DOI:10.13652/j.spjx.1003.5788.2024.80001

基于近红外光谱的翡翠贻贝重金属铅污染识别

姜 微[1] 刘忠艳[1] 刘 瑶[2] 熊建芳[1] 曾绍庚[1]

（1.岭南师范学院计算机与智能教育学院,广东 湛江 524048;
2.岭南师范学院电子与电气工程学院,广东 湛江 524048）

摘要:[**目的**]通过近红外光谱技术解决贻贝重金属铅污染问题。[**方法**]应用近红外反射光谱结合模式识别的方法进行重金属铅污染检测。首先获得了在950～1 700 nm范围内的健康贻贝和重金属铅污染贻贝光谱数据,应用基于随机变量组合的变量重要性分析(variable importance analysis based on random variable combination,VIAVC)波段选择算法对光谱数据降维,筛选最佳波段子集。针对检测健康贻贝和重金属铅污染贻贝是一个不平衡的分类问题,研究探索一种基于万有引力的固定半径最近邻(gravitational fixed radius nearest neighbor,GFRNN)方法用于贝类重金属铅污染识别。[**结果**]相较于传统的K最近邻法、固定半径近邻法和支持向量机算法,研究提出的VIAVC-GFRNN方法在检测重金属铅污染方面表现出更优异的性能,并且不受样本不平衡率的影响。VIAVC-GFRNN模型的接收者操作特征曲线下面积值达到了0.988 6,检测精度和几何均值均达99.17%。[**结论**]近红外光谱结合模式识别方法在检测贻贝中铅污染方面具有很大的潜力。

关键词:近红外光谱;贻贝;重金属检测;不平衡分类

Identification of heavy metal Pb pollution in *Perna viridis* based on near-infrared spectroscopy

JIANG Wei[1] *LIU Zhongyan*[1] *LIU Yao*[2] *XIONG Jianfang*[1] *ZENG Shaogeng*[1]

（1. *School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang, Guangdong* 524048, *China*; 2. *School of Electronic and Electrical Engineering, Lingnan Normal University, Zhanjiang, Guangdong* 524048, *China*）

Abstract: [**Objective**] Addressing the heavy metal lead pollution in oysters using near-infrared spectroscopy technology.[**Methods**] This study proposed the use of near-infrared reflectance spectroscopy combined with pattern recognition for detecting Pb contamination. Initially, spectral data of healthy mussels and Pb-contaminated mussels in the range of 950～1 700 nm were collected. The wavelength selection algorithm of variable importance analysis based on the random variable combination (VIAVC) was utilized to reduce the dimensionality, and selected the optimal subset of wavelengths. Considering the detection of healthy mussels and Pb-contaminated mussels as an imbalanced classification problem, the gravitational fixed radius nearest neighbor (GFRNN) method based on universal gravity was explored for identifying Pb contamination in mussels. [**Results**] The experimental results demonstrated that the proposed VIAVC-GFRNN method outperformed traditional algorithms such as K-nearest neighbor, fixed radius nearest neighbor, and support vector machine algorithms in detecting Pb contamination, while remaining unaffected by the imbalance ratio. The area under the receiver operation curve value of the VIAVC-GFRNN model reached 0.988 6, with a detection accuracy and geometric mean of 99.17%. [**Conclusion**] Near-infrared spectroscopy combined with pattern recognition methods has great potential for detecting Pd pollution in mussels.

Keywords: near-infrared spectroscopy; mussels; heavy metal detection; unbalanced classification

---

基金项目:国家自然科学基金青年科学基金项目(编号:62005109);广东省科技创新战略专项资金竞争性项目(编号:2023A01025);岭南师范学院红树林生态系统智能监测创新团队项目

通信作者:刘瑶(1982—),女,岭南师范学院副教授,博士。E-mail:liuyao@lingnan.edu.cn

收稿日期:2024-01-01 **改回日期:**2024-05-18

贻贝是一种富含营养物质的海产品，但被生存环境中重金属污染的贻贝可能对人体健康造成危害。传统的一些化学检测手段如石墨炉原子吸收光谱法[1-2]、空气—乙炔火焰原子吸收光谱法[3]、电感耦合等离子体质谱法[4]等，已被用于检测贻贝中的重金属含量。然而，这些方法都需要在高温微波消解系统中使用浓硝酸进行样品消解，然后进行稀释和过滤，最后才进行重金属分析，设备昂贵、样品预处理因素多、需要经验丰富的技术人员，具有破坏性，时间消耗大且无法处理众多样品。

近红外光谱（near-infrared spectroscopy，NIRS）技术作为一种无损、实时的分析方法，已逐渐被广大科研人员钟爱并应用在了食品质量和安全检测评估中，如水果[5]、蔬菜[6]、乳制品[7]、谷类[8]、海鲜[9]、肉类[10]等。但重金属不吸收近红外区域的能量且没有相关波长[11]，无法直接利用光谱特性来检测贻贝中的重金属，主要是通过贻贝在受到污染时触发的复杂防御机制所引起的相关生物分子结构和浓度变化来间接反映[12]。因此，通过贻贝的反射光谱，有望间接检测到隐藏的重金属污染情况。目前，使用NIRS技术探测贝类中的重金属，相关的研究较少。Liu等[13-14]借助近红外光谱法区分贻贝样品重金属污染研究。上述应用NIRS研究的方法，在提高检测精度、检测速度方面取得了较好的效果，但仍有一些关键问题需要进一步探索，例如波段选择方法和模型的泛化能力。

在NIRS数据建模过程中，波长选择是一个关键问题。在检测重金属铅污染时，有些波长是不相关或多余的，对模型的预测能力造成影响。为了建立可靠的校准模型，需要进行适当的波长选择，以提高模型的预测准确性。因此，研究采用VIAVC法解决波长选择的问题，该方法允许提取与样本特定信息相关的有效波长，从而优化建模过程。在实际应用中，贻贝重金属污染样本分类不均衡是一个常见的问题，即被污染样本数量较少。研究拟将GFRNN法[15]应用于贝类重金属污染识别应用中。该算法的目标是以类间样本之间的相似度来进行分类，通过引入万有引力的概念，算法可以更准确地区分健康贻贝和受污染贻贝样本，使得分类器在处理不平衡数据集时更加稳健。

