·信号与信息处理·

基于拉曼光谱技术的可燃液体快速检测研究

刘倩莹[1],贺鹏飞[1],冯巍巍[2,3,4*],杨思节[2,3],蔡宗岐[2,3],王焕卿[2,3]

(1.烟台大学,山东 烟台 264005;2.中国科学院海岸带环境过程与生态修复重点实验室(烟台海岸带研究所),山东 烟台 264003;
3.中国科学院海洋大科学研究中心,山东 青岛 266071;4.中国科学院大学,北京 100049)

摘 要:拉曼光谱作为一种分子"指纹"图谱,能够根据物质分子间的振动对物质进行定性分析,广泛应用在很多领域。拉曼光谱可应用于便携式监测系统,但其数据量偏大,如果不对其数据处理,会增加后续的分析时间,影响自动识别的速度。文中选取九种常见可燃液体90#汽油、93#汽油、97#汽油、丙酮、二甲苯、甲醇、乙醇、乙二醇、叔丁醇为例,对其进行数据预处理分析,特征峰提取效果显著,进而对数据进行512点的压缩,然后选取支持向量机(support vector machine,SVM)分类算法和随机森林分类算法进行模型训练。研究结果表明,随机森林算法的识别可燃液体样品的交叉验证精度高于SVM算法,随机森林算法的均方误差的结果也都优于SVM算法。运用拉曼光谱技术可明显检测出可燃液体样品的谱峰,对数据进行压缩,提高分析速度,可为后续仪器小型化提供技术参考。

关键词:拉曼光谱;可燃液体;随机森林;支持向量机;谱峰

中图分类号:TP317.4 **文献标识码**:A **文章编号**:1673-1255(2021)-01-0058-06

Research on Rapid Detection of Flammable Liquid Based on Raman Spectroscopy Technology

LIU Qian-ying[1], HE Peng-fei[1], FENG Wei-wei[2,3,4*], YANG Si-jie[2,3], CAI Zong-qi[2,3], WANG Huan-qing[2,3]

(1. Yantai University, Yantai 264005, China; 2. Key Laboratory of Coastal Environmental Processes and Ecological Remediation, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China;
3. Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao 266701, China;
4. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: As a kind of molecular fingerprint spectrum, Raman spectroscopy can perform qualitative analysis of substances based on the vibration between the molecules of substances, and is widely used in many fields. Raman spectroscopy can be applied to portable monitoring systems, but the amount of data is too large. If the data is not processed, it will greatly increase the subsequent analysis time and affect the speed of automatic identification. Nine common flammable liquids such as 90#, 93# and 97# gasoline, acetone, xylene, methanol, ethanol, ethylene glycol and tert-butanol are selected, data preprocessing analysis is performed on them, and the characteristic peak extraction effect is significantly, the data is compressed by 512 points, and then the support vector machine (SVM) and the random forest (RF) classification algorithms are selected for model training. Research results show that the cross-validation accuracy of the RF algorithm for identifying combustible liquid samples is higher than that of the SVM algorithm, and the mean square error results of the RF algorithm are also better than that of the SVM al-

收稿日期:2020-10-29

基金项目:中国科学院科技服务网络计划(STS计划)项目(KFJ-STS-ZDTP-077);山东省重点研发计划(2019JZZY010810);烟台市重点研发计划项目(2018ZDCX007)

作者简介:刘倩莹(1996-)女,山东烟台人,烟台大学硕士研究生,主要研究方向为信号与信息处理. E-mail: liuqianying960418@126.com;通信作者:冯巍巍(1980-),男,山东烟台人,博士,中国科学院烟台海岸带研究所研究员,主要从事海洋环境监测技术与仪器研发方面的研究. E-mail: wwfeng@yic.ac.cn

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

gorithm. The use of Raman spectroscopy technology can clearly detect the peaks of combustible liquid samples, and compress the data to increase the analysis speed, which can provide technical references for subsequent instrument miniaturization.

