文章编号 1672-6634(2021)03-0018-10 **DOI** 10.19728/j.issn1672-6634.2021.03.003

网络应用程序分类的多样化组合特征选择算法

蒋胜利[1],张文祥[2],张军英[2]

(1.洛阳师范学院 电子商务学院,河南 洛阳 471934;2.西安电子科技大学 计算机科学与技术学院,陕西 西安 710071)

摘 要 考虑到网络用户数量的快速增长、日益复杂的网络环境以及网络应用程序多元化的现状,识别网络中的具体应用程序(诸如 Google、Facebook、Skype、MSN 等)是网络应用的重要研究方向,通过提取网络流量特征并利用机器学习方法识别网络应用程序是其主流方法,但由于网络流量特征多且复杂,经特征选择所获特征用于分类的性能往往严重依赖于所选用的分类器而不能很好反映应用程序的个性特征。为此本文提出了一种基于多样化组合特征选择的网络应用程序分类方法,通过组合特征重要性筛选和递归特征消除,获取这两种特征选择方法选择到的特征进行并操作,后用皮尔逊相关系数进一步去除冗余特征。对有 87 个特征的总计 3577296 个实例的网络数据集实验结果表明,与传统的诸如 VT、RFE、L1 正则化逻辑回归等特征选择方法相比,组合特征选择方法在 KNN、SVM、RF、GBDT、XGBoost、LighGBM 等各类分类器上均性能优异(分类准确率提升了 0.5%-3.0%),且性能基本不受所采用的分类器的影响,表明所选出的特征能更客观反映网络应用程序的特性,同时所需运行时间也极大缩减(缩减了 20%-90%),提升了网络应用程序实时监管的效率。

关键词 特征选择;Web 应用程序分类;机器学习

中图分类号 N39

文献标识码 A

开放科学(资源服务)标识码(OSID)

Diversified Combination Feature Selection Algorithm for Web Application Classification

JIANG Shengli[1], ZHANG Wenxiang[2], ZHANG Junying[2]

(1.School of E-commerce, Luoyang Normal University, Luoyang 471934, China; 2. School of Computer Science and Technology, Xidian University, Xian 710071, China)

Abstract The rapid growth of network users, the increasingly complex network environment and the diversification of web applications require identifying specific web applications in the internet (such as Google, Facebook, Skype, MSN, etc.) Extracting web traffic features is the mainstream method for the identification. Different feature subsets obtained by different feature selection methods have a great impact on classification accuracy of web applications, making the selected features heavily dependent to the follow-up selected classifier. This indicates that the selected features are not the personal features of the problem.

收稿日期:2020-11-04

基金项目:国家自然科学基金项目(11674352)资助

通讯作者:张军英,女,汉族,博士,教授,博士生导师,研究方向:模式分析、机器学习、癌症相关计算生物信息学、精准医疗,E-mail:jyzhang@mail.xidian.edu.cn。

This paper proposes a diversified feature combination selection method for web application program classification. Through important feature screening and recursive feature elimination with diverse feature selection methods and diverse types of classifiers as well as tree-based classifiers which are claimed to robust to noises, the features are selected and merged, and redundant features are further removed based on Pearson correlation analysis to obtain the final solution of selected features. Experiments of the proposed method and comparison with some typical counterpart feature selection methods such as VT, RFE and L1 regularized logistic regression on a web application program identification problem show that the proposed method outperforms its counterparts when the six main stream classifiers are applied, KNN, SVM, RF, GBDT, XGBoost, and lighGBM (the classification accuracy is increased by 0.5%-3.0%). This indicates that the selected features by the proposed approach is a better representation of the web application programs, and the running time for the identification is greatly reduced, by 20%-90%, which greatly improves the possibility of real-time monitoring of network applications.

Key words　Feature selection; Web application program classification; Machine learning

0　引言

互联网高速发展，据中国互联网络信息中心(CNNIC)公布的数据，2020 年 1-6 月，移动互联网接入流量消费达 745 亿 GB(第 46 次中国互联网络发展状况统计报告，http://www.gov.cn/xinwen/2020-09-29/5548176/files/1c6b4a2ae06c4ffc8bccb49da353495e.pdf)，各种各样的网络应用程序层出不穷。互联网的应用从简单的文字信息交流和门户网站的信息展示，发展到现在的游戏、视频、旅游等各种不同的网络应用程序。多样的互联网应用极大的丰富了人们的生活，随之而来网络流量的增大，使对网络环境的监控与管理要求越来越高，不仅需要分析网络流量，更需要对网络上所运用的应用程序进行实时分类，是较流量分析的更深层次分析的需要[1]。

本文的选题意义和价值表现在：所提出的方法能够快速高效地从数据中找出多样化网络应用的特征表示，并基于此对网络应用程序进行高质量分类，从而为对大量、多样、高速的网络应用环境的进一步分析、高效监督与有效控制，奠定基础和技术保障。

表 1　部分特征名称及其含义

特征名称	描述
Source.Port	源端口
Destination.Port	目的端口
Source.IP	源端口 IP 地址
Destination.IP	目的端口 IP 地址
Average.Packet.Size	数据包的平均大小
Flow.Bytes.s	每秒流的字节数
Flow.Duration	流的持续时间
Fwd.Packet.Length.Std	正向数据包长度标准差
Bwd.Packets.s	每秒反向数据包数
Bwd.IAT.Total	反向发送两个数据包之间的总时间
Init_Win_bytes_forward	在初始窗口中向前发送的字节总数
ACK.Flag.Count	ACK 标志计数

在对网络应用程序分类与识别这一问题上，当前使用报文解析的方式得到各种包含大量网络流量特征数据集，具体的流量特征及其含义部分示于表 1 中。针对这些网络流量特征，文献[2]提出了基于特征选择的业务流网络应用分类算法，但所选择的特征严重依赖于后续所采用的分类器的类型，从而不能表征和解释网络应用程序的特征；文献[3]提出了一种基于方差系数的特征选择方法，然后通过方差是在高斯分布假设下的

