大米拉曼光谱不同预处理方法的相近产地鉴别研究

王亚轩[1]，谭　峰[2*]，辛元明[2]，李　欢[2]，赵肖宇[2]，鹿保鑫[3]

1. 黑龙江八一农垦大学土木水利学院，黑龙江 大庆　163319
2. 黑龙江八一农垦大学电气与信息学院，黑龙江 大庆　163319
3. 黑龙江八一农垦大学食品学院，黑龙江 大庆　163319

摘　要　用相近产地的大米代替独有的地理因素形成的地域品牌大米，消费者难以辨别。基于拉曼光谱技术，试验对比不同预处理方法包括一阶导数＋平移平滑、二阶导数＋平移平滑，小波变换＋去除基线三种常用的预处理方法，另外提出一种改进的分段多项式拟合＋去除基线共四种预处理方法，分别结合偏最小二乘法实现相近产地大米的鉴别分析，提出一种最佳的鉴别相近产地大米的预处理方法。首先用拉曼光谱仪采集了黑龙江省依安县3个相近产地大米的150个拉曼位移为200～3 100 cm^{-1}的大米光谱样本，再对原始拉曼光谱分别用一阶导数＋平移平滑、二阶导数＋平移平滑、小波变换＋去除基线、分段多项式拟合＋去除基线进行光谱预处理。分别从每个产地选取33个样本进行训练，并对未知的51个样本建立了基于偏最小二乘法的鉴别分析模型，在训练集中一阶导数＋平移平滑的预处理方法相关系数值最大、均方误差和均方根误差最小，小波变换＋去除基线的预处理方法相关系数值最小、均方误差和均方根误差最大；在测试集中采用3点2次拟合＋去除基线的预处理方法的相关系数值最大、均方误差和均方根误差最小，二阶导数＋平移平滑的预处理方法最差。最后再通过PLS建模结果得知，在训练集中，采用四类九种预处理的方法对三个产地大米的总识别率均为100％；在测试集中，采用3点2次拟合＋去除基线对三个产地大米总识别率为100％，采用5点2次拟合＋去除基线对三个产地大米总识别率为52.9％，其他分段多项式拟合介于二者之间；采用一阶导数＋平移平滑、二阶导数＋平移平滑和小波变换的总识别率分别为88.2％，86.2％和96.1％；从中发现，分段式多项式拟合中的3点2次拟合＋去除基线的优势明显，与其相关系数、均方误差、均方根误差结果吻合，总体识别率高，鉴别效果稳定。

关键词　拉曼光谱；基线去除；大米；预处理方法

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)02-0565-07

引　言

大米是我国主要的主食来源[1-2]，全国大米种植区域广、种类多，土壤、环境和水质等差异形成地域因素会导致大米的品质发生变化。如五常大米、牡丹江响水大米，独特的地理环境形成特有的口感和营养价值，使其成为具有鲜明地理标识的大米产品。但一些商家为了追求更高的利润，用相近产地的大米代替地域品牌大米，购买时仅通过消费者肉眼判断很难区分，这不仅损害了粮农的利益，也不利于品牌产业链的健康发展。因此，研究相近产地大米的快速准确无损鉴别的方法能为鉴别地理标识大米提供理论和技术支持。

拉曼光谱通过物质内部分子对可见单色光的散射强度不同来识别分子结构，从而对物质内部官能团进行特定指纹标定，光谱谱峰强度与分子浓度有关。目前已广泛应用在食品、药材、化工、宝石等多个领域进行定性或定量的检测。拉曼光谱用于农产品检测方面，主要集中在对粮食、奶制品、果蔬类、食用油等的研究上，通过拉曼光谱分析产品内部是否掺杂其他物质，进行农产品质量和年份的鉴定。拉曼光谱应用于大米检测方面[3-6]，主流做法是通过光谱采集样本的原始特征光谱，再去除荧光和噪声，将样本分为训练集和测试集，结合主成分分析和偏最小二乘法进行数学模型的

