第41卷，第8期
2021年8月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 41，No. 8，pp2469-2473
August，2021

基于拉曼光谱技术的海水微塑料快速识别技术研究

杨思节[1，2]，冯巍巍[2，3，4*]，蔡宗岐[2，3]，王 清[2，3]

1. 哈尔滨工业大学(威海)，山东 威海 264200
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 近年来由于塑料的大量使用和排放，这些塑料经环境作用破碎变成微塑料大量汇聚到海洋中，导致海洋中聚集大量微塑料。微塑料形状较小，难以识别其来源与种类。激光拉曼探测技术具有快速、无损、且各物质指纹峰明显易被精确识别等优点，近年来被广泛应用。本文基于拉曼光谱探测技术，提出了一种结合小波处理、随机森林算法实现海水中微塑料快速识别的智能分类方法。针对六种典型的海水微塑料标准样品(丙烯腈(A)-丁二烯(B)-苯乙烯(S)的三元共聚物(ABS)、聚酰胺(PA)、聚对苯二甲酸乙二醇酯(PET)、聚丙烯(PP)、聚苯乙烯(PS)、聚氯乙烯(PVC))，采用激光拉曼探测技术进行光谱数据收集，对获取的拉曼光谱采用小波基为DB7、分解次数为3的小波，标准差归一化进行了拉曼光谱预处理。为了提高识别速度，同时还需要对光谱数据进行数据压缩预处理，分别进行了数据压缩点为64，128，256，512和1 024点的数据压缩比较，它们的决策树算法识别精度分别为91.51%，91.67%，92.35%，93.17%和93.21%，随机森林算法识别精度分别为93.12%，93.92%，94.83%，96.81%和96.81%，实验结果表明，微塑料的拉曼光谱压缩为512点时为效率和精度的最佳压缩点，可以为实际工程应用中微塑料拉曼数据压缩提供参考。分别采用决策树、随机森林两种算法进行微塑料拉曼光谱识别研究。研究结果表明，基于拉曼光谱数据，随机森林算法的识别微塑料交叉验证精度高于决策树算法。为进一步提高识别精度，进行了模型参数(折次k)优化研究，采用经过优化后的模型参数($k=20$)，随机森林算法识别微塑料的交叉验证精度可以达到97.24%。可以为实际海水中微塑料的快速识别提供技术参考。

关键词 微塑料；激光拉曼；小波分析；决策树；随机森林

中图分类号：X834 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)08-2469-05

引 言

自从发现微塑料在海洋和海洋生物中无处不在，全球对微塑料的关注已大大增加[1-3]。2015年第二届联合国环境大会上，微塑料污染被列为与全球气候变化、臭氧耗竭和海洋酸化并列的重大全球环境问题[4]。研究表明微塑料已经大量存在于各个大洋、海湾中，例如在北冰洋中发现了高浓度的微塑料[5]，天津近岸海域微塑料污染严重[4]，山东桑沟湾微塑料丰度很高[6]。但是目前国内还没提出成熟的快速智能识别海水中微塑料的方法。

微塑料是指粒径小于5 mm的塑料颗粒，由于其粒径较小，微塑料的识别鉴定仍然是一个挑战[2]。目前研究表明，光谱分析法(FTIR、Raman)和热分析法(Py-GC-MS、TED-GC-MS)应用于微塑料的识别检测频率最高。热分析法容易破坏微塑料的属性，红外光谱分辨率较低且容易受海水的干扰，而拉曼光谱作为红外互补光谱，近年来受到越来越多的关注。通过拉曼光谱的基团频率振动峰对微塑料进行分类鉴别，指纹峰明确易于识别，而且不需要制样、为非破坏性，避免了样品制备过程中可能造成的污染和保持样品的完整性[7]。因此本文基于拉曼光谱探测技术，提出了一种结合小波处理、随机森林算法实现海水中微塑料快速识别的智能分类方法。

收稿日期：2020-08-05，修订日期：2020-12-16

基金项目：国家重点研发计划项目(2019YFD0901101)，山东省重点研发计划项目(2019JZZY010810)资助

作者简介：杨思节，女，1997年生，哈尔滨工业大学(威海)与中国科学院烟台海岸带研究所联合培养硕士研究生
e-mail：19S030089@stu.hit.edu.cn *通讯作者 e-mail：wwfeng@yic.ac.cn