统计量,而实际情况完全有可能违背这个假设。文献[4]针对网络流量的时间特征,提出了基于时间特征的网络应用程序分类算法,但这种只处理时间特征的方法忽略了非时间特征对分类结果的影响。随着深度学习在图像、自然语言处理、推荐算法等领域的成功应用,文献[5]提出了一种将网络流量特征转换为图像的深度学习框架来识别网络应用程序。这是一种具有革命性的创新,充分利用了深度学习在图像处理任务上的优势,但存在对硬件性能的要求较高、结果的可解释性差和高时间成本等劣势,不适于网络分析的实时性要求。

1 研究动机

对于数据科学家而言,大多希望通过模型的微调以获得模型的最佳性能,所以模型的可解释性被认为是实验改进的指导准则和方向,从而尽可能地得到最优的结果。而在工业界中,利用机器学习以及深度学习解决问题需要与用户建立信任关系,无论最终解决方案的目标是什么,终端用户都需要可解释、可关联或可理解的解决方案。网络应用程序分类任务,同样需要模型的可解释性作为网络流量相关业务的数据原理支撑,只有知道具体的每一个特征所代表的含义才能有的放矢的给出相应的解决方案。从表 1 可以看出,网络流量特征名称是清晰且有对应的含义的,但数据量大,所以常使用降低特征维度的方式来缩减数据量,以提高模型的分类效果、可解释性及泛化能力。

常用的降维方法有特征选择法[6]和特征提取法[7]。特征选择和特征提取方法存在同一个目标,即减少数据集中特征的数量,起到数据降维、提升模型学习效果的作用。但它们各自的原理和使用的技术不同:特征提取是通过原始特征的组合与计算获得新的特征,这些新特征不再具有原始特征的含义,从而用这些新特征训练的模型可解释性差,使这一方法在对解释性要求较高的场合并不适用;特征选择则从原始特征中选择对模型训练效果好的特征子集,由于原始特征的含义清晰,用决策树等解释性强的集成模型后,可明确分析出哪些特征对分类起正作用,哪些起负作用,从而仅用起正作用的特征进行分类,这对模型分类正确率和泛化能力的提升,以及对实际业务问题的分析有很大的帮助。

目前大多数网络应用程序分类采用特征选择方法,但是方法单一,效果不足,尤其是很多特征选择方法的性能严重依赖于其后续的分类器的选择,使所选择出的特征不能较好地表征网络应用程序的特性[8,9]。

实际上,特征选择至少有两个不同的目的[10],(1) 以分类目的,即选出最优的特征子集,使用在这个特征子集上的分类器可对样本进行最好的分类;(2) 以表征为目的,即选出的最优特征子集,能够对样本特性进行最好的表征。这两个目的既不同又联系,前者更强调分类性能,而分类性能常常与所选用的分类器有关,后者更强调表征性能,从而在此基础上的分类性能应基本与分类器的选用基本无关。一定程度上,以表征为目的的特征选择应是特征选择的最终目标,因为其对样本特性的优秀表征能力,使在此基础上的分类性能,应具有更为优秀的分类性能。

本文试图给出一种特征选择方法,以期选出的特征较少依赖后续分类器选择,从而能更好地表征网络应用程序的特征。为此,提出了一种多样化组合特征选择的网络应用程序分类算法 FIRFEP(Feature Importance and Recursive Feature Elimination with Pearson Correlation Coefficient)。该算法组合了两种不同的特征选择方法(特征重要性筛选法和递归特征消除法),每种方法在具体方法上都体现了广泛的多样化的特点,并将这些方法选出的特征进行并运算并去除其中的冗余特征,且每种方法都采用了具有较好鲁棒性的算法,达到了最终选择出能更好表征网络应用程序特征的目的。与目前的 3 种主流特征选择算法和 6 种主流分类算法在网络应用程序分类的实验结果比较表明,算法在泛化性能和运行效率上都得到了明显提升,特别地,运行效率还得到了极大的提升,使对网络应用程序的实时监控成为可能。

2 FIRFEP 算法

我们认为,现有各种特征选择方法,要么与分类器无关却用于分类时性能欠佳,要么与分类器有关却存在分类器偏好(即换种分类器的性能会较大幅度下降),难于反映数据中的本质特征。从这个意义上讲,充分利用好各种特征选择方法的优势,通过由它们选出的特征的组合,并去除相关特征,则更能反映数据本身的本质。从这个意义上讲,通过现有特征选择方法所选择出特征的多样性组合,是选出尽可能不存在分类器偏好的数据本质特征的重要途径。图 1 给出了 FIRFEP 算法的基本架构。

本文提出的多样化组合特征选择(FIRFEP)的原则是：第一，所采用的特征选择方法应尽可能不同，从而所选出的特征尽可能地形成优势互补；第二，对于分类器依赖的特征选择方法，其所依赖的分类器应尽可能不同，从而其所选出的特征构成不同的分类器依赖；第三，分类器的选取选用具有良好鲁棒性、泛化能力、易于对分类过程进行解释的树形结构分类器。这三点导致了特征选择的多样性。将这些方法选择出来的特征组合起来(通过并运算)，有可能存在冗余特征，因此还需要去除它们之间具有相关特性的冗余特征，获得最终的特征选择结果。