收稿日期：2019-12-25，**修订日期**：2020-04-22

基金项目：黑龙江省自然科学基金重点项目(ZD2019F002)，中国博士后基金面上项目(2017M620123)，中央引导地方项目(ZY18B01)，黑龙江八一农垦大学三横三纵支持计划(TDJH201907)，黑龙江八一农垦大学人才科研启动计划(XDB2013-18)资助

作者简介：王亚轩，1980年生，黑龙江八一农垦大学土木水利学院讲师　　e-mail：wangyaxuan1980@163.com

＊通讯作者　　e-mail：tf1972@163.com

建立，来判别大米的产地、品种、新陈度等指标。黄嘉荣[3]等对广东大米、东北大米及糯米进行分类，识别率是97.9%，孙娟[4]等对大米进行产地分类，选择黑龙江大米、江苏大米、湖南大米三地大米识别率为94%以上。赵迎[5]等对储存三年以上和当年大米进行新陈大米进行分类，识别率为95%。从以上分析可以看出，研究主要是集中在不同品种大米的种类区分、对南方和北方产地大米的产地区分、不同年份大米的新陈度区分，而基于相近产地对大米进行分类鲜有研究。因为光谱鉴别不可避免要引入机器的噪声和荧光背景等干扰因素，因地域相近大米内部的淀粉、糖类等物质含量差异不大，从光谱中提取这些结构特征性片段难度很大，需要通过有效的预处理算法去除干扰，提取真实准确的拉曼特征峰。本文研究比较四类九种不同的预处理方法结合偏最小二乘法建模，提出一种鉴别相近产地大米的预处理方法，为大米产地鉴别提供新的理论依据。

1 实验部分

1.1 仪器与软件

光谱采集使用厦门奥谱天成光电有限公司制造的波长785 nm便携式拉曼光谱仪1台，检测范围在124.79～3 324.66 cm^{-1}，在最佳测量条件下，测量标准峰的位移值偏差为零，符合位移准确度不超过±4 cm^{-1}的使用要求；大米脱壳采用上海超星LJJM精米机1台，脱壳率≥99%，工作电压220 V，试验用量50～170 g；数据处理软件为matlab2010b。

1.2 样本

大米样本于2018年11月采自黑龙江依安县田间，分别是富饶乡黎明村(北纬47.389 49、东经125.406 00)、新兴镇东莱村(北纬47.752 09、东经125.187 28)、上游乡红五月村(北纬47.933 883、东经125.322 755)相同品种的粳米，依次用A，B和C表示上述的三个产地大米，每个产地均随机选取50个脱壳后的表面完好的大米作为试验样本，3个产地共计150个样本。其中选择每个产地样本数的2/3即33个样本用作训练集，剩余的1/3即17个样本用作测试集，共计51个样本用于测试。

1.3 光谱的获取

将从田间采集的带壳稻米装入尼龙网兜，在实验室晾晒10 d后，采用统一加工对其用精米机进行两次脱壳、每次脱壳50 s，再用100目筛子过筛，筛选出其中表面光滑完整的大米胚乳(去除胚芽)作为样本。拉曼光谱检测参数设置为：激光功率300 mW，激发波长785 nm，分辨率为6.58 cm^{-1}，积分时间为5 000 ms，扫描范围为200～3 300 cm^{-1}的波段，测试条件为室温，相对湿度为55%。每个样本选择米粒中间区域的背部或腹部采集数据，连续进行4次采集，取其平均值作为每个样本的存储数据。

1.4 光谱的预处理方法

光谱中普遍存在着荧光和背景噪声，仅靠仪器的精度和准确度来消除检测干扰受到仪器自身的限制，需要结合数学处理原始光谱数据来去除噪声和基线漂移，常用的方法有导数处理、平移平滑、多项式拟合、归一化等。导数处理主要是扣除仪器背景或漂移(散射)对信号的影响；平移平滑、多项式拟合能够非常有效的提高谱图信噪比，降低随机噪声的影响；归一化可以消除尺度差异过大带来的不良影响。