Key words: Raman spectroscopy; combustible liquid; random forest (RF); support vector machines (SVM); spectral peak

拉曼光谱是由印度科学家C V Raman和前苏联科学家曼杰斯塔姆分别于1928年发现的一种散射光谱。拉曼光谱提供了物质分子的简正振动频率及有关振动能级的信息,这些信息只与被测物质的分子振动和转动能级有关,因此拉曼光谱具有物质的指纹特征[1-2]。

随着人们对海洋资源的开发利用要求日渐提高,对海洋环境中的油品等资源勘测识别尤为重要,目前我国油品等可燃液体样品检测采用的光谱学技术主要包括拉曼光谱和近红外、中红外光谱。其中拉曼光谱能够反映可燃液体样品中所含苯、饱和烃、烯烃等有机大分子组分在拉曼光照射下的分子变化特征及结构信息,并具备样品前处理和光谱图采集等方面的优势。

实验利用拉曼光谱对样品进行无损分析,具有测试样品非接触性、非破坏性、检测灵敏度高、时间短、样品所需量小及样品无需制备的特点。随着测量仪器的精度越来越高,每次测量都会产生大量数据,这么多的数据会对后续的分析工作带来很大的麻烦[5-6]。而且,如果所有数据都被传输,将给系统带来很大的负担。因此有必要对数据进行压缩处理,选择有代表性的数据进行分析、传输和存储数据,这将带来极大的方便。

1　材料与方法

1.1　材料与仪器

拉曼光谱仪采用实验室自主搭建的便携式拉曼光谱仪(激光波长785 nm,激发功率400 mW,积分时间为500 ms),实验室用透明玻璃瓶。

可燃液体样品除汽油外均为实验室选取纯品,共九种,均为公共领域常见易燃品,包括:90#汽油、93#汽油、97#汽油、丙酮、二甲苯、甲醇、乙醇、乙二醇、叔丁醇。

如图1所示,采用的基于拉曼光谱的可燃液体的快速测量装置示意图。包括探测单元和信号处理传输单元。其中,探测单元包括光源控制模块、激光光源、拉曼探头以及载物装置模块,信号处理传输单元包括光谱采集模块和计算机,载物装置模块用于承载待测样品,光源控制模块输出端与激光光源相连,激光光源发出的激光经入射光纤通过拉曼探头传输至待测样品,待测样品产生的拉曼散射光再经拉曼探头进入信号处理传输单元的光谱采集模块,光谱采集模块将采集到的光谱信号传输至计算机,计算机再对拉曼光谱数据进行数据处理。

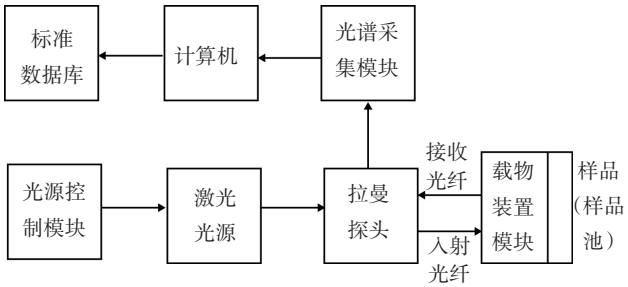

图1　基于拉曼光谱的可燃液体快速测量装置结构示意图

1.2　数据预处理

应用拉曼光谱技术对九种可燃液体样品进行分类识别研究,计算机的数据处理流程如图2所示。

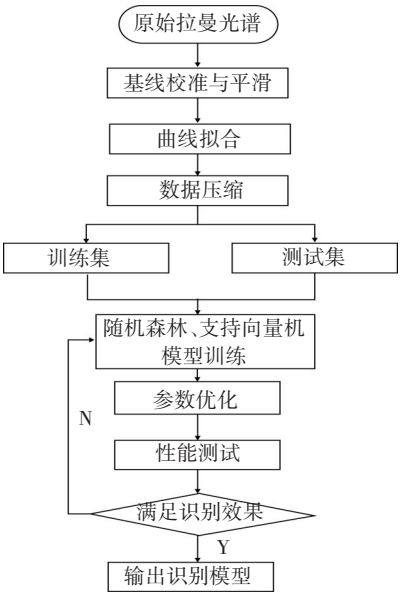

图2　数据处理流程图

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(1)光谱采集

在实验室避光室温环境下,将九份样品逐一取样,然后放入实验室用玻璃瓶中进行拉曼光谱测试。便携式拉曼光谱仪检测时激光波长为785 nm;扫描次数5次,最后获得5次扫描光谱的平均光谱。