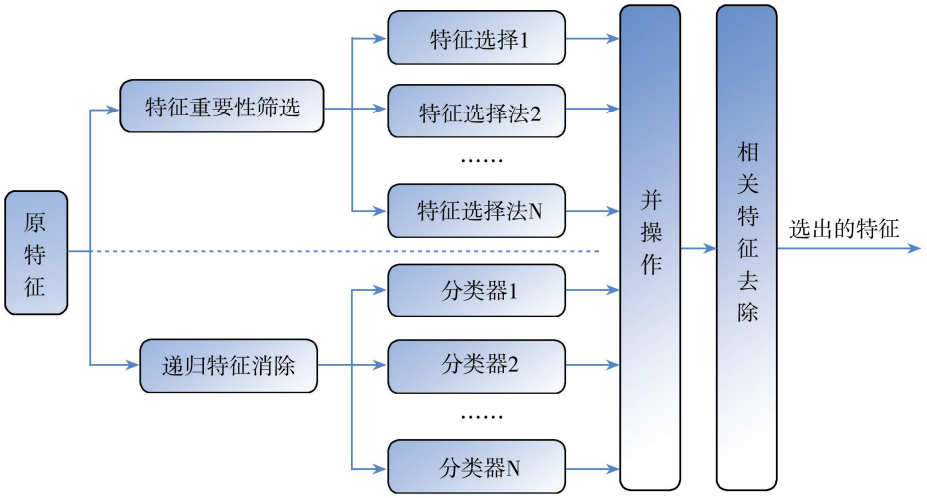


图 1　多样化组合特征选择法基本框架

特征选择方法通常分为过滤式(Filter)、包裹式(Wrapper)和嵌入式(Embedding)[10]，其中每类方法的典型算法分别为方差选择法(Variance Threshold，VT)[11]，递归特征消除法(FFE)[12]、基于逻辑回归带 L1 惩罚项的特征选择法(L1-based)[13]。鉴于以上原则，我们选用上述三种方法作为本文算法的特征选择基础算法，选用具有树形结构的鲁棒分类器 GBDT[14]、XGBoost[15]、LightGBM[16]等作为本文算法的基础分类器。

FIRFEP 算法分为两个阶段，第一阶段是分别使用基于特征重要性筛选法的 N 种特征选择方法和基于递归特征消除法中 M 种分类器方法对数据集中的特征分别进行初步筛选，并对筛选出的特征子集求取并集；第二阶段，采用皮尔逊相关系数法对第一阶段筛选出的特征集合进行特征相关性判断，丢弃具有高线性相关的特征，达到进一步删除冗余特征的目的，得到最终的特征选择集合，且所有的特征选择和分类器的选用都运用了具有良好鲁棒性和泛化能力的树结构模型。多种特征筛选方法的互补、多种分类器的互补，以及各类树模型分类器的良好鲁棒性和泛化能力，是本文提出的 FIRFEP 方法能够提取出良好特征的关键。

算法用基于特征重要性筛选法对单个重要的特征进行筛选，用基于递归特征消除法对联合起来重要的特征进行筛选。其中特征重要性筛选法执行效率高、可解释性强是首选的特征选择方法，但是存在手动指定特征筛选阈值而导致分类算法性能下降、容易丢失联合起来重要的特征的问题；递归特征消除法则较好解决了联合起来重要的特征筛选，但所选出的特征严重依赖于其中所选用的分类器类型。FIRFEP 算法则将这两种方法结合，通过对选出特征的并集运算，使所选出的特征既可以满足高可解释性，又不丢失联合起来重要的特征，同时其性能又不依赖于所采用的分类器，从而能够更加客观地反映被识别目标的特性，并运用皮尔逊相关系数筛选法，通过丢弃特征线性相关性大于设定门限的冗余特征，达到进一步提升算法性能的目的。

RIRFEP 算法的一个重要特点是特征选择的多样化，通过多样化的特征选择方法选出的特征的并集，并去除相关的特征，获取最终的特征选择结果。RIRFEP 算法的基本原理具体如下。

2.1　特征重要性筛选

特征重要性也称为基尼重要性或基尼指数[17]。基尼指数与信息熵有类似特质，均可用来衡量信息的不确定性。依据基尼指数对特征进行筛选的方法在决策树模型建模时常常用到。决策树中的每个节点都是关于某个特征对样本划分的条件，通过计算不纯度来选择最优的决策条件。对于多个决策树来说，计算每个特征不纯度下降程度的总和，把它们的归一化值作为特征的重要性，基尼重要性公式如

$$Gini\ importance = \frac{N_t}{N} * (impurity - \frac{N_{tR}}{N_t} * right\_impurity - \frac{N_{tL}}{N_t} * left\_impurity), \tag{1}$$

其中 N 是样本总数，N_t 是当前节点的样本数，N_{tR}、N_{tL}是当前节点右孩子、左孩子的样本数，$impurity$ 是当前节点的不纯度，即基尼指数，$right\_impurity$、$left\_impurity$ 是当前节点右孩子、左孩子的不纯度。

假设有 K 个类，每个样本都有其类标，对于某一样本，其属于第 k 类的概率 p_k 可用属于这类的样本数占样本总数的比例来计算，则这一特征的基尼指数的公式如

$$G(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2\ , \tag{2}$$

特征的基尼指数越大,说明该特征对表征目标特性的重要性越大。

基于树模型的特征选择方法具有鲁棒性好、可解释性强、易于使用等特点,本文采用了3种较新的基于树模型的特征选择方法进行特征选择(分别是GBDT[14]、XGBoost[15]、LightGBM[16]),依据这些方法的特征重要性对特征进行排序获得其各自的特征选择结果,并将这些结果进行并操作,得到基于特征重要性的特征筛选结果。采用多种特征选择方法的主要考虑是,每种特征选择方法都有各自的偏向,而对这些方法各自选出的特征进行并操作则会使所获得的特征筛选结果具有很好的互补性。

2.2 递归特征消除法

Guyon等人在对癌症分类的基因选择研究中首次提出了递归特征消除法(RFE)[12],该方法通过对产生的特征子集按照分类中特征的重要性进行排序,可以删除冗余和无效的特征,提升分类算法的分类性能。

递归特征消除法是一种寻找最优特征子集的贪心算法,其主要思想是反复的构建分类器,对所有特征的特征重要性进行排序,去掉当前特征集中最不重要的特征或保留当前重要性最高的特征,然后在剩余的特征集上重复这个过程,直到减少特征会造成分类器分类性能损失或增加新的特征分类器但分类性能没有提升为止。这个过程最后留下的特征集合就是最终所选择的特征集合。RFE的具体过程为(1) 训练分类器,优化目标损失函数;(2) 根据特征重要性准则计算特征排名;(3) 移除特征重要性最低的特征或保留特征重要性最高的特征;(4) 重复上述步骤直到分类器性能不再提升或所有特征都已遍历。

