近红外光谱的海水微塑料快速识别

吴 雪[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 264209
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 光谱技术与机器学习算法结合快速识别微塑料，为微塑料的现场检测提供了极大的技术支持，是一个得到极大关注的新领域。近红外光谱检测技术具有检测速度快、灵敏度高、不损坏样品，且可以在不对样品进行预处理的情况下直接检测等特点，在化学分析、质量检测等领域广泛应用。本文基于近红外光谱检测技术，研究比较了结合 Support Vector Machine(SVM)和 Extreme Gradient Boosting(XGBoost)两种机器学习分类算法，构建微塑料的高速有效识别分类模型。采用微型近红外光谱仪采集了 20 种常见的微塑料标准样品的光谱数据，为了防止过拟合，对每种样品多次采样，共收集了 1 260 个微塑料样本，每个样本包含 512 个数据点。利用 XGBoost 算法进行特征重要性排序，共提取了对识别准确率影响较大的 65 个数据点。分别采用 SVM 算法和 XGBoost 算法对数据降维后提取的 65 个数据点建立微塑料快速识别模型，并运用网格搜索(GridSearchCV)对 XGBoost 算法影响较大的超参数进行选取，确定 n_estimators，learning_rate，min_child_weigh，max_depth，gamma 的最佳超参数分别为 700，0.07，1，1，0.0。为了提高模型的稳定性，识别速率和泛化能力，对模型采用 10 折交叉验证和混淆矩阵评估；研究结果表明，XGBoost 模型对微塑料的识别准确率为 97%，而 SVM 模型对微塑料的识别准确率为 95%；XGBoost 模型对微塑料识别的正确率优于 SVM 模型。综上所述，XGBoost 模型微塑料识别整体性能优于 SVM 模型，为实际微塑料快速识别提供技术支撑。

关键词 微塑料；近红外光谱；XGBoost；SVM

中图分类号：O434 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)11-3501-06

引 言

塑料制品在日常生活中随处可见，迄今为止全球生产的 83 亿吨塑料中，有 10%以塑料碎片的形式积累在海洋和淡水系统中，使塑料污染成为快速增长的环境问题[1]。塑料经过降解，变为微塑料，微塑料可能作为吸附污染物、病毒等的载体对人类和其他生命形式具有潜在的危害。为了研究微塑料在环境中的运输过程以及对环境的污染情况，在现场对微塑料进行识别检测是非常有必要的[2]。

现阶段，微塑料的识别检测方法大部分为目视法，光谱法和热分析法[3]，然而目视法具有很大的主观性，热分析法在检测过程容易损坏样品，光谱法包括拉曼光谱法和近红外吸收光谱法，拉曼光谱不仅需要大量的数据预处理[4]，而且由于荧光作用的影响，需要对样品进行前处理[5]。近红外光谱检测技术利用近红外吸收带探测聚合物官能团的拉伸和弯曲模式，通过微塑料独特的化学成分和成键模式识别微塑料[2]。机器学习算法在数据处理方面表现出强大的性能，利用机器学习算法与近红外光谱结合，可以实现现场实时快速检测，具有快速，高效，无损等特点。

采用近红外光谱检测结合 XGBoost 机器学习分类算法可快速识别海水中的微塑料，不仅操作简单，适用范围广，而且携带方便，可以实现现场实时快速检测[6]。

收稿日期：2021-09-06，**修订日期**：2022-04-06

基金项目：国家重点研发计划项目(2019YFD0901101)资助

作者简介：吴 雪，1995 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：20S030091@stu.hit.edu.cn *通讯作者 e-mail：wwfeng@yic.ac.cn

1 实验部分

1.1 微塑料样品近红外光谱检测系统

近红外光谱检测系统由照明系统、分光系统以及接收系统组成，图 1 为微型近红外光谱检测系统结构示意图。首先利用带光源的积分球测得未放置样品时的光通量，然后放置样品进行测量；测得的样品光谱数据通过接收光纤经光谱收集模块和光谱处理模块进行处理，处理后的光谱数据经光电转换模块进行光电转换后进入数据处理模块进行数据处理。