(2)光谱基线校准与平滑

光谱预处理方法首先进行波长和拉曼位移之间的转换,通过下式

$$(\frac{1}{\lambda_0}-\frac{1}{\lambda})\times 10^7 \tag{1}$$

其中,λ_0为激发波长,$\lambda_0<\lambda$,求得拉曼光谱图的横坐标。

基线漂移现象是现代仪器在测量信号过程中的常见问题,在拉曼光谱基线漂移情况比较严重的情况下,会阻碍对拉曼光谱的进一步观察和分析,对后续的结果也会造成一定影响,一阶导数(first derivative)和二阶导数(second derivative)是光谱分析中常见的预处理方式。导数光谱法可以很好的消除基线漂移现象以及荧光背景的干扰,有效的提高分辨率和灵敏度,增强样本光谱特征[9-13]。采用Savitzky-Golay滤波器对数据进行平滑处理,提高光谱平滑性,降低噪音干扰,并对拉曼光谱图进行归一化处理,对多峰进行拟合,通过find peaks自动寻峰算法对峰位进行标记。

选取谱图信息比较丰富的波段0~1 100 cm^{-1},然后采用Savitzy-Golay滤波器对谱线进行平滑处理,平滑窗口大小设置为9,最后对谱图进行归一化处理,如图3所示。

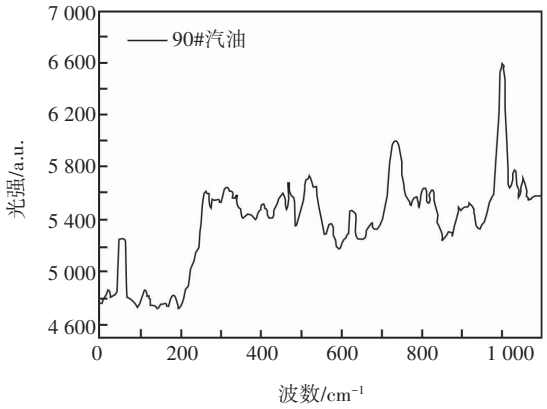

图3 90#汽油原始拉曼光谱图

(3)特征峰提取

为判别重叠峰,特征峰提取效果显著,在提取特征峰前对谱图进行曲线拟合,选用非线性最小二乘算法(Levenberg-Marquardt)实现高斯曲线拟合,然后逐一标记特征峰,如图4~图13分别为90#汽油、93#汽油、97#汽油、丙酮、二甲苯、甲醇、乙醇、乙二醇、叔丁醇的九种典型样品寻峰后的光谱图,可以显示明显的谱峰。表1表示由上述拉曼光谱图分析整理后得到的9种可燃液体的特征峰位。通过图4~图13及表1可以看出,不同种类的可燃液体存在明显的峰位的数量、位置及峰高比的区别,能够为后续自动识别提供数据支持。