文献[18]指出该方法严重依赖分类器本身来选择特征,选取的分类器不同,最终递归特征消除法得到的特征子集也会不同,从而所得到的特征子集不能客观反映各分类对象的特征。

本文采取多种不同的分类器来解决该方法对单一分类器的依赖,提高所选特征的泛化能力。鉴于树模型的分类器具有良好的鲁棒性,在运用RFE进行特征选择过程中,我们分别选用2个具有树结构的分类器(RF分类器[19]和极度随机树Extra Trees[20])进行分类,用基尼重要性对特征进行排序,并对选出的特征进行并操作,获得特征筛选结果。

2.3 皮尔逊相关系数法

通过对2.1和2.2中的特征筛选结果的进一步的并操作,得到一个特征子集。由于在2.1和2.2中均未考虑特征之间的相关性问题,然而特征子集中可能存在冗余特征,为此通过特征之间的相关性分析来进一步获得更加紧凑的特征子集。

皮尔逊相关系数法是一种最简单的、能帮助理解特征与特征之间统计线性关系的统计学方法[21]。该方法衡量的是变量之间的线性相关性,结果处于区间[−1,1],−1表示完全的负相关,+1表示完全的正相关,0表示没有线性相关性,其计算公式为

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{3}$$

其中n是样本数量,x_i,y_i分别代表下标为i的样本,$\bar{x}$,$\bar{y}$是样本的均值。

对进一步并操作后的特征子集,计算其中两两之间的皮尔逊相关系数,一旦其相关系数的绝对值超过一个门限,则仅取其中一个特征,从而最终获得更为紧凑的特征子集,作为本文提出的组合特征选择的最终特征选择结果。

3 实验与结果分析

实验数据为网上公开数据,实验内容包括各种特征选择方法选出特征后运用各类分类器进行分类的泛化能力(分类准确率)和识别过程的实时性(运行时间),目的是验证本文提出的FIRFEP方法所选出的特征,对各类分类器的适应性是否其分类准确率基本不受后续分类器类型选择的影响,从而是更具客观地反映识别对象的特征。

表 2 示出了实验环境，表 3 示出了实验中使用的部分超参数。这里重点报告两个实验的结果，实验 1 对比了本文 FIRFEP 算法与常用的特征选择算法在网络应用程序的识别与分类任务上，不同分类算法的准确率；实验 2 验证了 FIRFEP 算法在运行效率上的提升效果。

表 2　实验环境

类别	参数
操作系统	Windows 10，64 bit
处理器	Intel i5-7500 3.40 GHz
Jupyter Notebook	5.5 版本
Python	3.5 版本

表 3　分类模型的参数设置

类别	参 数
KNN	n_neighbors=10，weights='distance'
SVM	C=10，　decision_function_shape='ovo'
RF	n_estimators=100
GBDT	n_estimators=100，learning_rate=1，subsample=1
XGBoost	n_estimatores=100，max_depth=4，multi='softmax'，num_class=6
LightGBM	n_estimators=20，objective='multiclass'，boosting='goss'

3.1　实验数据

数据来源于哥伦比亚波帕扬考卡大学公开网络数据集[22]，该数据总共收集了六天不同时段的共 3，577，296 个实例，其网络流的统计特征由 CICFlowmeter[23] 获得（IP 地址，端口，到达时间等），网络应用层协议由 ntopng[24]（DPI 深度报文检测）处理所获得，共计特征 87 个。每条数据都保存了由网络设备生成的 IP 流的信息，即源和目标 IP 地址，端口，到达时间，在该流上用作类的第 7 层协议（应用程序）等 87 个属性特征。大多数属性都是数字类型，由于有时间戳，也有字符串类型和日期类型。

本文旨在通过组合特征选择法分类基于网络流特征的具体应用程序，而不是简单地通过分析网络流五元组来识别网络应用程序，且每一个类别的各个特征与数据采集的时间无关，因此丢弃源 IP 地址（Source.IP）、目的 IP 地址（Destination.IP）、时间戳（TimeStamp）等特征。为了验证本算法得到模型的泛化能力和普适性，人工的选取了各个网络应用数据量差别较大、数据本身无缺漏且具有代表性的六类网络应用，分别是：'TWITTER'，'OFFICE_365'，'MS_ONE_DRIVE'，'EBAY'，'TEAMVIEWER'，'TWITCH'。它们涵盖了社交软件、办公软件、云端存储、购物应用、远程控制应用和在线直播等多方面网络应用程序。

3.2　实验设计

特征选择方法通常分为过滤式（Filter）、包裹式（Wrapper）和嵌入式（Embedding），我们选取了这三大类方法中每类方法的典型算法——方差选择法（Variance Threshold，VT），递归特征消除法（RFE）、基于逻辑回归带 L1 惩罚项的特征选择法（L1-based），与我们的方法进行比较，它们各选出的特征并分别用 6 种主流机器学习分类算法：KNN（k-Nearest Neighbor，K 最近邻）、SVM（Support Vector Machine，支持向量机）、RF（Random Forest，随机森林）、GBDT（Gradient Boosting Decision Tree，梯度提升决策树）、XGBoost（eXtreme Gradient Boosting，极端梯度提升），LightGBM（Light Gradient Boosting Machine，轻量级的梯度提升器）进行分类，以考察本文提出的 FIRFEP 方法与其它特征选择方法以及对各类分类器的适应性，即是否分类器的分类性能基本保持较少受到分类器类型的影响。图 2 示出了比较实验的实验框架。其中，方差选择法（VT）[11] 计算各个特征的方差，然后根据人工预定的阈值，选择方差大于阈值的特征。实验 1 中阈值为 0.8；递归特征消除法（FRE）[12] 中使用的是随机森林分类器进行特征消除，所有参数均为默认值；L1-based 特征选择法[13] 用基于逻辑回归带 L1 正则化项进行特征选择，其 L1 正则化法具有稀疏解的特性，天然具备特征选择的特性，实验中正则化系数 C 取为 1。