1 实验部分

激光拉曼系统可实现对微塑料的直接测量，不需要对样品进行预处理，并且检测速度快，可以很好地实现微塑料的快速识别。图 1 为激光拉曼探测系统完成微塑料光谱数据收集的过程。有光源控制电路、探测单元和信号处理传输单元，其中探测单元包括激发光源、入射光纤、探头、接收光纤、光谱采集模块，信号处理传输单元包括光谱处理模块、光电转换模块、数据处理模块和数据传输接口。采用 785 nm 的激发光源。

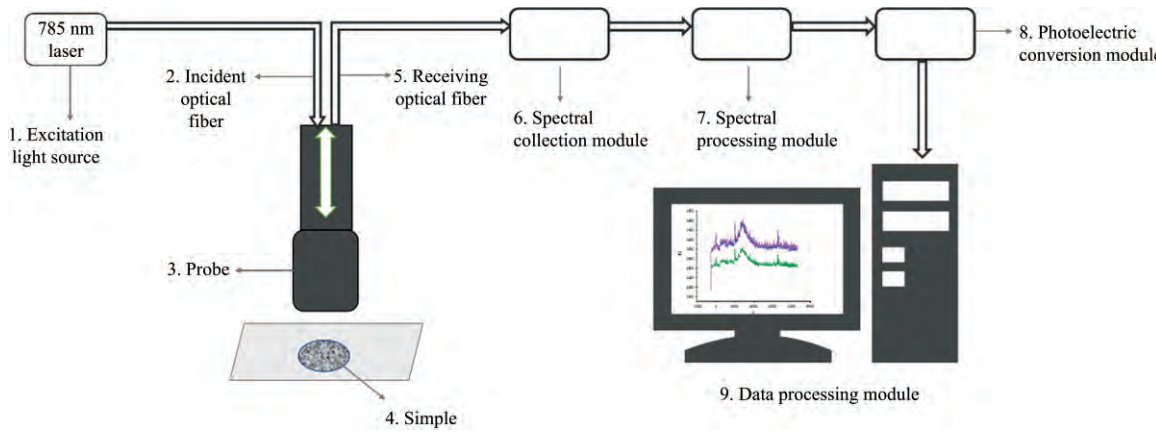

图 1 激光拉曼测微塑料测量系统

1：激发光源；2：激发光纤；3：探头；4：样品；5：接收光纤；
6：光谱收集模块；7：光谱处理模块；8：光电转化模块；9：数据处理模块

Fig. 1 Laser Raman microplastic measuring system

1: Excitation light source; 2: Incident optical fiber; 3: Probe; 4: Simple; 5: Receiving optical fiber
6: Spectral collection module; 7: Spectral processing module; 8: Photoelectric conversion module; 9: Data processing module

1.1 原始拉曼数据获取

选取环境中比较常见的六种微塑料：丙烯腈(A)-丁二烯(B)-苯乙烯(S)的三元共聚物(ABS)、聚酰胺(PA)、聚对苯二甲酸乙二醇酯(PET)、聚丙烯(PP)、聚苯乙烯(PS)、聚氯乙烯(PVC)。选取激发波长为 785 nm 的激光探测器固定在距离标准样品 2 cm 处进行测量，光谱采集模块的光谱范围为 768～1 190 nm，拉曼光谱的积分时间为 500 ms。

1.2 数据预处理

1.2.1 标准差归一化处理

标准差归一化是对拉曼光谱数据进行中心平移变换和无量纲压缩处理，可以用来消除拉曼光谱中激光光源功率变化、光强衰减等影响。分别取波数在 0～4 000 cm^{-1} 共 1 745 个光谱数据进行标准差归一化运算。