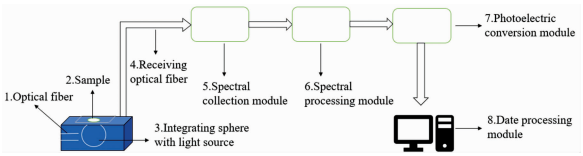

图 1 近红外光谱检测系统

1：光纤；2：样品；3：带光源的积分球；4：接收光纤；5：光谱收集模块；6：光谱处理模块；7：光电转化模块；8：数据处理模块

Fig. 1 Near Infrared microplastic measuring system

1：Optical fiber；2：Sample；3：Integrating sphere with light source；4：Receiving optical fiber；5：Spectral collection module；6：Spectral processing module；7：Photoelectric conversion module；8：Date processing module

1.2 模型评价方法

K 折交叉验证将原始数据集划分为相等的 K 份(“折”)，选取其中的一份作为测试集，其他部分作为训练集，重复 K 次，通过训练集来计算模型的准确率，取平均准确率为模型的最终准确率[4]。其中 10 折交叉验证不仅能准确描述模型的泛化能力，而且具有较好的稳定性和识别速率[7]。本工作采用 10 折交叉验证。

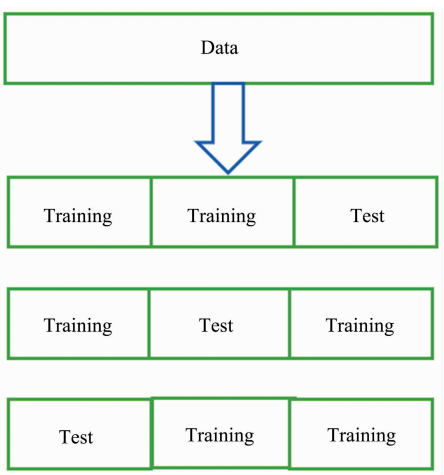

图 2 K 折交叉验证结构示意图

Fig. 2 K-fold cross-validation structure

混淆矩阵(confusion matrix)又称误差矩阵，一种特定的矩阵呈现算法性能的可视化效果，每一列代表预测值，每一行代表实际的类别，混淆矩阵能够全面的反映模型的性能[8]。

1.3 分类识别算法构建

1.3.1 数据降维

原始的近红外光谱在 900～1 750 nm 范围内具有 512 个数据波段，不同的近红外光谱数据波段对模型识别的准确率的重要性程度不同，XGBoost 算法属于集成算法，在特征考虑方面相对全面。因此利用 XGBoost 算法对近红外光谱数据波段的重要性程度进行评估，筛选出重要性程度高的特征波段，达到提高模型识别准确率和速度的目的[9]。

1.3.2 SVM，XGBoost 模型的构建

支持向量机算法利用超平面分离数据点，通过最大化超平面到两个子类中两个最近数据点的距离(即边距 m)[10]，达到分类的目的。

XGBoost 是一种基于树的集成算法，内部决策树采用回归树[11]，该算法已被证明是一种可靠、高效的机器学习问题解决器[12]。XGBoost 算法不断通过误差添加回归树进行拟合，然后把这些回归树进行集成划分进行分类。

图 3 为 XGBoost 算法和 SVM 算法识别微塑料模型建立的流程，数据集输入后利用 XGBoost 算法进行特征提取，然后进行重组，将预处理后的数据集分为测试集和训练集两部分，利用 SVM 算法和 XGBoost 算法对训练集进行学习建模[13]，利用测试集评估模型的整体性能。

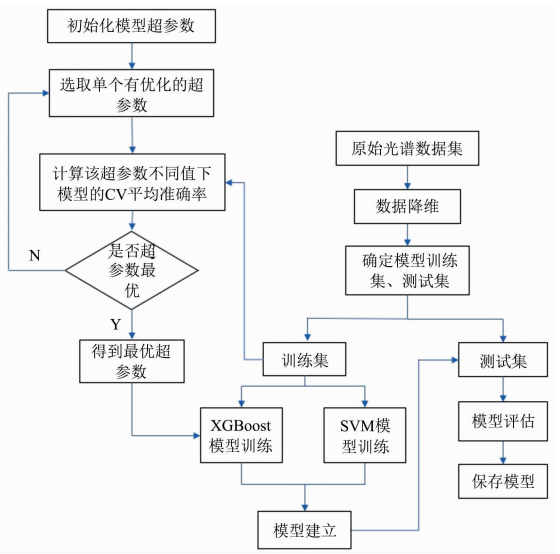

图 3 XGBoost 和 SVM 算法识别微塑料

Fig. 3 XGBoost and SVM for microplastic detection

1.3.3 模型参数的选择

XGBoost 包含较多的超参数，目前对超参数的调整选择并没有明确的规则[11]，本文运用网格搜索(GridSearchCV)对模型影响较大的超参数 n_estimators 即迭代次数、learning_rate 即学习率、min_child_weigh 即最小的叶子节点权重、max_depth 即树的最大深度、gamma 即叶子节点分裂时所需要的最小的损失减小量进行选取。