3.3　评价指标

本文采用机器学习中的通用指标——准确率[25-28] 对实验结果进行评价

$$\text{准确率：}accuracy=\frac{TP+TN}{TP+TN+FP+FN},\tag{4}$$

其中 TP（True Positive）被正确划分为正例的样本数；TN（True Negative）被正确划分为负例的样本数；FP

(False Positive)被错误划分为正例的个数；*FN*(False Negative)被错误划分为负例的个数。

本文用每种分类算法的运行时间来衡量算法处理网络应用程序分类问题的实时性。

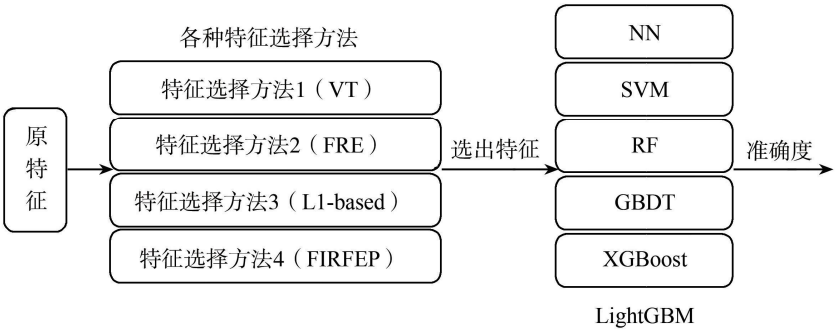
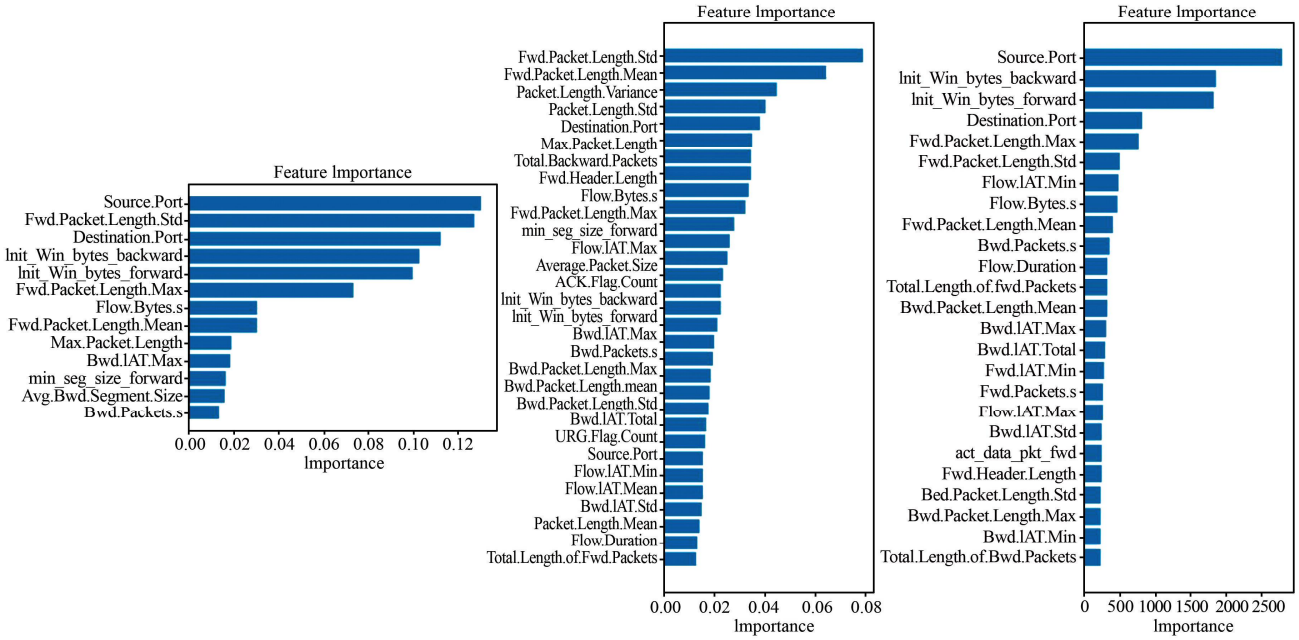


