步提高网络流量识别的正确性和稳定性,提出基于 gcForest 的网络流量分类方法,并在 Moore 数据集上进行了训练和测试,最终结果表明该分类器对网络流量分类的整体准确率达到了 96.36%,网络流量分类的性能好于传统 SVM 和 RF 机器学习分类方法。

参考文献(References)

[1] Marsch P, Bulakçı Ö, Queseth O, et al. 5G system design: architectural and functional considerations and long term research[M]. USA: John Wiley & Sons, 2018.

[2] Dhote Y, Agrawal S, Deen A J. A survey on feature selection techniques for Internet traffic classification[C]// Proceedings of International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2015: 1375 – 1380.

[3] Moore A W, Zuev D, Crogan M L. Discriminators for use in flow-based classification [R]. London: Queen Mary University of London, 2005.

[4] Aggarwal R, Singh N. A new hybrid approach for network traffic classification using SVM and Naïve Bayes algorithm[J]. International Journal of Computer Science and Mobile Computing, 2017, 6(6): 168 – 174.

[5] 王勇, 周慧怡, 俸皓, 等. 基于深度卷积神经网络的网络流量分类方法[J]. 通信学报, 2018, 39(1): 14 – 23.
WANG Yong, ZHOU Huiyi, FENG Hao, et al. Network traffic classification method basing on CNN[J]. Journal on Communications, 2018, 39(1): 14 – 23. (in Chinese)

[6] Wang C, Xu T G, Qin X. Network traffic classification with improved random forest [C]// Proceedings of 11th International Conference on Computational Intelligence and Security (CIS). IEEE, 2015: 78 – 81.

[7] Fernández-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. Journal of Machine Learning Research, 2014, 15(1): 3133 – 3181.

[8] Zhou Z H, Feng J. Deep forest [J]. National Science Review, 2018, 6(1): 74 – 86.

[9] Zhou M, Zhang S Z, Qiu Y J, et al. Entropy-based spammer detection [C]// Proceedings of the 10th International Conference on Internet Multimedia Computing and Service. ACM, 2018: 43.

[10] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques [J]. ACM SIGMETRICS Performance Evaluation Review, 2005, 33(1): 50 – 60.

[11] 高文, 钱亚冠, 吴春明, 等. 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 Yaguan, WU Chunming, et al. The divide-conquer and voting strategy for traffic feature selection[J]. Acta Electronica Sinica, 2015, 43(4): 795 – 799. (in Chinese)