1.2.2 小波分析处理

拉曼采集微塑料光谱数据时存在的噪声和荧光背景是影响分析拉曼光谱的主要问题。本文利用小波分析来降低采集的微塑料拉曼光谱的噪声。小波变换(wavelet transform，WT)通过伸缩平移运算对信号(函数)逐步进行多尺度细化，可以局部化分析非平稳信号[8]。根据常用去噪小波函数选取了 Daubechies(DBN)小波。实验发现用 DB7 小波基，分解次数选择 3 次分析微塑料的拉曼光谱最合适。图 2 分别是聚丙烯(PP)原始光谱和经过标准差归一化、DB7 小波分析后的拉曼光谱图。

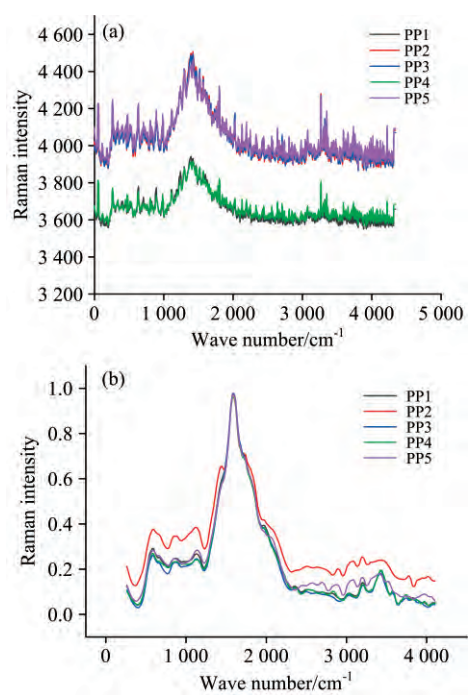

图 2 聚丙烯原始拉曼光谱和经预处理后的拉曼光谱

Fig. 2 Original Raman spectra of polypropylene (a) and Raman spectra after pretreatment (b)

1.2.3 数据压缩预处理

原始拉曼光谱具有 1 745 个数据点，不同的属性对光谱分析具有不同的重要程度，为了提高模型识别速度需要对原始光谱进行数据压缩，利用随机森林算法能评估各个属性在

分类问题上的重要性程度，选出重要性重要程度高的属性，达到数据压缩的目的。

1.3　构建分类识别算法

选择机器学习中的决策树算法和随机森林算法分别构建识别模型，他们都比较适合小样本集的分类识别。决策树(decision tree，DT)算法实现分类的原理[9-10]：构造一种模型，使模型能够从样本数据的特征属性中，通过学习简单的决策规则，从而预测目标变量的值。随机森林(random forest，RF)算法是建立了多个决策树，并将它们合并在一起，最终叶节点是分类问题的多数类。

利用训练数据根据损失函数最小化的原则建立决策树模型。把输入数据集划分成训练集(train)和测试集(test)两部分，模型通过 fit 方法从训练数据集中学习，然后调用 score 方法在测试集上进行评估，打分；从分数上我们可以知道模型当前的训练水平如何。用精度(accuracy)来判断分类(classification)模型的好坏。其中决策树分割算法选择 ID3。

随机森林算法中树的个数需要事先指定，这种需要人工选择的参数称为超参数。超参数选择不恰当，就会出现欠拟合或者过拟合的问题。使用网格搜索(GridSearchCV)来找到一个合适的树个数。最终用 GridSearchCV 确定随机森林算法中树的个数为 100 个。为了调整超参数，测试集的数据会“泄漏”给模型。选择交叉验证(cross-validation，CV)作为精度测试方法，可以很好的解决这些问题。常用 k 折交叉验证，即数据集被划分成 k 个子集，每次训练的时候，用其中 $k-1$ 份作为训练数据，剩下的 1 份作为测试，重复 k 次，然后取 k 次精度的平均值。交叉验证通过多次划分，大大降低了这种由一次随机划分带来的偶然性，同时通过多次训练，模型也能遇到各种各样的数据，从而提高其泛化能力。