图 2　各种特征选择方法在各种分类器上的识别性能比较实验

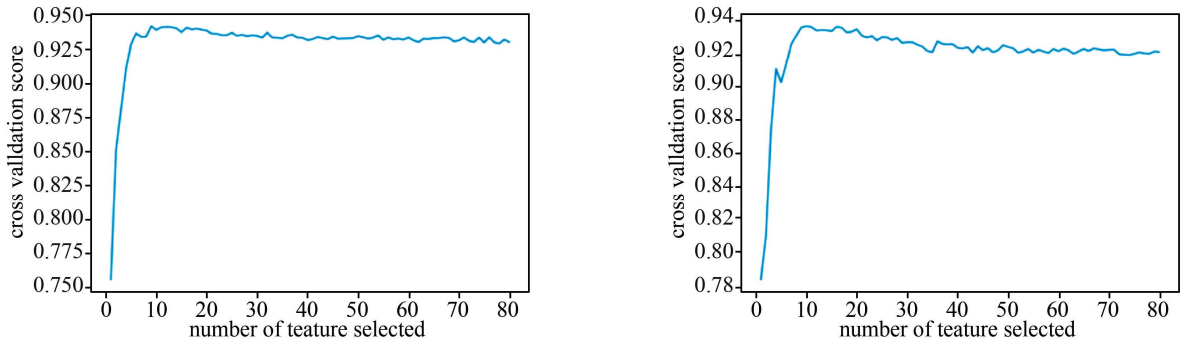
3.4　实验结果与比较

本文所提出的多样化组合特征选择方法 FIRFEP，其在这组数据集上的具体做法如下。对数据集分别运用 GBDT、XGBoost、LightGBM 三种特征选择方法所获得的特征筛选结果如图 3 所示，其中图(a)、(b)、(c)分别为 GBDT 分类器、XGBoost 分类器以及 LightGBM 分类器最终得到的特征重要性排序结果。这些结果中只包含了特征重要性大于平均特征重要性的特征，丢弃了其余特征，最终三种不同的决策树模型分别保留了 15、26、20 个特征。从图中可以看出，这三种不同模型保留的特征集合中部分特征相同，如“Source.Port”、“Flow.Bytes.s”、“Flow.Duration”等，但同时也存在单个特征集合独有的特征，如通过 GBDT 模型筛选出的“Subflow.Fwd.Bytes”特征、XGBoost 模型筛选出的“ACK.Flag.Count”特征以及 LightGBM 模型筛选出的“Bwd.IAT.Total”特征。为提高算法筛选特征的容错率，本文采用将三组特征子集的并集作为特征重要性筛选法的结果，获得特征子集 A，这一过程获得的 A 中共有 28 个特征。



(a) GBDT　　(b) XGBoost　　(c) LightBGM

图 3　三种特征选择方法的特征重要性筛选结果及相应特征的特征重要性



(a) RF 递归特征消除过程　　(b) Extra Trees 递归特征消除过程

图 4　递归特征消除过程示意图

对数据集运用 RFE 方法，其中的分类器分别选为 RF 分类器和 Extra Trees 所得到的特征集合，其识别准确率示于图 4 中，其中(a)为 RF 分类器的结果，(b)为 Extra Trees 分类器的结果。由图可知，RF 分类器在保留 9 个特征后，分类器的分类性能就不再上升，Extra Trees 分类器在保留了 10 个特征后分类器的分类

性能就不再上升了。我们对这 9 个特征和这 10 个特征进行并操作，获得特征子集 B。

将特征子集 A 与 B 进行并操作，并进行相关性分析，对于任两个特征变量，若其相关系数的绝对值超过设定门限（实验中门限设为 1），意味着这两个特征在数学上存在严重的正或者负线性相关，那么这两个特征对应的实际含义从某种程度上来说是完全一样的，只保留其中一个特征，最终获得的特征子集记为 C，其中共保留了 33 个特征。图 5 是经过相关性分析后的部分特征相关性热力图。

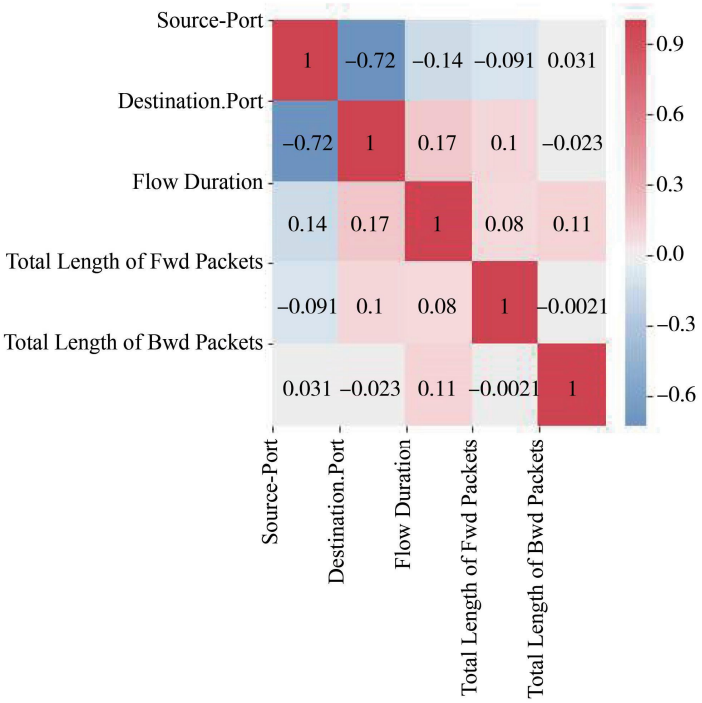


图 5　部分特征相关系数热力图

我们将所选出的 33 个特征与用其他典型的 VT，RFE（其中的分类器选用 RF）和 L1-based 等特征选择方法所选出的分别为 32、10、22 个特征在 KNN、SVM、RF、GBDT、XGBoost、LightGBM 等 6 种主流分类算法上进行了网络应用程序分类的实验和性能比较，分类准确率的实验结果示于图 6 中。由图可知，使用不同特征选择方法及不同的机器学习分类算法最终得到的分类准确率都不相同，其中 KNN 模型、RF 模型、XGBoost 模型使用 VT 特征选择法最终分类效果最好；SVM 模型、GBDT 模型、LightGBM 模型使用 L1-base 特征选择方法最终分类效果最好；而 RFE 特征选择法在所有分类模型上最终分类结果表现都不是最佳的，这一实验结果验证运用特征递归消除法进行特征选择的结果高度依赖其中所使用到的分类器这一缺点，所以 FIRFEP 算法采取多种不同的分类器来解决递归特征消除法对单一分类器的依赖是很有必要的。使用本文提出的 FIRFEP 算法对数据集特征进行特征选择，在上述六种分类模型上均具是最好的表现，分类准确率上均有 0.5%-3.0%的提升。