1　材料与方法

1.1　试验制备

研究中所用贻贝样本来自中国广东湛江海鲜市场。根据NY 5073—2006《无公害食品　水产品中有毒有害物质限量》的规定，可以将健康的贻贝定义为铅含量符合国家限量标准（Pb≤1.0 mg/kg）的贻贝，而将超过这一标准的贻贝定义为非健康的贻贝。这样的定义符合实际应用的需求，确保了对贻贝健康状况的准确分类和有效监测。试验前为了确定金属溶液[$Pb(CH_3COO)_2 \cdot 3H_2O$]的浓度，参照国家海水养殖水质标准要求和相关文献[16]报道中的养殖环境中的重金属铅浓度不高于0.05 mg/L，并根据急性毒性试验方法，根据前期多次预试验的结果，最终确定试验的铅溶液质量浓度为0.9 mg/L模拟重金属铅污染的环境[17]。贻贝样本被暴露在30‰盐度的海水中，实验室环境温度维持在26～29 ℃。海水的pH值调整为8.0，溶解氧含量为6.5 mg/L。试验过程中，另设一组贻贝样本在没有添加任何重金属的海水中饲养，作为对照组。为了保持海水质量，所有海水都通过水族箱泵进行过滤处理。每天给贻贝样本提供螺旋藻藻类作为食物。经过连续暴露10 d后，共采集了160个样本，其中80个样本为非污染的贻贝样本，另外80个样本为受重金属铅污染的贻贝样本，用于进行光谱获取和后续的分析研究。在样本的采集过程中，确保样本的代表性和一致性，并严格按照试验设计将样本分组。

贻贝从海水中取出后立即打开，将其软组织部分置于壳的一侧。然后通过近红外光谱系统获取贻贝的光谱信息。该系统主要包括近红外光谱仪、卤素光源、Y型光纤、计算机和可移动平台，其中近红外光谱仪为台湾光兴电子股份有限公司的SW2520-050-NIRA型近红外光谱仪。在系统的辅助下，使用Spectramart软件收集光谱数据，并在950～1 700 nm的范围内测量贻贝反射光谱，总共有114个波长。此外，为降低噪声影响，预先进行了黑白矫正操作[18]。

将光谱采集后的贻贝根据GB 5009.268—2016《食品安全国家标准　食品中多元素的测定》中的等离子质谱法测定翡翠贻贝中铅的含量。

1.2　VIAVC波长选择方法

为了去除原始波段中不重要的或者不相关的特征波长，引入模型集群分析思想，达到降维目的。设翡翠贻贝光谱数据包含n个样本，m个波段。基于模型集群分析思想对每个变量进行重要性分析，即VIAVC [19-20]，主要包括4个步骤：

（1）采用二进制矩阵重采样（binary matrix resampling，BMR）产生大量随机的变量组合，可以保证每个变量以相同的概率被选择。BMR流程[20]如图1所示。

（2）采用偏最小二乘—线性判别分析（partial least squares linear discriminant analysis，PLS-LDA）对BMR数据子集建立子模型，使用接收器工作曲线（receiver operating characteristic，ROC）及曲线下面积（area under the curve，AUC）值作为分类器模型度量指标，对每个变量的重要性进行评估。以第j个变量为例，第j个变量的重要性评估是通过包含和排除同一个第j个变量建模对其进行比较来获得，而其他变量的状态保持不变。对于第j个变量，通过将$\boldsymbol{M}$的第j列中所有“1”改为“0”，所有

矩阵中,行表示抽样运行次数,列表示变量;1表示包含该变量用于建模,0表示不包含该变量

图1　BMR流程

Figure 1　BMR process diagram

“0”改为“1”,同时保持$\boldsymbol{M}$的其他列不变,来获得新矩阵$\boldsymbol{M}'$,变量状态变化过程如图1(b)和图1(c)。采用PLS-LDA的10折交叉验证(cross validation,CV)建模,根据公式生成$\phi_{\text{include},j}$和$\phi_{\text{exclude},j}$,评估第j个变量的重要性。

$$\phi_{\text{include},j}=\begin{cases}E_{\text{AUCCV},0}\text{的第}i\text{个分量} & \text{如果}M_{ij}=1\\ E_{\text{AUCCV},j}\text{的第}i\text{个分量} & \text{如果}M'_{ij}=1\end{cases},i=1,2,\cdots,Q, \tag{1}$$

$$\phi_{\text{exclude},j}=\begin{cases}E_{\text{AUCCV},0}\text{的第}i\text{个分量} & \text{如果}M_{ij}=0\\ E_{\text{AUCCV},j}\text{的第}i\text{个分量} & \text{如果}M'_{ij}=0\end{cases},j=1,2,\cdots,m, \tag{2}$$

式中:

M_{ij}——$\boldsymbol{M}$的第i行和第j列中的值;

M'_{ij}——$\boldsymbol{M}'$的第i行和第j列中的值;

$\phi_{\text{include},j}$——收集包含第j个变量建模的$E_{\text{AUCCV},0}$和$E_{\text{AUCCV},j}$的值;

$\phi_{\text{exclude},j}$——收集不包含第j个变量用于建模的$E_{\text{AUCCV},0}$和$E_{\text{AUCCV},j}$的值;

$E_{\text{AUCCV},0}$——对$\boldsymbol{M}$的所有行进行建模,每行获得AUC的CV值;

$E_{\text{AUCCV},j}$——对$\boldsymbol{M}'$的所有行进行建模,每行获得AUC的CV值;

i——抽样次数;

j——变量。

$\phi_{\text{include},j}$和$\phi_{\text{exclude},j}$都包含Q个模型的结果,对Q个子模型总体,根据$\phi_{\text{include},j}$和$\phi_{\text{exclude},j}$之间的两个分布来评估变量重要性。$\phi_{\text{include},j}$分布和$\phi_{\text{exclude},j}$分布平均值的差值为:

$$D_{\text{MEAN},j}=M_{\text{EAN,exclude}}-M_{\text{EAN,include}}, \tag{3}$$

式中:

$D_{\text{MEAN},j}$——$\phi_{\text{exclude},j}$和$\phi_{\text{include},j}$两个分布平均值的差值;

$M_{\text{EAN,exclude}}$——$\phi_{\text{exclude},j}$分布的平均值;

$M_{\text{EAN,include}}$——$\phi_{\text{include},j}$分布的平均值。

采用配对t检验和多重比较Bonferroni-Holm[21]校正来考察$\phi_{\text{include},j}$和$\phi_{\text{exclude},j}$的两种分布之间的差异是否显著,每个变量得出一个P值。P值越小,说明两种分布的差异越显著。预定义阈值为P=0.05,通过式(3)和假设检验将变量分为4类:强信息变量($D_{\text{MEAN},j}>0$,$P<0.05$)、弱信息变量($D_{\text{MEAN},j}>0$,$P>0.05$)、无信息变量($D_{\text{MEAN},j}<0$,$P>0.05$)和干扰变量($D_{\text{MEAN},j}<0$,$P<0.05$)。强信息变量会对建模产生有益的影响,且非常显著;弱信息变量引起了正差异,但不显著;干扰变量对建模有显著的不良影响;而非信息变量对建模也有不良影响,但不显著。因此,区分干扰性和非信息变量是十分必要的。

(3)迭代去除非信息和干扰变量,保留强信息和弱信息变量;最后,利用P值对保留的信息变量进行排序。

(4)将保留的排序后的变量依次按秩增加到子集中,直到变量都被包含,针对每个子集分别建立PLS-LDA模型,并用10折双交叉验证[22](DCV),检验各子集的预测性能。采用DCV可以克服新样本预测误差与模型参数优化之间的依赖性。根据序列增加得到的子集预测精度最好的,即为最佳变量子集。其流程图如图2所示。

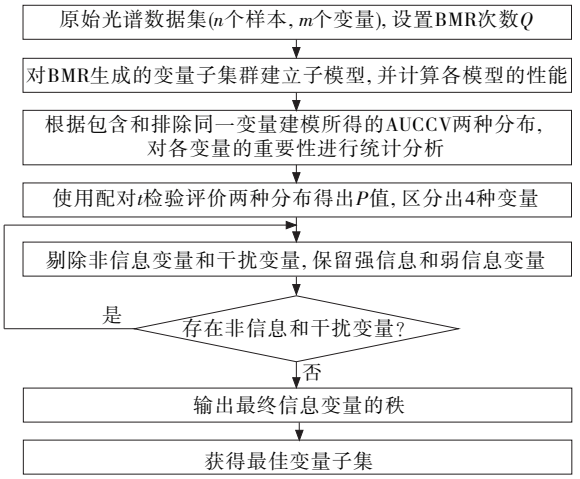


图 2　基于 VIAVC 算法流程图

Figure 2　Flow chart of VIAVC algorithm

1.3　GFRNN 模型的分类思想

GFRNN 法将万有引力思想引入近邻分类策略中，是一种有效的处理数据不平衡问题的机器学习算法[15]。

首先，假设存在一个二元类不平衡数据集，已知输入为训练集样本 X_{All}，测试集样本 Y_{All}，输出为预测集 Y_{All} 的类标签 $\tilde{\varphi}$。将正类(positive class，PC)即少数类训练集定义为 $X_{\mathrm{PC}}=\{(x_1,\varphi_1),(x_2,\varphi_1),\cdots,(x_{n_{\mathrm{PC}}},\varphi_1)\}$，将负类(negative class，NC)即多数类训练集定义为 $X_{\mathrm{NC}}=\{(x_{n_{\mathrm{PC}}+1},\varphi_2),(x_{n_{\mathrm{PC}}+2},\varphi_2),\cdots,(x_{n_{\mathrm{PC}}+n_{\mathrm{NC}}},\varphi_2)\}$，其中 $\varphi_i\in[\{-1,1\}]$表示训练集样本的类标签，1 表示 PC，−1 表示 NC。n_{PC} 和 n_{NC} 分别表示训练集中正类和负类样本的数目。训练集样本总数量可以写成：$n_{\mathrm{All}}=n_{\mathrm{PC}}+n_{\mathrm{NC}}$。

在式(4)中定义了距离测量函数 $d(\cdot)$，用于测量两个样本之间的距离，如：

$$d(x_{\mathrm{p}},x_{\mathrm{q}})=\|x_{\mathrm{p}}-x_{\mathrm{q}}\|_2, \tag{4}$$

式中：

x_{p}，x_{q}——训练样本；

$d(x_{\mathrm{p}},x_{\mathrm{q}})$——样本 x_{p} 到 x_{q} 的距离。

然后，通过 FRNN 算法来寻找在给定半径 R 内与测试样本 y 近邻的候选者集合 x_{candi}，FRNN[23]定义如式(5)所示。

$$F_{\mathrm{RNN}}(y,X_{\mathrm{All}},R)=\{x_{\mathrm{p}}\in X_{\mathrm{All}},d(y,x_{\mathrm{p}})<R\}, \tag{5}$$

式中：

$F_{\mathrm{RNN}}(y,X_{\mathrm{All}},R)$——给定半径 R 内训练集样本与测试样本 y 近邻的候选者集合；

y——测试样本(表示待搜索的点)；

x_{p}——训练样本；

X_{All}——训练集样本；

$d(y,x_{\mathrm{p}})$——测试样本 y 到训练样本 x_{p} 的距离；

R——固定的半径值(通常用常数表示)。

R 影响 FRNN 搜索的范围，半径越大搜索的样本越多，半径越小则搜索的样本数量越少。确定半径 R 是获得候选者集合 x_{candi} 的关键问题。参照文献[24]，将训练样本对之间的均方欧氏距离作为 R 的取值，计算如式(6)所示。

$$R=\frac{1}{2}n_{\mathrm{All}}(n_{\mathrm{All}}-1)\sum_{x_{\mathrm{p}},x_{\mathrm{q}}\in X_{\mathrm{All}}}d(x_{\mathrm{p}},x_{\mathrm{q}}), \tag{6}$$