GridSearchCV 不仅可以遍历每一种参数的可能性，找到最佳参数，而且可以利用交叉验证有效的避免偶然性[14]。

2　结果与讨论

2.1　原始近红外光谱数据获取

采用微型近红外光谱仪对丙烯腈、丁二烯、苯乙烯的三元共聚物(acrylonitrile butadiene styrene，ABS)，聚丙烯腈(polyacrylonitrile，PAN)，聚碳酸酯(polycarbonate，PC)，聚对苯二甲酸乙二醇酯(polyethylene glycol terephthalate，PET)，聚甲基丙烯酸甲酯(polymethyl methacrylate，PMMA)，聚丙烯(polypropylene，PP)，聚苯乙烯(polystyrene，PS)，聚氯乙烯(polyvinyl chloride，PVC)，热塑性聚氨酯(thermoplastic polyurethane，TPU)，乙烯-醋酸乙烯酯共聚物(ethylene-vinyl acetate copolymer，EVA)，聚对苯二甲酸丁二醇酯(polybutylene terephthalate，PBT)，聚己内酯(polycaprolactone，PCL)，聚醚砜(polyethersulfone，PES)，聚乳酸(polylactic acid，PLA)，聚甲醛(polyoxymethylene，POM)，聚苯醚(polyphenylene oxide，PPO)，聚苯硫醚(polyphenylene sulfide，PPS)，聚四氟乙烯(poly tetra fluoroethylene，PTFE)，聚乙烯醇(polyvinyl alcohol，PVA)，苯乙烯-丁二烯-苯乙烯嵌段共聚物(styrenic block copolymers，SBS)20 种常见的微塑料样品采集近红外光谱数据。选取 900～1 750 nm 近红外光谱波段，可准确地检测出 PC，PET，PS，TPU，PBT，PES，PPO，PPS 和 SBS 的苯环吸收振动峰以及 PC，PET，PMMA，TPU，EVA，PBT，PCL 和 PLA 的酮羰基吸收振动峰。积分时间 150 ms，对 20 种微塑料测取了 1 260 个样本数据。每种微塑料样品的部分近红外光谱图如图 4 所示。