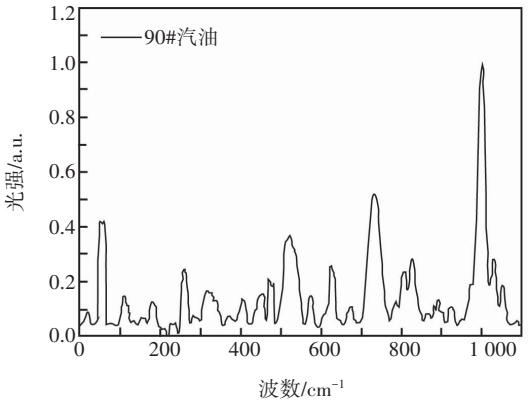

图4 90#汽油预处理后拉曼光谱图

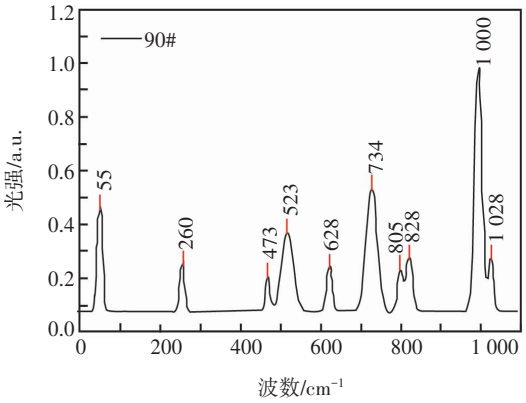

图5 90#汽油的特征峰位

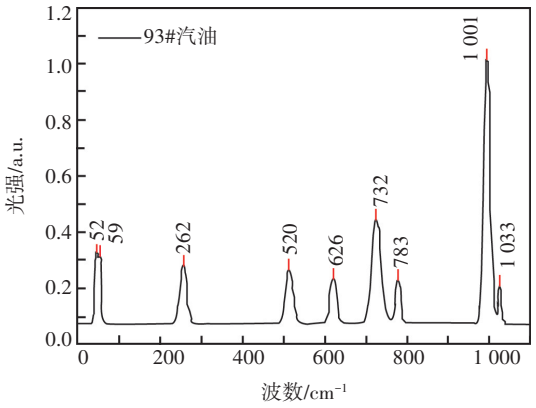

图6 93#汽油的特征峰位

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

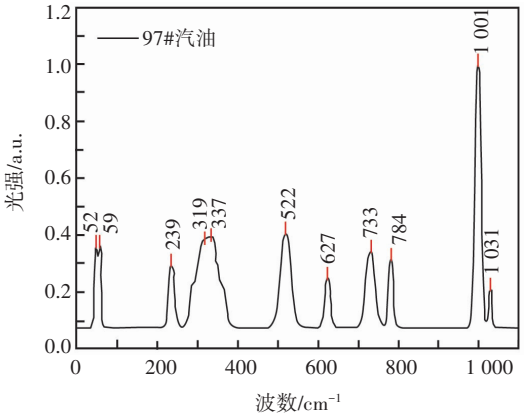

图7　97#汽油的特征峰位

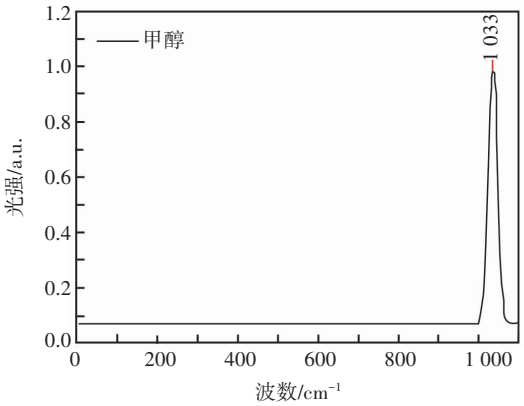

图8　甲醇的特征峰位

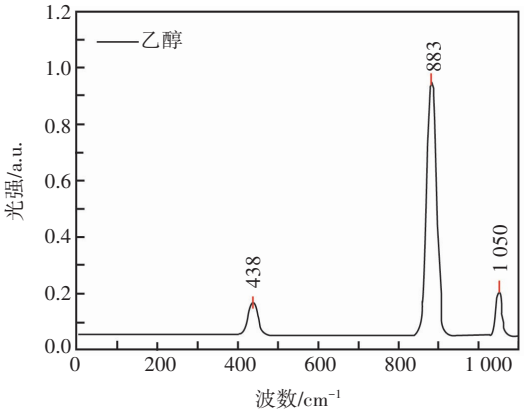

图9　乙醇的特征峰位

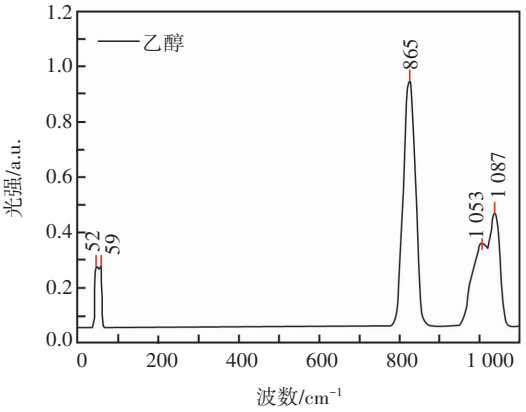

图10　乙二醇的特征峰位

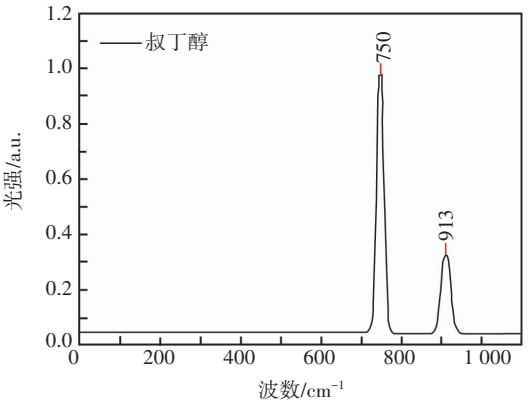

图11　叔丁醇的特征峰位

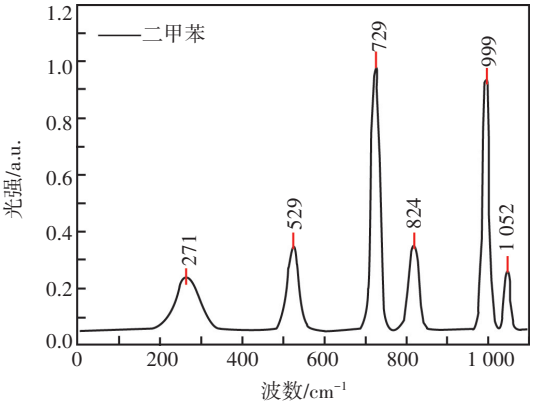

图12　二甲苯的特征峰位

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

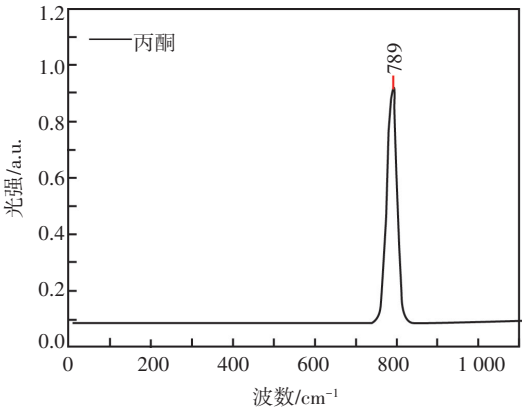

图13 丙酮的特征峰位

表1 9种可燃液体样品的特征峰位

物质名称	特征峰位/(cm^{-1})
90#汽油	55,260,473,523,628,734,850,828,1000,1028
93#汽油	52,59,262,520,626,783,1001,1033
97#汽油	52,59,239,319,337,522,627,733,784,1001,1031
甲醇	1033
乙醇	438,883,1050
乙二醇	52,59,865,1053,1087
叔丁醇	750,913
二甲苯	271,529,729,824,999,1052
丙酮	789

表1表示由上述拉曼光谱图分析整理后得到的9种可燃液体样品的特征峰位。

1.3 拉曼光谱分类决策算法

文中选取SVM和随机森林两种算法构建分类识别模型。支持向量机(support vector machines,SVM)与传统方法相比,支持向量机算法是建立在统计学习理论的VC维理论和结构风险最小原理基础上,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以期获得最好的推广能力。决策树(decision tree,DT)是一种简单但是广泛使用的分类器。通过训练数据构建决策树,可以高效的对未知数据进行分类。决策树算法是一种典型的分类算法,具有很好的健壮性,本质上是一种归纳算法。它从大批无规则、无序的样本中推导出分类规则,并将这些规则以树结构的形式呈现出来[14]。