式中：

n_{All}——训练集样本的总数量；

x_{p}，x_{q}——训练样本；

X_{All}——训练集样本；

$d(x_{\mathrm{p}},x_{\mathrm{q}})$——样本 x_{p} 到 x_{q} 的距离。

在获得候选集合 x_{candi} 后，将样本集中的每个样本视为具有相应质量的实体，基于万有引力的方法计算候选集中每个样本对测试样本的万有引力。式(7)定义了万有引力函数 $D(\cdot)$，用于计算测试样本 y 与候选集样本 x_i 之间的万有引力。

$$D(y,x_i)=G\frac{m_y m_{x_i}}{d(y,x_i)^2}, \tag{7}$$

式中：

$D(y,x_i)$——测试样本 y 与候选集样本 x_i 之间的万有引力；

G——万有引力系数；

m_y——测试样本 y 的质量；

m_{x_i}——候选集样本 x_i 的质量；

$d(y,x_i)$——测试样本 y 到候选集样本 x_i 的距离。

由于参数 G 和 m_y 对最终分类结果没有影响，为了简化式(7)，均设为 1。为了使模型更好地应用于不平衡数据分类，减轻样本不平衡性对分类结果的影响，考虑用 X_{All} 的不平衡比例来表示候选样本之间的平衡，将训练样本的质量 m_{x_i} 根据不平衡率设置如下：

$$m_{x_i}=\begin{cases}\dfrac{n_{\mathrm{NC}}}{n_{\mathrm{PC}}}, & x_i\in X_{\mathrm{PC}}\cap X_{\mathrm{candi}}\\ 1, & x_i\in X_{\mathrm{NC}}\cap X_{\mathrm{candi}}\end{cases}, \tag{8}$$

式中：

n_{NC}——训练集中负类样本的数目；

n_{PC}——训练集中正类样本的数目；

m_{x_i}——训练样本 x_i 的质量；

X_{PC}——正类样本集；

X_{NC}——负类样本集；

X_{candi}——候选集样本。

最后，通过函数 $F(\cdot)$ 计算作用在测试样本 y 上万有引力的合力，此处不考虑万有引力的方向，只是简单地将所有候选样本都位于一条直线上，并将样本的每个特征的

权重视为1。合力大小计算如式(9)所示。

$$F(y)=\sum_{i=1,2,x_i\in X_{\text{candi}}}\varphi_i D(y,x_i), \tag{9}$$

式中:

$F(y)$——测试样本y上万有引力的合力;

$D(y,x_i)$——测试样本y与候选集样本x_i之间的万有引力;

φ_i——训练集样本的类标签;

X_{candi}——候选集样本。

计算出每一类样本对测试样本的万有引力合力之后,根据合力的大小进行分类。最终分类结果偏向合力较大的一类。如果$F(y)<0$,说明候选集中负类样本对测试样本的万有引力合力最大,测试样本属于负类;而$F(y)>0$,则测试样本属于正类。算法具体流程:

GFRNN算法:

输入:训练集样本X_{All},测试集样本Y_{All}。

输出:预测集Y_{All}的类标签$\tilde{\varphi}$。

Step1:设置一个合适的半径R。根据式(6)得到X_{All}的固定半径R。

Step2:选择候选者样本。根据式(5),通过FRNN获取候选集合x_{candi}。

Step3:对于测试集Y_{All}中的每一个样本:

3-1:由式(7)和式(8)计算当前测试样本$y\in Y_{\text{All}}$与来自x_{candi}的每个候选样本之间的引力;

3-2:根据式(9)计算万有引力的合力$F(y)$;

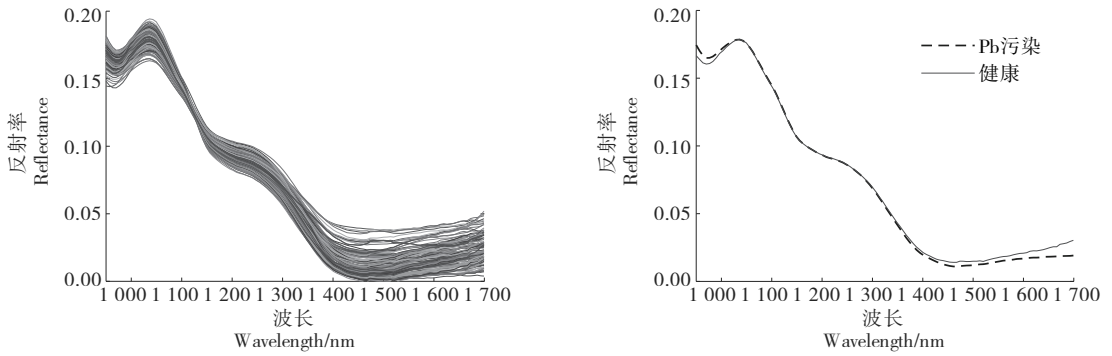
3-3:如果$F(y)<0$,则$\tilde{\varphi}=-1$;否则,$\tilde{\varphi}=1$。

Step4:输出$\tilde{\varphi}$。

2　分析与讨论

2.1　贻贝光谱数据处理与分析

图3给出了受铅污染和健康的贻贝样本的光谱曲线以及平均光谱曲线。图3(a)揭示了160个样品的近红外光谱,其中80个是受铅污染的样品,另外80个则为未受污染的样品。因为这些样品都源自同一物种,其光谱表现出高度相似的吸收带和形状。然而,观察图3(b),可以明显看出,铅污染样品和非污染样品的平均光谱在950~1 130 nm的反射率值有所不同,铅污染样品的反射率略高。进一步来看,在1 380~1 700 nm的区间,可以发现未受污染样品的反射率值超过了铅污染样品。在其他的波长范围,两种样品的平均光谱曲线几乎可以看作是重叠的,且均在1 130 nm附近有明显的吸收峰。受铅污染贻贝的主要吸收峰位于1 140,1 470,1 610 nm处。