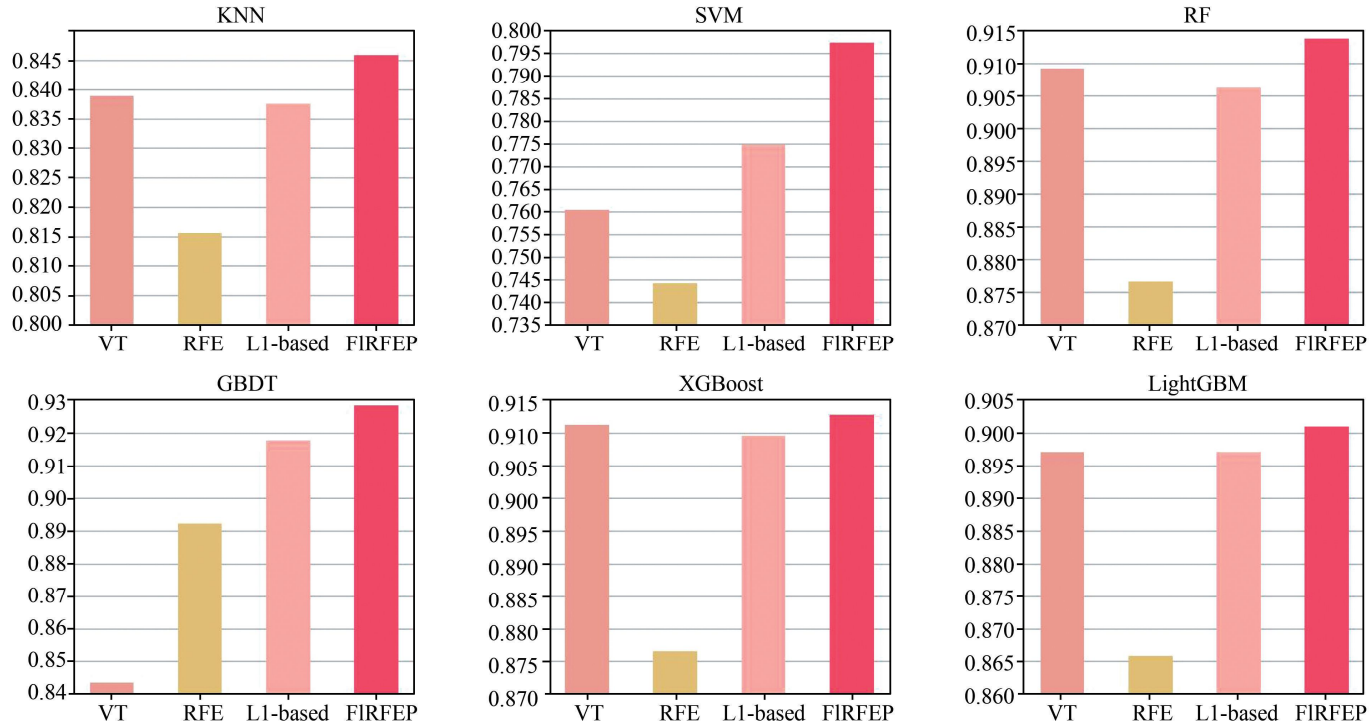


图 6　FIRFEP 算法与常用特征选择方法分类准确率对比图

图 7 给出了不同的特征选择方法选择出特征用各个分类器进行分类的分类性能比较，其中 Accuracy 是分类的平均准确率，Variance 是分类准确率的方差，从而图 7 反映了所选择出的特征对后续分类器的依赖情况。从图中可以看出，本文提出的 FIRFEP 特征选择方法，具有最高的分类平均准确率和最小的分类准确率方差，表明了使用所提出的 FIRFEP 方法，相对于其它特征选择方法，选出了最独立于后续分类器的特征。

表4示出各个分类器在运用本文提出的特征选择方法后的模型运行时间对比情况,可以看出,采用FIRFEP算法进行特征选择后再进行分类任务和不采用该方法直接进行分类任务相比,各机器学习分类算法进行分类的平均执行时间提升量达到了50%以上,最大执行时间提升量达到了近90%,极大的缩减了各分类模型在完成网络应用程序分类任务上的执行时间,提高了算法的实时性,能够更好适应对快速变化的网络环境进行实时管理的要求。

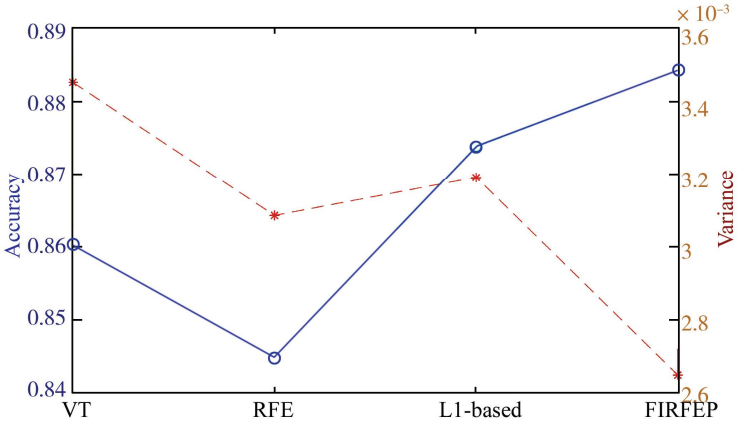


图7 几种特征选择方法选择出的特征的分类性能比较

上述实验结果表明,本文提出的FIFREP算法,保留了与网络应用程序识别相关的重要特征,提升了分类模型在网络流量应用程序识别的识别性能,极大缩减了网络应用程序识别的实时性,有望在更大规模网络应用程序分类数据集上取得更好的效果。

表4 模型运行时间对比

模型	不采用FIRFEP算法/ms	采用FIRFEP算法/ms	运行时间缩减量/%
KNN	242.33	168.87	30.31
SVM	565.08	180.81	68.00
RF	204.11	192.12	58.74
XGDT	529.19	55.83	89.45
XGBoost	877.12	692.03	21.10
LightGBM	365.35	37.23	89.81

4 结语

本文提出了一种组合特征选择法FIRFEP算法来实现网络应用程序的分类,其最大特点是:特征选择方法的多样性和分类器的多样性和鲁棒性。其中特征选择方法的多样性表现在:在基于特征重要性筛选时采用多种特征选择方法,在基于递归特征消除法进行特征选择时采用多种分类器;分类器的多样性表现在:所有分类器均选用 优良鲁棒性和泛化能力的树模型分类器(诸如GBDT、XGBoost、LightGBM、RF、Extra Trees等),并对所选择的特征进行并操作和相关性分析,保证了所选出的特征具有很好的互补性,既没有单一特征选择方法所导致的特征偏好,也没有单一分类器所导致的分类器偏好,从而能更加客观地反映网络应用程序的特征。