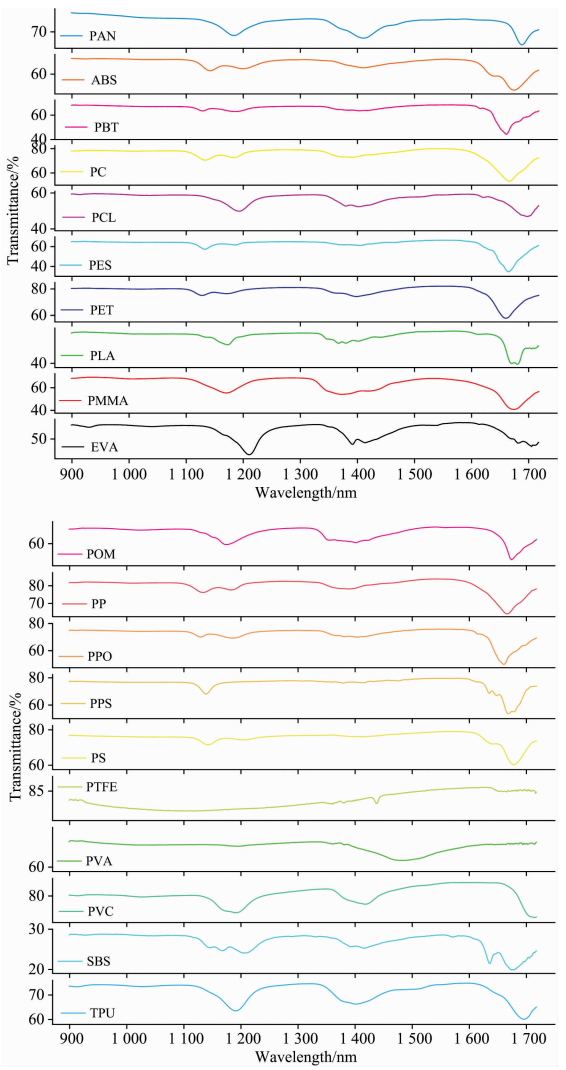

图 4　微塑料样品原始近红外光谱

Fig. 4　Original Near Infrared spectra of microplastics

2.2　数据降维

利用 XGBoost 算法对近红外光谱 512 个特征波段的重要性进行评估，筛选出了重要性程度高的 65 个特征波段[4]，对数据进行降维，图 5 为 XGBoost 筛选出的重要性程度位于前 30 的数据点。

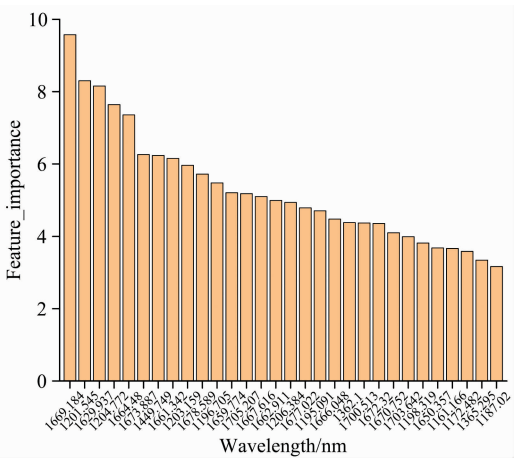

图 5　重要程度位于前 30 的光谱变量

Fig. 5　The 30 most important spectral variables

2.3　XGBoost 参数对比评价

运用 GridSearchCV 对 n_estimators，learning_rate，gamma 分别在 100～900，0.01～0.09，0.00～0.40 中网格搜索选取最佳超参数，其中 min_child_weigh 和 max_depth 一起调参，在 1，3，5 中进行最佳超参数选择。如图 6 所示，确定 n_estimators 取 700，learning_rate 取 0.07，min_child_weigh＝1 和 max_depth＝1，gamma 取 0.0 为最佳超参数。

2.4　SVM 与 XGBoost 模型评估对比

使用已进行降维操作后的测试集样本对已建立的 SVM 和 XGBoost 模型进行评估。由图 7 的 SVM 混淆矩阵可以看出，SVM 模型对 11 种微塑料的识别准确率达到 100%，有 4 种微塑料的识别准确率达到 90%以上，有 4 种微塑料的识别准确率达到 80%以上，1 种微塑料的识别准确率为 76%。由图 7 的 XGBoost 混淆矩阵可以看出，XGBoost 模型对 15 种微塑料的识别准确率达到 100%，识别准确率达到 90%以上的有 3 种，2 种微塑料的识别准确率达到 83%以上。

由表 1 和图 8 可以看出，在同等条件下，XGBoost 模型的识别准确率为 97%，而 SVM 模型的识别准确率为 95%；且 XGBoost 模型的 Accuracy score，Precision score，Recall