(a) 原光谱　　(b) 平均光谱

图3　受铅污染和健康的贻贝样本的光谱曲线

Figure 3　Spectrum curves of Pb-contaminated and healthy mussel samples

铅的污染会对贻贝的生理和代谢产生负面影响,可能会导致蛋白质、酶、DNA以及其他重要生物分子的结构发生变化,这些变化将会反映在光谱曲线上。因此,这些光谱的差异为区分铅污染样品和非污染样品提供了可能。然而,并非所有的光谱差异都容易理解和解析,特别是当一些特定的化学成分的波长与其他重要的背景信息相重叠时,这一任务就会变得困难。然而,即便面临这样的挑战,依然可以借助于这些光谱差异来探索贻贝受铅污染的影响,从而更深入地理解和识别铅污染。

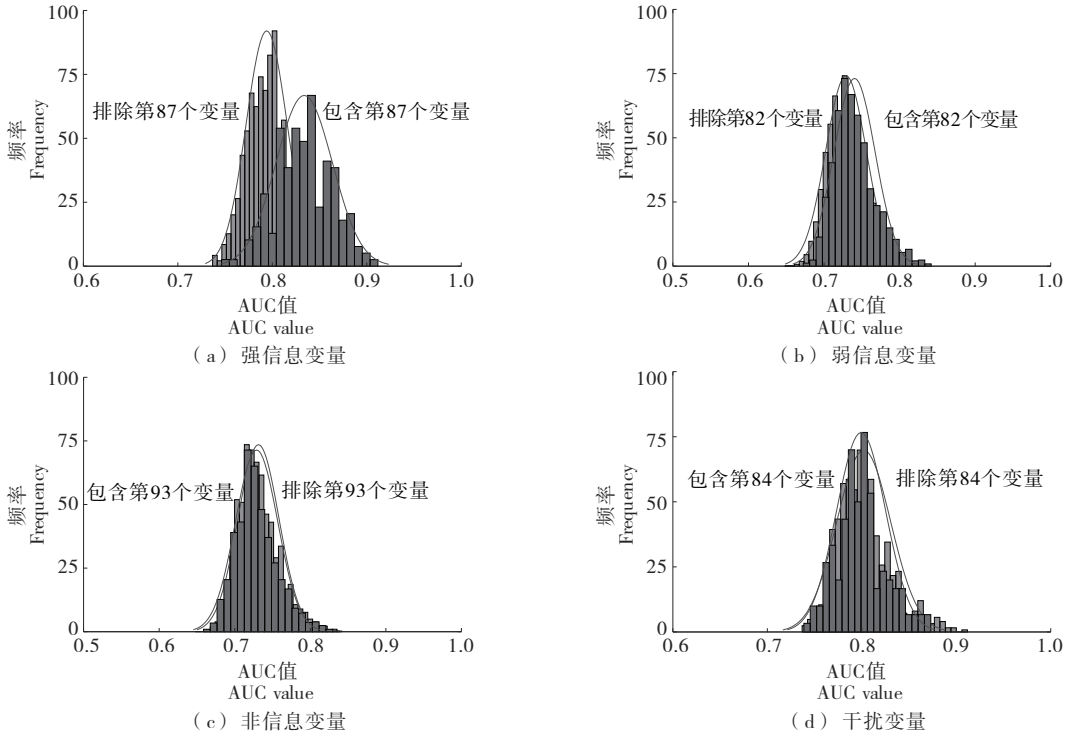
在近红外光谱系统工作过程中,受环境条件和样本结构特性变化的影响,光谱基线漂移、随机噪声和多重散射效应等问题常常会出现。为了有效降低这些影响,通常采用预处理技术来处理光谱数据。研究中,采用了Savitzky-Golay平滑结合自适应迭代惩罚最小二乘(adaptive iteratively reweighted penalized least squares, AIRPLS)方法,以提高光谱的信噪比,减少噪音干扰,并突出光谱中的特征,以便进行更精确的光谱分析。在评估预处理方法的实际效果时,选择分类性能作为衡量指标。通过对预处理后的光谱数据进行分类分析,提高模型的准确性和可靠性。

2.2　波段选择试验结果与分析

翡翠贻贝近红外光谱数据维数为114,高维数据带来

了较大的计算资源浪费，模型计算识别速度较慢，不利于实际应用。采用VIAVC对光谱数据进行降维，对算法控制参数进行大量试验测试，采样次数Q=1 000，交互检验次数为10。图4给出了VIAVC运行第一轮迭代的4种变量（深灰色分布表示包含第j个变量，浅灰色分布表示排除第j个变量），其中第87个波长变量（$D_{MEAN}>0$且$P=5.283\,9\times10^{-13}\ll0.05$）被认为是强信息变量，第82个变量（$D_{MEAN}>0$且$P=0.475\,1>0.05$）为弱信息变量，第93个变量（$D_{MEAN}<0$且$P=1.010\,7>0.05$）是非信息变量，第84个变量（$D_{MEAN}<0$且$P=1.686\,7\times10^{-9}\ll0.05$）被归类为干扰变量。经过7次迭代后，VIAVC保留了15个变量，5次迭代中的变量数分别为60，37，29，23，20，18，15。这些变量是根据包含和排除对应变量得到的两种分布之间差异的P值进行排名的，其他99个变量被定义为非信息变量或干扰变量。

为了获得最佳变量子集，根据保留的15个信息变量的秩，利用DCV方式按照排序向前选择变量来建立子模型，并考察每个子集的预测结果如图5所示。由图5可以看出，当变量数为6时，预测精度最高的6个变量为最佳变量子集。利用前6个变量构建PLS-LDA模型的交叉验证准确率为97.5%，敏感性为1，特异性为0.933 3，AUC值为0.954 4。