数据处理模块流程图如图 3 所示。

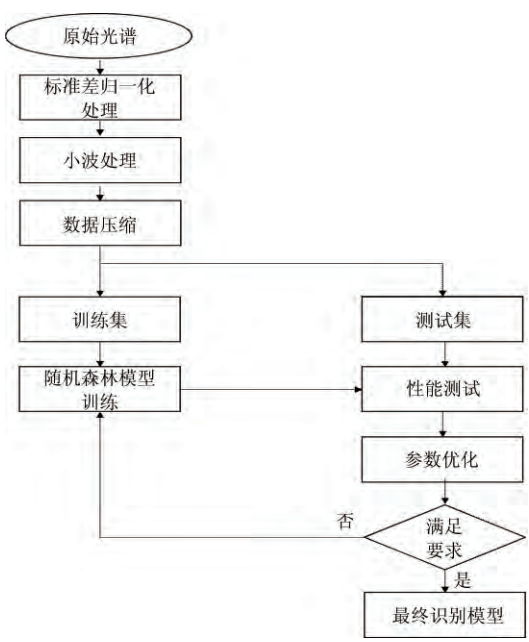

图 3　数据处理模块流程图

Fig. 3　Flow chart of data processing module

2　结果与讨论

采用精度(accuracy)、交叉验证精度(CV accuracy)、均方误差(MSE)作为判定决策树算法、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。

2.1　数据压缩结果与讨论

利用随机森林算法中的属性重要性排列(feature_importances)返回特征的重要性，feature_importances 越高代表特征越重要，然后保留重要程度高的属性，去掉重要程度低的属性，达到数据压缩的目的。

利用拉曼光谱 1 745 个光谱点中的排名前 64，128，256，512 和 1 024 的光谱点分别形成的数据作为决策树算法和随机森林算法的训练数据集，结果如图 4 所示，可以为实际工程应用中选择数据压缩点数提供参考。

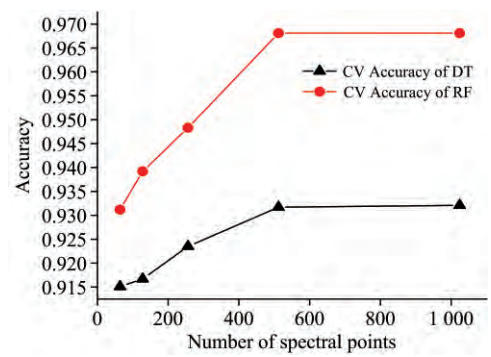

图 4　不同光谱点个数决策树(DT)算法和随机森林(RF)算法的交叉验证精度

Fig. 4　CV accuracy of decision tree (DT) and random forest (RF) with different spectral points number

由图 4 可以看出光谱点个数在 512 之前，随着光谱点个数的增多交叉验证精度增加幅度较大，而在 512 个光谱点之后随着点个数的增加，决策树算法和随机森林算法的交叉验证精度都基本维持不变。最终选取 512 个光谱点，此时的光

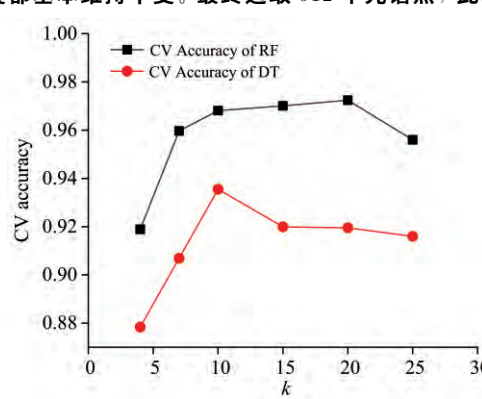

图 5　不同 k 值时决策树(DT)算法模型和随机森林(RF)算法模型交叉验证精度

Fig. 5　CV accuracy of decision tree (DT) and random forest (RF) with different k values

谱点个数较少，既能提高计算速度且又能保证微塑料识别的交叉验证精度，有利于实际工程应用。

2.2 折次(参数 k)对模型精度影响分析

k 折交叉验证中数据集被划分成 k 个子集，每次训练的时候，用其中 $k-1$ 份作为训练数据，剩下的 1 份作为测试，重复 k 次，然后取 k 次精度的平均值。不同模型具有不同的最优 k 值。实验分别取 $k=4$，7，10，15，20，25 对比交叉验证精度，如图 5，来选取识别微塑料模型的最优 k 值。