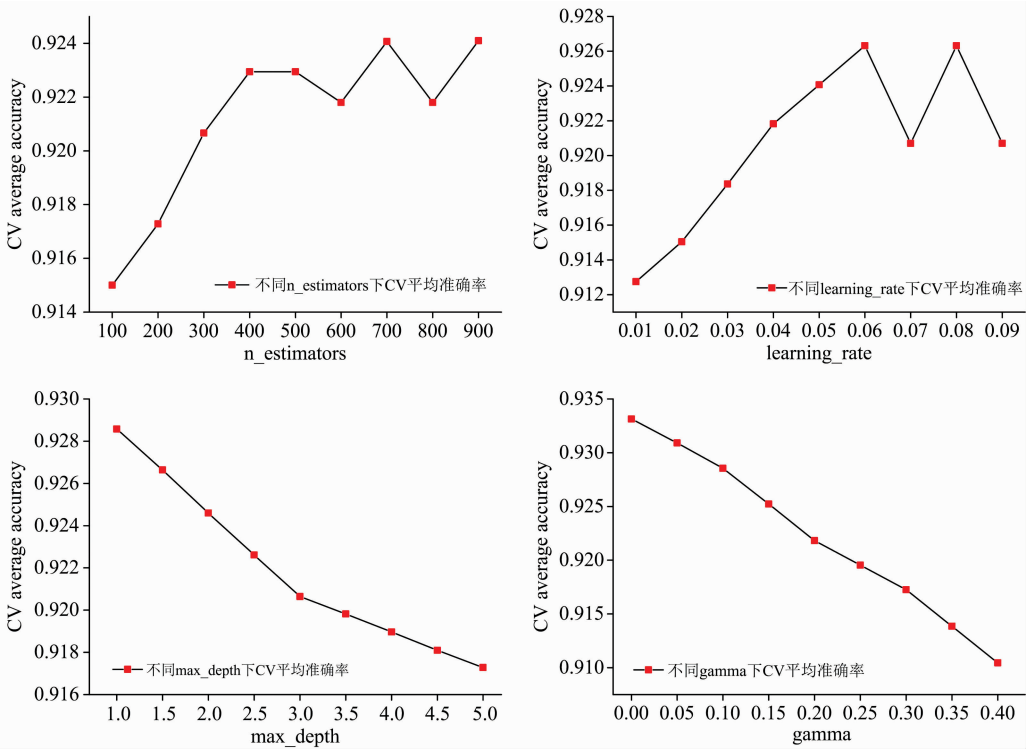

图 6　不同 n_estimators，leasning_rate，max_depth，gamma 下 CV 平均准确率

Fig. 6　CV average accuracy of n_estimators，learning_rate，max_depth，gamma

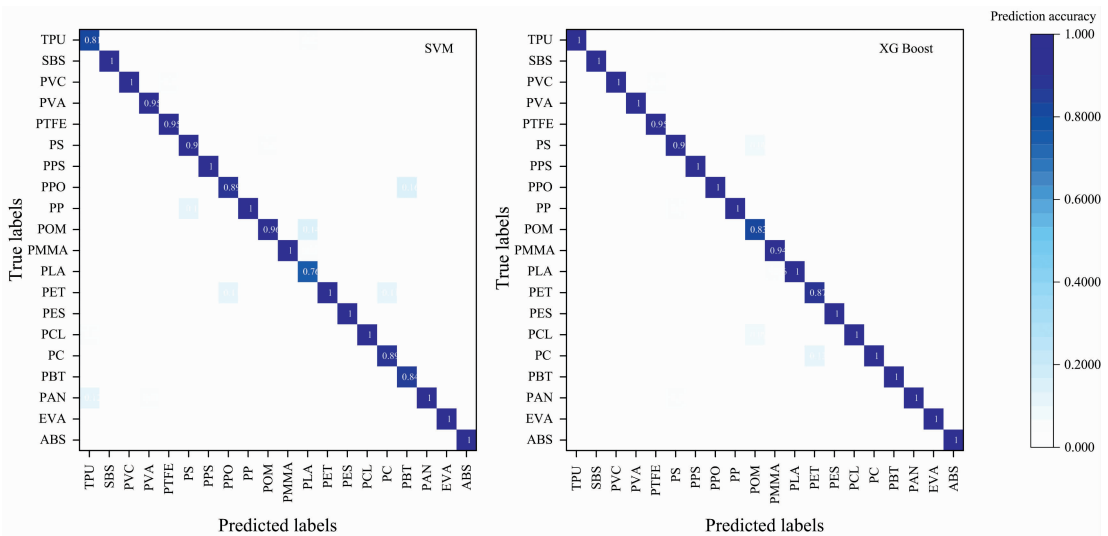

图 7　SVM 和 XGBoost 混淆矩阵

Fig. 7　SVM and XGBoost confusion matrixes

表 1　XGBoost 和 SVM 算法运行结果对比

Table 1　Comparison of operation results between XGBoost and SVM algorithms

Microplastic category	Precision of XGBoost	Recall of XGBoost	F1-score of XGBoost	Precision of SVM	Recall of SVM	F1-score of SVM
ABS	1.00	1.00	1.00	1.00	1.00	1.00
EVA	1.00	1.00	1.00	1.00	1.00	1.00
PAN	0.93	1.00	0.97	0.82	1.00	0.90
PBT	1.00	1.00	1.00	1.00	0.84	0.91