（a）强信息变量　（b）弱信息变量　（c）非信息变量　（d）干扰变量

深灰色柱形表示包含第j个变量分布，浅灰色柱形表示排除第j个变量分布

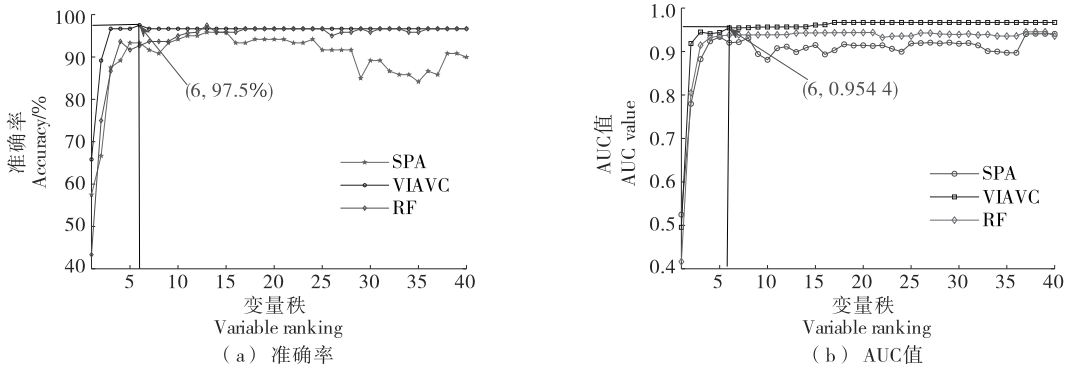
图4　样本应用VIAVC算法的4种变量预测误差分布

Figure 4　Illustrates the distribution of variable prediction errors in four scenarios when VIAVC algorithm is applied to samples

研究还应用子窗口排列分析法（subwindow permutation analysis，SPA）[25]和随机蛙跳法（random frog，RF）[26]对变量进行排序，并与VIAVC法进行比较。结果表明，不同方法会生成不同的变量排序，且排序差异较大。根据SPA和RF方法的变量排序，采用PLS-LDA逐步添加顶变量的方式进行建模，利用DCV来检验各子集的预测性能，结果如图5(a)和图5(b)所示。可以看出，VIAVC和RF算法分别选择了6和13个变量作为最优子集，预测准确率最高为97.50%。两者具有相同的预测精度，但VIAVC的AUC值0.954 4优于RF的AUC值0.942 5，且RF比VIAVC选择了更多的变量。因此，VIAVC优于RF。对于SPA方法选择了前37个变量为最佳子集，预测精度为95.83%，AUC值为0.941 0。从预测准确率和AUC值可以看出，VIAVC对信息变量的识别能力更强。

2.3　GFRNN模型检测性能分析

为了评估GFRNN模型检测重金属铅污染贻贝的性能，将GFRNN模型与FRNN算法[23]和KNN算法[27]进行比较（最近邻系数K设置为5），另外还与经典的SVM算法[28]进行了比较。SVM模型采用径向基核函数，其参数C的取值范围为$\{10^{-2},10^{-1},0,10^{1},10^{2}\}$，$\gamma$从集合



（a）准确率　　（b）AUC值

图5　按照降序选择变量集合的准确率和AUC值曲线

Figure 5　Accuracy and AUC values of variable sets selected in descending order

$\{10^{-3},10^{-2},10^{-1},0,10^{1},10^{2},10^{3}\}$中选择。

使用VIAVC算法对光谱数据进行降维,其中设置Q参数为1 000,得到了6个波长变量分别为1 522.54,1 672.47,1 516.0,1 079.59,1 476.69,1 052.85 nm。将选取的波长变量代入到GFRNN、FRNN、KNN和SVM模型中建立分类模型,结果如表1所示。试验中,贻贝训练集和测试集样本数比例为6:2。

表1　不同分类模型的性能比较

Table 1　Performance comparison based on different classifiers

分类模型	准确率/%	几何均值	AUC
GFRNN	99.17	0.991 7	0.988 6
FRNN	97.50	0.975 1	0.933 3
KNN	91.67	0.922 4	0.921 9
SVM($C=1,\gamma=1$)	97.50	0.975 9	0.946 2

从表1可以看出,4种模型对铅污染贻贝的识别效果都较好,分类准确率均>91.67%,其中GFRNN在贝类重金属铅污染数据集上的性能表现更加优异,FRNN和SVM模型次之,KNN模型最差。结果表明,与FRNN、KNN和SVM分类器相比,利用GFRNN模型在分类精度、G-mean和AUC值上具有较强的竞争力,其模型分类准确率为99.17%,G-mean为0.991 7,AUC值为0.988 6。

通常情况下,受重金属铅污染的贻贝数量相对较少,导致采集的样本中健康贻贝和受重金属铅污染贻贝的数量呈不平衡状态。为了比较所提出的VIAVC结合GFRNN模型与其他模型在重金属铅污染贻贝检测方面的性能,对重金属铅污染贻贝数据集在不同不平衡比例下进行了试验。

表2~表4分别给出了GFRNN与其他对比模型的检测准确率、G-mean和AUC值随训练集中健康贻贝和重金属铅污染贻贝的数量比例变化的试验结果。其中,训练集包含健康贻贝样本60个,而铅污染样本的数量从50个减少到10个,每次减少10个。从表2~表4可以看出,4种分类模型的准确率、G-mean和AUC值均随着铅污染样本数量的减少而下降,且均是开始下降幅度非常小,然后逐渐增大。VIAVC-GFRNN模型的准确率最高,其分类准确率变化最小。当铅污染训练样本数量减少到40时,各分类模型检测性能降低不太明显;当减小到20个时,VIAVC-GFRNN模型的准确率能达到96.25%,VIAVC-SVM模型的准确率减小到85%;当只有10个时,VIAVC-GFRNN模型的准确率达到94.29%,其他模型的准确率低于85%,VIAVC-KNN模型的表现最差。结果表明,与其他3类模型相比,VIAVC-GFRNN模型的分类准确率受类别不平衡的数据集影响较小,具有较好的稳健性。因此,利用VIAVC-GFRNN模型在铅污染贻贝不平衡数据集上检测是有效的,将该方法应用于大规模贝类重金属污染检测识别是可行的。