首先将1 000点的拉曼光谱数据进行512点的数据压缩,将数据集分为训练集(Train)和测试集(Test)两部分,前者用于训练,后者用于评估。采用python中的Scikit-Learn模块,Scikit-Learn提供了一种非常简单的评估方法,即scoring参数,它可以同时用于GridSearchCV和cross_val_score,只需要提供一个字符串,用于描述要使用的评估指标即可。然后对算法进行训练,评估模型,用精度(Accuracy),即表示正确分类的样品所占的比例来判断选择的一系列分类模型的好坏。采用精度(Accuracy)、交叉验证精度(CV accuracy)、均方误差(MSE)作为判定支持向量机算法、随机森林算法识别模型的指标,模型的普通精度、交叉验证精度越接近1,均方误差越接近0,表明模型具有越好的识别精度和性能。

表2是选取3组不同的拉曼光谱数据集时,决策树(DT)和随机森林(RF)算法对同一数据集训练后的普通精度、交叉验证精度和均方误差结果对比。

表2 支持向量机(SVM)和随机森林(RF)算法运行结果对比

数据集序列号	支持向量机(SVM)精度	支持向量机(SVM)交叉验证精度	支持向量机(SVM)均方误差	随机森林(RF)精度	随机森林(RF)交叉验证精度	随机森林(RF)均方误差
1	0.777 8	0.916 7	0.555 6	1.000 0	0.972 2	0.000 0
2	0.875 0	0.906 3	0.500 0	1.000 0	0.968 6	0.000 0
3	0.714 3	0.928 6	0.714 3	0.998 7	0.928 6	0.000 0

可以看出,在同等条件下,随机森林算法的普通精度和交叉验证精度始终都高于SVM算法,且随机森林算法的均方误差低于SVM算法。所以在基于拉曼光谱识别分类可燃液体样品时,选取随机森林算法建立快速识别模型。随机森林能处理高维特征,不容易产生过拟合,模型训练速度比较快,特别是对于大数据而言,对大数据集的适应能力强:既能处理离散型数据,也能处理连续型数据,数据集无需规范化。

2 结 论

针对九种典型的可燃液体样品90#汽油、93#汽油、97#汽油、丙酮、二甲苯、甲醇、乙醇、乙二醇、叔丁醇为例,采用了二阶导数法进行基线校正Savitzky-Golay滤波器对数据进行平滑处理,选用非线性最小二乘算法(Levenberg-Marquardt)实现高斯曲线

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

拟合,最终识别谱峰效果明显,对数据进行512点的压缩。分别比较了支持向量机、随机森林两种算法识别可燃液体,建立随机森林和支持向量机分类模型,通过参数精度、交叉验证精度、均方误差对模型进行判别,结果验证随机森林的各项指标均优于支持向量机,为拉曼光谱技术应用于可燃液体等油气资源的快速检测奠定了基础。后续的工作在此研究的基础上需要进一步扩充样本数据量,为海洋油气资源的探测提供快速检测的技术手段。

参考文献

[1] 席杨.拉曼光谱预处理与拉曼成像插值算法的研究[D].兰州:兰州大学,2018.

[2] 王亚鸽.基于近红外光谱技术的枸杞产地鉴别及品质快速检测研究[D].西安:西北农林科技大学,2015.

[3] 邓之银.基于拉曼光谱的油品快速检测技术与实验研究[D].杭州:浙江大学,2014.

[4] 包丽丽.几种常用油品拉曼光谱的检测及分析[J].光谱学与光谱分析,2012,32(2):394-397.

[5] 李鹏鹏.基于小波分析的太赫兹光谱信号分析研究[D].郑州:河南工业大学,2018.