由图 5 可以看出，不论决策树算法模型还是随机森林算法模型，并不是 k 值越高精度越高，而是随着 k 值的增加精度都会出现拐点，具体模型出现拐点的 k 值可能会有差异。实验结果表明，针对微塑料拉曼光谱识别决策树算法模型智能识别的最优 k 值是 10，此时交叉验证精度可以达到 93.55%。随机森林算法模型智能识别塑料拉曼光谱的最优 k 值是 20，此时交叉验证精度可以达到 97.24%。

2.3 决策树、随机森林算法比较分析

表 1 是选取 5 个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 1 决策树(DT)和随机森林(RF)算法运行结果对比

Table 1 Comparison of operation results between decision tree (DT) and random forest (RF) algorithm

Data set number	Accuracy of DT	CV accuracy of DT	MSE of DT	Accuracy of RF	CV accuracy of RF	MSE of RF
1	0.906 3	0.738 1	1.093 8	1.000 0	0.873 0	0.000 0
2	0.906 3	0.841 3	1.093 8	1.000 0	0.873 0	0.000 0
3	0.957 5	0.838 2	0.383 0	0.936 1	0.924 1	0.368 1
4	0.925 9	0.907 7	0.463 0	0.963 0	0.963 0	0.092 6
5	0.890 3	0.927 4	0.683 8	0.987 0	0.972 0	0.004 5

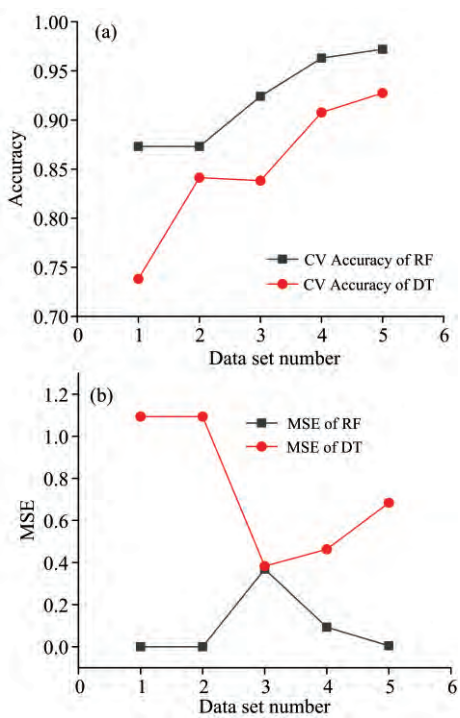

图 6 决策树(DT)和随机森林(RF)算法的运行结果

Fig. 6 Operation results of decision tree (DT) and random forest (RF)

由表 1 和图 6 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决策树算法，且随机森林算法的均方误差低于决策树算法。所以在基于拉曼光谱识别分类微塑料时，选取随机森林算法来建立快速识别模型。这是由于随机森林算法采用自举随机采样技术，而且通过交叉验证避免随机采样结果的偶然性，对非平衡数据具有较好的模型预测性能。

3 结 论

利用激光拉曼检测系统对海水中常见的六种微塑料样品进行了分析，利用 DB7 小波分析方法，标准差预处理对拉曼光谱数据集进行了预处理，为了提高识别速度，同时对光谱数据进行了数据压缩，分别进行了数据压缩点为 64，128，256，512 和 1 024 点的数据压缩比较，它们的决策树算法识别精度分别为 91.51%，91.67，92.35%，93.17% 和 93.21%，随机森林算法识别精度分别为 93.12%，93.92%，94.83%，96.81%和 96.81%。基于精度和效率考虑，最终光谱数据压缩点数选择 512 个点。研究了参数 k 对识别精度的影响。分别比较了决策树、随机森林两种算法识别微塑料。研究结果表明，针对海水中典型的微塑料样品，当 k 值为 20，随机森林算法可以达到 97.24%。可以为实际海水中微塑料的快速识别提供技术参考。

References

[1] Zahra Sobhani, Md Al Amin, Ravi Naidu, et al. Analytica Chimica Acta, 2019, 1077: 191.
[2] Shan Jiajia, Zhao Junbo, Zhang Yituo, et al. Analytica Chimica Acta, 2019, 1050: 161.
[3] Meghnad Pirsaheb, Hooshyar Hossini, Pouya Makhdoumi. Process Safety and Environmental Protection, 2020, 142: 1.