表2　健康贻贝和铅污染贻贝在不同数量比例下的检测准确率

Table 2　Demonstrates the detection accuracy of healthy mussels and Pb-contaminated mussels at different ratios of quantity　%

分类模型	60:50	60:40	60:30	60:20	60:10
GFRNN	99.09	98.00	96.67	96.25	94.29
FRNN	97.27	94.00	85.56	83.75	78.57
KNN	90.91	89.00	83.33	77.50	72.86
SVM	96.36	95.00	91.11	85.00	81.43

3　结论

研究开发了一个将近红外光谱与VIAVC-GFRNN法相结合的模型以识别重金属铅污染的贻贝。首先,确定VIAVC算法选择的特征波段(6个)可以有效区分健康贻贝和铅污染的贻贝。其次,研究提出将GFRNN模型用于解决重金属污染不平衡数据集检测的问题,以准确率、G-mean和AUC值作为性能指标,GFRNN模型的性能优

表 3　健康贻贝和铅污染贻贝在不同数量比例下的 G-mean 值

Table 3　Presents the G-mean values of healthy mussels and Pb-contaminated mussels at different quantity ratios

分类模型	60∶50	60∶40	60∶30	60∶20	60∶10
GFRNN	0.990 1	0.979 2	0.953 5	0.943 1	0.858 5
FRNN	0.969 5	0.934 6	0.832 4	0.770 7	0.632 5
KNN	0.912 9	0.883 4	0.812 4	0.705 7	0.587 2
SVM	0.962 3	0.944 0	0.891 1	0.786 2	0.659 4

表 4　健康贻贝和铅污染贻贝在不同数量比例下的 AUC 值

Table 4　Shows the AUC values of healthy mussels and Pb-contaminated mussels at different quantity ratios

分类模型	60∶50	60∶40	60∶30	60∶20	60∶10
GFRNN	0.988 1	0.972 0	0.960 5	0.942 5	0.852 6
FRNN	0.942 4	0.931 9	0.829 7	0.789 3	0.630 5
KNN	0.912 2	0.863 4	0.806 9	0.701 1	0.579 1
SVM	0.948 3	0.936 9	0.885 4	0.784 6	0.661 2

于 FRNN、KNN 和 SVM 算法。最后，对 VIAVC-GFRNN 模型在不同比例不平衡数据集中的检测性能进行了分析。试验结果表明，VIAVC-GFRNN 模型实现了更好的分类性能，且对健康贻贝和铅污染贻贝的比例不敏感。利用近红外光谱与 VIAVC-GFRNN 方法相结合来检测贻贝中重金属污染是可行的，在未来的研究中，可以设计更有效的计算测试样本与训练集中每个样本距离的策略，来改进算法以提高效率。NIRS 与模式识别方法相结合，有望在贝类和其他海鲜的监测中发挥重要作用。

参考文献

[1] 张舒玄, 卢海燕, 李优琴, 等 . 农产品中重金属的检测方法研究进展[J]. 理化检验(化学分册), 2019, 55(8): 976-983.
ZHANG S X, LU H Y, LI Y Q, et al. Recent advances of researches on detection methods of heavy metals in agricultural products[J]. Physical Testing and Chemical Analysis Part B (Chemical Analysis), 2019, 55(8): 976-983.

[2] 李万杰, 马春, 张新欣, 等 . 微波消解—石墨炉原子吸收法检测海产品中痕量重金属[J]. 大连工业大学学报, 2015, 34(2): 111-113.
LI W J, MA C, ZHANG X X, et al. Determination of trace heavy metal in marine products by microware digestion-graphite furnace atomic absorption spectrometry[J]. Journal of Dalian Polytechnic University, 2015, 34(2): 111-113.

[3] YAP C K, ISMAIL A, TAN S G, et al. Assessment of different soft tissues of the green-lipped mussel Perna viridis (Linnaeus) as biomonitoring agents of Pb: field and laboratory studies[J]. Water Air Soil Pollut, 2004, 153(1): 253-268.

[4] 孙玲玲, 宋金明, 刘瑶, 等 . 四极杆碰撞反应池 ICP-MS 同时测定贻贝中的 Mo 等 12 种重金属[J]. 海洋环境科学, 2020, 39(3): 453-459.
SUN L L, SONG J M, LIU Y, et al. Simultaneous determination of molybdenum and other heavy metals in Mytilus edulis by inductively coupled plasma mass spectrometry with quadrupole collision cell technology[J]. Marine Environmental Science, 2020, 39(3): 453-459.

[5] 张欣欣, 李尚科, 李跑, 等 . 近红外漫反射光对水果的穿透能力研究[J]. 中国食品学报, 2022, 22(1): 298-305.
ZHANG X X, LI S K, LI P, et al. Studies on the penetration ability of near infrared diffuse light on fruits[J]. Journal of Chinese Institute of Food Science and Technology, 2022, 22(1): 298-305.

[6] 王海华, 李长缨, 李民赞 . 基于近红外反射光谱的洋葱可溶性固体物检测[J]. 光谱学与光谱分析, 2013, 33(9): 2 403-2 406.
WANG H H, LI C Y, LI M Z. Detection of onion soluble solids content based on the near-infrared reflectance spectra[J]. Spectroscopy and Spectral Analysis, 2013, 33(9): 2 403-2 406.

[7] 黄明月, 吴海云, 靳皓, 等 . 基于近红外透射—漫反射光谱掺杂牛奶判别[J]. 光谱学与光谱分析, 2020, 40(S1): 85-86.
HUANG M Y, WU H Y, JIN H, et al. Discrimination of adulterated milk based on near infrared transmission and diffuse reflectance spectroscopy[J]. Spectroscopy and Spectral Analysis, 2020, 40(S1): 85-86.