[6] 高颖.基于拉曼光谱的天然气组分分析方法与系统开发[D].杭州:浙江大学,2019.

[7] Moore D S, Scharff R J. Portable Raman explosives detection[J]. Analytical and Bioanalytical Chemistry, 2009, 393(6-7): 1571-1578.

[8] 田芳明.基于拉曼光谱与有机成分分析的大米身份识别[D].长春:吉林大学,2018.

[9] 章颖强,董伟,张冰,等.基于拉曼光谱和最小二乘支持向量机的橄榄油掺伪检测方法研究[J].-光谱学与光谱分析,2012,32(6):1554-1558.

[10] 王毅.便携式拉曼光谱仪及数据处理关键技术研究[D].成都:电子科技大学,2019.

[11] 刘铭晖,董作人.基于Voigt函数拟合的拉曼光谱谱峰判别方法[J].中国激光,2017,44(5):051103-1-6.

[12] YU Xiao, XU Liang, MO Jian-qing, et al. Raman spectroscopy de-noising based on EEMD combined with VS-LMS algorithm[J]. Optoelectronics Letters, 2016, 12(1): 16-19.

[13] 俞潇.光谱信号及光纤振动信号预处理算法的应用研究[D].乌鲁木齐:新疆大学,2016.

[14] 黄立贤,沈志学.基于决策树的Landsat多光谱影像分类方法[J].光电技术应用,2011(3):49-52.

(上接第38页)

依据第2节关于带隙光纤各模式弯曲损耗的取值标准,可以看出,弯曲半径为10 cm时,所设计的光纤各模式每千米最大的弯曲损耗可以达到3 dB,说明所设计的7-cell HC-PBGF在工程使用中可以保证抗弯损特性。

4 结　论

设计了一种7-cell HC-PBGF,对该光纤的色散、模场面积、限制损耗和弯曲损耗等特性进行仿真分析。所设计的光纤在色散方面具有多零色散点,可以实现特定波长的光传输,同时具有较大的有效模式面积,在波长2 000 nm处,LP_{01}、LP_{02}和LP_{21}模式的有效模式面积可以达到3.248×10^{-10} m^2、3.911×10^{-10} m^{2}和6.525×10^{-10} m^2。该光纤的最佳传输带宽为94 nm,即从波长1 984~2 078 nm,在此波长范围内LP_{01}、LP_{21}和LP_{02}的限制损耗均很小。当弯曲半径为10 cm时,所设计的光纤各模式每千米最大的弯曲损耗可以达到3 dB,说明所设计的7-cell HC-PBGF在工程使用中可以保证抗弯损特性。

参考文献

[1] 迟楠,胡昉辰,周盈君.高速可见光通信技术的挑战与展望[J].中兴通讯技术,2019:1-10.

[2] Cregan R F, Mangan B J, Knight J C, et al. Single-mode photonic band gap guidance of light in air[J]. Science, 1999, 285(5433): 1537-1539.

[3] Mangan B, Farr L, Langford A, et al. Low loss (1.7 dB/km) hollow core photonic bandgap fiber[C]//Optical Fiber Communication Conference, 2004.

[4] Amezcua-Correa R, Leon-Saval S G, Broderick N G R, et al. Control of surface modes in low loss hollow-core photonic bandgap fibers[J]. Optics Express, 2008, 16(2): 1142-1149.

[5] Wheeler N V, Petrovich M N, Slavik R, et al. Wide-bandwidth, low-loss, 19-cell hollow core photonic band gap fiber and its potential for low latency data transmission[C]//National Fiber Optic Engineers Conference, 2012.

[6] Baddela N K, Petrovich M, Jung Y, et al. First demonstration of a low loss 37-cell hollow core photonic bandgap fiber and its use for data transmission[C]//Lasers & Electro-optics, 2014.

[7] 宋镜明,孙康,徐小斌.空芯光子带隙光纤散射损耗特性及结构优化研究[J].中国激光,2015(11):121-126.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net