[4] BAI Lu, LIU Xian-hua, CHEN Yan-zhen, et al(白　璐, 刘宪华, 陈燕珍, 等). Environmental Chemistry(环境化学), 2020, 39(5): 1161.
[5] Obbard R W, Sadri S, et al. Earths Future, 2014, 2(6): 315.
[6] Wang Jun, Lu Lin, Wang Mingxiao, et al. Science of the Total Environment, 2019, 667: 1.
[7] XU Xin-xia, SHEN Xue-jing, YANG Xiao-bing, et al(徐昕霞, 沈学静, 杨晓兵, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(6): 1929.
[8] ZHENG Xia, HU Dong-bin, LI Quan(郑　霞, 胡东滨, 李　权). Acta Scientiae Circumstantiae(环境科学学报). DOI: 10.13671/j.hjkxxb.2020.0123(2020.0123:1-8)
[9] Zhou Xiaoyi, Lu Pan, Zheng Zijian, et al. Reliability Engineering and System Safety, 2020, 200: 106931.
[10] LI Ling-ling, LI Yun-mei, LÜ Heng, et al(李玲玲, 李云梅, 吕　恒, 等). Environmental Science(环境科学). DOI: 10.13227/j.hjkx.202003266.
[11] Kappler A, Fischer D, Oberbeckmann S, et al. Anal. Bioanal. Chem., 2016, 408: 8377.
[12] DONG Xin, LI Guo-long, HE Kun, et al(董　鑫, 李国龙, 何　坤, 等). Journal of Mechanical Engineering(机械工程学报), 2020, 56(11): 96.
[13] WANG Zhi-fang, WANG Shu-tao, WANG Gui-chuan(王志芳, 王书涛, 王贵川). Acta Phonica Sinica(光子学报), 2019, 48(4): 0412004.

Study on Rapid Recognition of Marine Microplastics Based on Raman Spectroscopy

YANG Si-jie[1,2], FENG Wei-wei[2,3,4*], CAI Zong-qi[2,3], WANG Qing[2,3]
1. Harbin Institute of Technology (Weihai), Weihai　264200, China
2. CAS 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　266071, China
4. University of Chinese Academy of Sciences, Beijing　100049, China

Abstract　Due to a large amount of use and discharge of plastics, these plastics are broken into microplastics by the environmental effect and gather in the ocean in large quantities, leading to the accumulation of a large number of microplastics in the ocean, inrecent year. Microplastics are small in shape and difficult to identify their source and type. Laser Raman detection technology has been widely used in recent years which have fast, nondestructive and easy identification. In this paper, based on Raman spectral detection technology, an intelligent classification method combining wavelet processing and random forest algorithm is proposed to realize the rapid recognition of microplastics in seawater. The spectral data were collected by using laser Raman detection technology from six typical seawater microplastics standard samples(ABS, PA, PET, PP, PS, PVC), and the obtained spectra were pretreated by wavelet base DB7 and decomposition times 3 and standard deviation normalization. In order to improve the recognition speed, the spectral data is compressed at the same time. The data are respectively compressed to 64, 128, 256, 512 and 1 024 points, and their decision tree algorithm identification accuracy was 91.51%, 91.67%, 92.35%, 93.17% and 93.21% respectively. The random forest algorithm identification accuracy was 93.12%, 93.92%, 94.83%, 96.81% and 96.81%, respectively. The experimental results show that the Raman spectral compression of microplastics is the best compression point for efficiency and precision when the Raman spectral compression is 512 points, which can provide a reference for the Raman data compression of microplastics in practical engineering applications. Two recognition algorithms, decision tree and random forest, were used to study the Raman spectrum recognition of microplastics. The results show that the cross-validation accuracy of the random forest is higher than that of the decision tree. In order to further improve the identification accuracy, the model parameter optimization was carried out, and the cross-validation accuracy of the random forest method for identifying microplastics could reach 97.24% by using the optimized model parameters. It can provide a technical reference for the rapid identification of microplastics in seawater.

Keywords　Microplastics; Laser Raman; Wavelet analysis; Decision tree; Random forest

* Corresponding author　(Received Aug. 5, 2020; accepted Dec. 16, 2020)