[8] 吕程序, 姜训鹏, 张银桥, 等 . 基于变量选择的小麦粗蛋白含量近红外光谱检测[J]. 农业机械学报, 2016, 47(S1): 340-346.
LU C X, JIANG X P, ZHANG Y Q, et al. Variable selection based near infrared spectroscopic quantitative analysis on wheat crude protein content[J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(S1): 340-346.

[9] 姜微, 刘瑶, 刘忠艳, 等 . 间隔影响分析波长选择算法在近红外光谱鉴别贝类毒素中的应用[J]. 食品与发酵工业, 2023, 49(2): 271-279.
JIANG W, LIU Y, LIU Z Y, et al. Application of margin influence analysis wavelength selection algorithm in the identification of shellfish toxins by near infrared spectroscopy [J]. Food and Fermentation Industries, 2023, 49(2): 271-279.

[10] 唐鸣, 田潇瑜, 王旭, 等 . 基于近红外特征波段的注水肉识别模型研究[J]. 农业机械学报, 2018, 49(S1): 440-446.
TANG M, TIAN X Y, WANG X, et al. Recognition model of water-injected meat based on characteristic spectrum extraction of infrared spectroscopy[J]. Transactions of the Chinese society for Agricultural Machinery, 2018, 49(S1): 440-446.

[11] GARCÍA-MARTÍN J F, BADARÓ A T, BARBIN D F, et al. Identification of copper in stems and roots of *Jatropha curcas* L. by hyperspectral imaging[J]. Processes, 2020, 8(7): 822-831.

[12] CHEN Y N, SUN D W, CHENG J H, et al. Recent advances for rapid identification of chemical information of muscle foods by hyperspectral imaging analysis[J]. Food Eng Rev, 2016, 8(3): 336-350.

[13] LIU Y, XU L L, ZENG S G, et al. Rapid detection of mussels contaminated by heavy metals using near-infrared refectance spectroscopy and a constrained diference extreme learning machine[J]. Spectrochimica Acta Part A, 2022, 269: 120776.

[14] 曾绍庚, 刘瑶, 刘忠艳. 基于近红外光谱技术和LSPTSVM模型的镉污染贻贝检测研究[J]. 环境工程, 2024, 42(1): 235-242.
ZENG S G, LIU Y, LIU Z Y. Detection of mussels contaminated with cadmium based on near-infrared spectroscopy and lsptsvm[J]. Environmental Engineering, 2024, 42(1): 235-242.

[15] ZHU Y J, WANG Z, GAO D Q. Gravitational fixed radius nearest neighbor for imbalanced problem[J]. Knowledge-Based Systems, 2015, 90: 224-238.

[16] 程波, 彭嘉琪, 王新月, 等. 中国水产品质量安全标准体系现状研究[J]. 中国渔业质量与标准, 2023, 13(5): 53-66.
CHENG B, PENG J Q, WANG X Y, et al. Research on the current situation of China's aquatic product quality and safety standard system[J]. Chinese Fishery Quality and Standards, 2023, 13(5): 53-66.

[17] 葛奇伟. 养殖贝类重金属特征污染物的筛选及其风险评价[D]. 宁波: 宁波大学, 2013: 19-23.
GE Q W. Studies on characteristic selection of heavy metal pollutants for cultured molluscs and their risk assessment[D]. Ningbo: Ningbo University, 2013: 19-23.

[18] 刘莉, 陶红燕, 方静, 等. 基于近红外高光谱的梨叶片炭疽病与黑斑病识别[J]. 农业机械学报, 2022, 53(2): 221-230.
LIU L, TAO H Y, FANG J, et al. Identifying anthracnose and black spot of pear leaves on near-infrared hyperspectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(2): 221-230.

[19] 孙俊, 张跃春, 毛罕平, 等. 基于计算机视觉的土壤镉胁迫生菜叶片污染响应分析[J]. 农业机械学报, 2018, 49(3): 166-172.
SUN J, ZHANG Y C, MAO H P, et al. Responses analysis of lettuce leaf pollution in cadmium stress based on computer vision[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(3): 166-172.

[20] YUN Y H, FU L, DENG B C, et al. Erratum to: Informative metabolites identification by variable importance analysis based on random variable combination[J]. Metabolomics, 2016, 12(2): 1-13.

[21] HOLM S. A simple sequentially rejective multiple test procedure[J]. Scandinavian Journal of Statistics, 1979, 6(2): 65-70.

[22] WESTERHUIS J A, HOEFSLOOT H C, SMIT S, et al. Assessment of PLSDA cross validation[J]. Metabolomics, 2008, 4(1): 81-89.

[23] 周鹏, 伊静, 朱振方, 等. 面向不平衡分类的固定半径最近邻逐步竞争算法(FRNNPC) [J]. 山东大学学报(理学版), 2019, 54(3): 102-109.
ZHOU P, YIN J, ZHU Z F, et al. Fixed-radius nearest neighbor progressive competition algorithm for imbalanced classification[J]. Journal of Shandong University (Natural Science), 2019, 54(3): 102-109.

[24] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6 191): 1 492-1 496.

[25] 孙通, 吴宜青, 李晓珍, 等. 基于近红外光谱和子窗口重排分析的山茶油掺假检测[J]. 光学学报, 2015, 35(6): 1-8.
SUN T, WU Y Q, LI X Z, et al. Discrimination of camellia oil adulteration by nir spectra and subwindow permutation analysis[J]. Acta Optica Sinica, 2015, 35(6): 1-8.

[26] 郑剑, 周竹, 仲山民, 等. 基于近红外光谱与随机青蛙算法的褐变板栗识别[J]. 浙江农林大学学报, 2016, 33(2): 322-329.
ZHENG J, ZHOU Z, ZHONG S M, et al. Chestnut browning detected with near-infrared spectroscopy and a random-frog algorithm[J]. Journal of Zhejiang A & F University, 2016, 33 (2): 322-329.

[27] COVER T M, HART P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[28] COEN T, SAEYS W, RAMON H, et al. Optimizing the tuning parameters of least square support vector machines regression of NIR spectra[J]. Journal of Chemometrics, 2006, 20: 184-192.