通过真实数据实验和测试,这种多样化组合特征选择法在网络应用程序识别这一任务中,具有良好的泛化能力、识别准确率及实时识别的性能,适用于用户量快速增长、环境日益复杂以及应用程序越发多元情况下对网络监控与管理的迫切需求,具有一定的实用性和有效性。进一步的研究包括开发新的特征选择方法以及特征中去除线性无关但是非线性相关的特征。

特征选择是机器学习的重要研究领域,如何从众多的特征中选出反映问题本质的特征,而较少或不受所采用技术的影响,多样化组合是重要解决方案,不仅对于网络应用程序分类问题,对于其它应用问题的特征选择,也有很好借鉴意义。

参 考 文 献

[1] ZHANG B Y,BIAN Y L,ZHANG H K,et al.Linear discriminant analysis in network traffic modeling[J].Journal of China Institute of Communications,2010,19(1):53-65.

[2] DONG Y N,ZHAO J J,JIONG J.Novel feature selection and classification of Internet video traffic based on a hierarchical scheme[J].

Computer Networks，2017，119：102-111.

[3] DONG Y N，YUE Q T，FENG M.An efficient feature selection method for network video traffic classification[C].// 17th IEEE International Conference on Communication Technology (2017 ICCT)，2017.

[4] LASHKARI A H，DRAPER G G，MAMUN M S I，et al.Characterization of tor traffic using time based features[C].// 2017 International conference on Information Systems Security and Privacy (2017 ICISSP)，2017.

[5] WANG W，ZHU M，ZENG X W，et al.Malware traffic classification using convolutional neural network for representation learning[C].// 2017 IEEE International Conference on Information Networking (2017 ICOIN)，2017.

[6] GUYON I，ELISSEEFF A.An Introduction to Variable and Feature Selection[J].Journal of Machine Learning Research，2003，3(6)：1157-1182.

[7] GUYON I M，GUNN S R，NIKRAVESH M，et al.Feature extraction foundations and applications[J].Studies in Fuzziness & Soft Computing，2006，205(12)：68-84.

[8] LIU B，TU H .P2P traffic classification using semi-supervised learning[C].// International Conference on Artificial Intelligence & Computational Intelligence，IEEE，2010.

[9] KISNER T，ESSOH A，KADERALI F.Statistical texture analysis methods for network traffic classification[C].//2007 Asian Conference on Communication Systems & Networks，ACTA Press，2007.

[10] JORDAN M I，MITCHELL T M.Machine learning：Trends perspectives and prospects[J].Science，349(6245)：255-260.

[11] ROBERTS A G K，CATCHPOOLE D R，KENNEDY P J.Variance-based feature selection for classification of cancer subtypes using gene expression data[C].// 2018 International Joint Conference on Neural Networks (IJCNN)，2018.

[12] GUYON I，WESTON J，BARNHILL S，et al.Gene selection for cancer classification using support vector machines[J].Machine learning，2002，46(1/3)：389-422.

[13] CHEN S B，ZHANG Y M，DING C H Q，et al.Extended adaptive Lasso for multi-class and multi-label feature selection[J].Knowledge-Based Systems，2019，173(6)：28-36.

[14] FRIEDMAN J H.Greedy Function Approximation：A Gradient Boosting Machine[J].Annals of Statistics，2001，29(5)：1189-1232.

[15] CHEN T Q，CARLOS G.Xgboost：A scalable tree boosting system[J].Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2016，22：785-794.

[16] KE G L，MENG Q，THOMAS F.LightGBM：a highly efficient gradient boosting decision tree[C].// Advances in Neural Information Processing Systems 30(NIP)，2017.

[17] GENUER R，POGGI J M，TULEAU M C.Variable selection using random forests[J].Pattern Recognition Letters，2010，31(14)：2225-2236.

[18] CHEN X W，JEONG J C.Enhanced recursive feature elimination[C].// 2017 International Conference on Machine Learning and Applications (2017 ICMLA)，2017.

[19] BREIREIMAN L.Random forest[J].Machine Learning，2001，45：5-32.

[20] GEURTS P，ERNST D，WEHENKE L.Extremely randomized trees[J].Machine Learning，2006，63(1)：3-42.

[21] RANGKUTI F R S，FAUZI M A，SARI Y A，et al.Sentiment analysis on movie reviews using ensemble features and pearson correlation based feature selection[C].//2018 International Conference on Sustainable Information Engineering and Technology (SIET)，2019.

[22] ROJAS J S.IP network traffic flows labeled with 75 apps labeled IP flows with their application protocol [EB/OL].(2018-09-16) http：// www.kaggle.com[2019-09-16].

[23] LASHKARI A H.CICF low Meter [EB/OL].(2018-03-07)http：// github.com [2019-11-05].

[24]ALFREDO C，LUCA D，SIMON M.NTOPNG [EB/OL].(2015-04-20) http：// github.com[2019-11-05].

[25] IOANNIS T，GIORGOS B，PAVLOS K，et al.A greedy feature selection algorithm for Big Data of high dimensionality [J].Machine Learning，2018，108：149-202.

[26] NICOLAS G P，CASTILLO J A R D，GONZALO C G.SI (FS)2：fast simultaneous instance and feature selection for datasets with many features[J].Pattern Recognition，2020，111-107723.

[27] YU N，WU M J，LIU J X，et al.Correntropy-based hypergraph regularized NMF for clustering and feature selection on multi-cancer integrated data[J].IEEE Transactions on Cybernetics，2020，99：1-12.

[28] ANGULO A P，SHIN K.Mrmr+ and cfs+ feature selection algorithms for high-dimensional data [J].Applied Intelligence，2019，49(5)：1954-1967.