续表 1

PC	0.86	1.00	0.92	1.00	0.89	0.94
PCL	0.91	1.00	0.95	0.95	1.00	0.98
PES	1.00	1.00	1.00	1.00	1.00	1.00
PET	1.00	0.87	0.93	0.85	1.00	0.92
PLA	0.95	1.00	0.98	1.00	0.76	0.86
PMMA	1.00	0.94	0.97	0.94	1.00	0.97
POM	1.00	0.83	0.90	0.88	0.96	0.92
PP	0.95	1.00	0.97	0.90	1.00	0.95
PPO	1.00	1.00	1.00	0.84	0.89	0.86
PPS	1.00	1.00	1.00	1.00	1.00	1.00
PS	0.90	0.90	0.90	0.95	0.90	0.93
PTFE	1.00	0.95	0.97	1.00	0.95	0.97
PVA	1.00	1.00	1.00	1.00	0.95	0.97
PVC	0.94	1.00	0.97	0.94	1.00	0.97
SBS	1.00	1.00	1.00	0.95	1.00	0.98
TPU	1.00	1.00	1.00	1.00	0.81	0.90

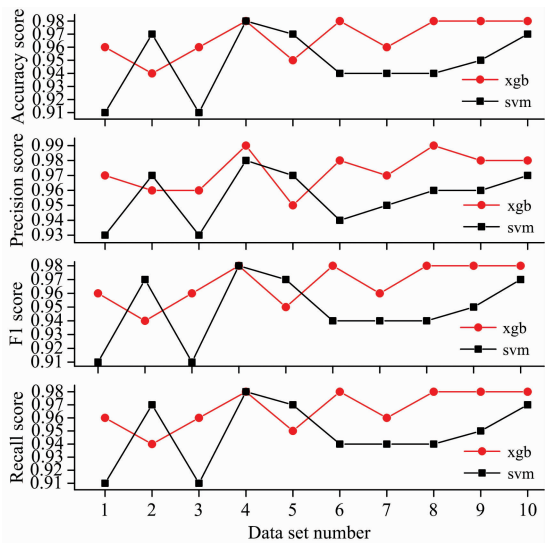

图 8　XGBoost 和 SVM10 折交叉验证下准确率，精准率，召回率，*F* 值

Fig. 8　Accuracy score, precision score, recall and F1-score of 10-fold cross validations of XGBoost and SVM

和 F1-score 的平均准确率均高于 SVM 模型。综上所述，SVM 模型的整体性能低于 XGBoost 模型。

3　结　论

微塑料可能作为吸附污染物、病毒等的载体对人类和其他生命形式具有潜在的危害，为了研究微塑料在环境中的运输过程以及对环境的污染情况，在现场对微塑料进行识别检测是非常有必要的。通过近红外光谱检测系统测得环境中常见的 20 种微塑料标准样品的光谱数据，利用 XGBoost 特征重要性排序，提取 65 个光谱数据点，对数据降维。运用 GridSearchCV 对影响 XGBoost 模型较大的超参数进行选取，确定 n_estimators，learning_rate，min_child_weigh，max_depth，gamma 的最佳超参数分别为 700，0.07，1，1，0.0。对 XGBoost 模型和 SVM 模型进行 10 折交叉验证评估和混淆矩阵评估，确定 XGBoost 模型、SVM 模型对 20 种微塑料的识别准确率分别为 97%和 95%；通过混淆矩阵可以看出 XGBoost 模型对微塑料识别的准确率优于 SVM 模型。综上所述，XGBoost 模型微塑料识别整体性能优于 SVM 模型，为实际微塑料快速识别提供技术支撑。

References

[1] Velimirovic M, Tirez K, Voorspoels S, et al. Analytical and Bioanalytical Chemistry, 2021, 413(24): 7.

[2] Michel A P M, Morrison A E, Preston V L, et al. Environ. Sci. Technol., 2020, 54(17): 10630.

[3] LUO Yong-ming, SHI Hua-hong, TU Chen, et al(骆永明, 施华宏, 涂　晨, 等). Chin. Sci. Bull.(科学通报), 2021, 66: 1547.

[4] YANG Si-jie, FENG Wei-wei, WANG Qing, et al(杨思节, 冯巍巍, 王　清, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2021, 41(8): 2469.

[5] Sommer C, Schneider L M, Nguyen J, et al. Marine Pollution Bulletin, 2021, 171: 112789.

[6] Liu Haitao, Niu Shuoran, Zhou Ying, et al. Micromachines, 2021, 12(6): 696

[7] LIANG Zi-chao, LI Zhi-wei, LAI Keng, et al(梁子超, 李智炜, 赖　铿, 等). Chinese Journal of Hospital Statistics(中国医院统计), 2020, 27(4): 289.

[8] Lemoine M, Piriou M, Charpentier A, et al. Small Ruminant Research, 2021, 202: 106469.

[9] CHE Hong-xin, WANG Tong, WANG Wei(车宏鑫, 王　桐, 王　伟). Data Analysis and Knowledge Discovery(数据分析与知识发现), 2021, 5(9): 107.

[10] Huu P N, Ngoc T P. Journal of Robotics, 2021, 2021: 3986497.
[11] WANG Xing-yu, LUO Yu, Osawa(王星宇, 罗 宇, 大 沢). Hot Working Technology(热加工工艺), 2021, http://doi.org/10.14158/j.cnki.1001-3814.20202994.
[12] Hu C A, Chen C M, Fang Y C, et al. BMJ Open, 2020, 10(2): e033898.
[13] LIU Wen-fang, HAN Jun, LIU Yan-feng, et al(刘文芳, 韩 军, 刘艳锋, 等). China Measurement & Test(中国测试), 2022, 48(1): 6.
[14] YE Tao, SI Qiao-rui, SHEN Chun-hao, et al(叶 韬, 司乔瑞, 申纯浩, 等). Journal of Drainage and Irrigation Machinery Engineering(排灌机械工程学报), 2021, 39(9): 884.

Study on Rapid Recognition of Microplastics Based on Infrared Spectroscopy

WU Xue[1, 2], FENG Wei-wei[2, 3, 4*], CAI Zong-qi[2, 3], WANG Qing[2, 3]
1. Harbin Institute of Technology, Weihai, Weihai 264209, China
2. Key Laboratory of Coastal Environmental Process and Ecological Restoration (Yantai Institute of Coastal Zone), 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 The combination of spectroscopic technology and machine learning algorithm for rapid identification of microplastics provides great technical support for microplastics' field detection, a new field that has attracted great attention. Nirs detection technology has the characteristics of fast detection speediness, highly sensitization, damage less, and can be directly detected without sample pretreatment, widely used in chemical analysis quality detection and other fields. This paper compares support vector machine (SVM) and Extreme Gradient Boosting (XGBoost), two machine learning classification algorithms based on the infrared spectrum, to build a classification model for high-speed and effective recognition of microplastics. Acrylonitrile butadiene styrene(ABS), Polyacrylonitrile (PAN), Polycarbonate (PC), Polyethylene glycol terephthalate(PET), Polymethyl methacrylate (PMMA), Polypropylene (PP), Polystyrene (PS), Polyvinyl chloride (PVC), Thermoplastic polyurethane (TPU), Ethylene-vinyl acetate copolymer (EVA), Polybutylene terephthalate (PBT), Polycaprolactone (PCL), Polyethersulfone (PES), Polylactic acid (PLA), Polyoxymethylene (POM), Polyphenylene Oxide (PPO), Polyphenylene sulfide (PPS), Poly tetra fluoroethylene (PTFE), polyvinyl alcohol (PVA), Styrenic Block Copolymers (SBS) 20 standard samples of microplastics were collected by using A miniature near-infrared spectrum. In order to prevent overfitting, 1 260 microplastic samples were collected, each sample containing 512 data points. The XGBoost algorithm was used to rank the importance of the logarithmic data points, and a total of 65 data points which greatly influenced the recognition accuracy were extracted. SVM algorithm and XGBoost algorithm are respectively used to establish a microplastic fast recognition model for 65 data points extracted after dimensionality reduction, and GridSearchCV is used to select the hyperparameters that have a great influence on XGBoost algorithm to determine n_estimators, learning_rate, The optimal hyperparameters for min_child_weigh, max_depth, and gamma are 700, 0.07, 1,1, 0.0, respectively. In order to improve the model's stability, recognition rate and generalization ability, a 10-fold cross-validation and confusion matrix were used to evaluate the model. The results show that the recognition accuracy of the XGBoost model is 97%, while that of the SVM model is 95%. The accuracy of the XGBoost model is better than the SVM model. In conclusion, the overall performance of the XGBoost model was better than that of the SVM model, which provides technical support for rapid identification of actual microplastics.

Keywords Microplastics; Near infrared spectrum; XGBoost; SVM

(Received Sep. 6, 2021; accepted Apr. 6, 2022)

* Corresponding author