用大米样本的原始光谱进行数据分析时，虽然可以用现有方法进行光谱数据预处理[7]，但其精度和准确度都达不到近地大米光谱鉴别的要求，试验对比四类九种不同预处理方法进行原始数据分析，包括一阶导数＋平移平滑、二阶导数＋平移平滑、小波变换＋去除基线三种常用的预处理方法，另外提出一种改进的分段多项式拟合＋去除基线共四种预处理方法进行平滑去噪和去除基线漂移，再用极差归一的方法进行单位统一，预处理后的数据分别采用偏最小二乘(pratial least squares，PLS)方法[8-9]进行建模分析，旨在探寻研究一种适合近地大米光谱的预处理方法。

2 结果与讨论

2.1 三个产地的大米原始拉曼光谱

不同产地大米的营养成分基本一致，但各自的含量差异导致强度不同。图1所示为200～3 300 cm^{-1}范围内三个产地的典型大米原始拉曼光谱，可见不同产地的大米峰值强度不同，但产生峰值位置基本相同。

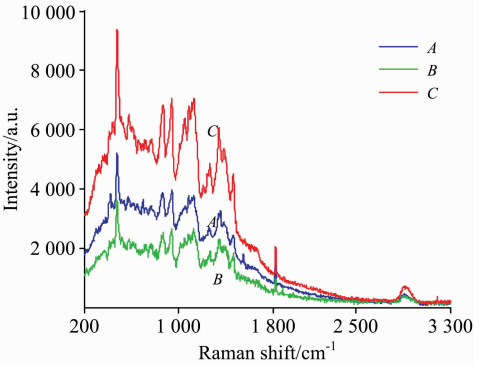

图1 三个产地大米原始光谱图

Fig. 1 Raw Raman spectrum of three producing area of rice

2.2 大米典型拉曼峰值指认

大米光谱特征峰[6,10]对应着内部化学键振动方式及大米中营养成分的差异，如图2所示，采用多项式拟合去除背景后的大米拉曼光谱的明显峰值出现在480，866，942，1 088，1 130，1 263，1 344，1 385，1 458，1 822和2 911 cm^{-1}处，峰值对应大米内部的主要营养物质，480 cm^{-1}为淀粉的骨架振动；866和942 cm^{-1}为支链淀粉的C—O—H和C—O—H变形振动；1 088 cm^{-1}为直链淀粉的C—O—H键弯曲振动；1 130 cm^{-1}为糖的C—O键伸缩振动和C—O—H键弯曲变形振动；1 263 cm^{-1}为蛋白质的酰胺Ⅲ带C—N键伸缩振动；1 344 cm^{-1}为糖的C—C键伸缩振动和C—O—H键弯曲变形振动；1 385 cm^{-1}为淀粉的C—C键伸缩振动；1 458 cm^{-1}为糖的C—H键弯曲振动；1 822 cm^{-1}为淀粉的O—C—O键

伸缩振动；2 911 cm^{-1}为淀粉的 H—C—H 键和 H—N—H 键伸缩振动；由此可见，主要特征峰出现在 200～1 900 和 2 800～3 000 cm^{-1}这两个位置区间，根据主要特征峰值出现的波段，选择 200～3 100 cm^{-1}的全波段进行建模分析。

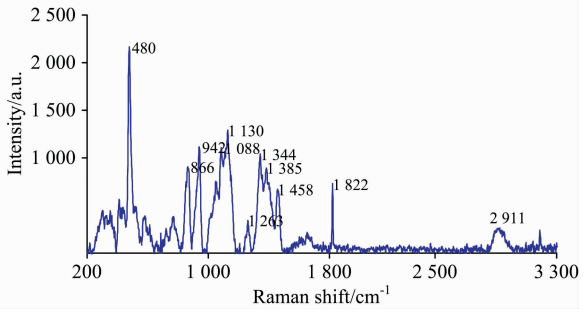

图 2　大米拉曼光谱主要特征峰

Fig. 2　Main characteristic peaks of rice Raman spectrum

2.3　大米拉曼光谱预处理方法

当前常用的预处理方法包括一阶导数、二阶导数、平移平滑、小波变换、多项式拟合等，结合大米光谱特征拉曼峰值的特点，下面选择四类九种预处理方法对光谱数据进行处理。

2.3.1　一阶导数＋平移平滑的预处理方法

一阶导数的数学表达式为

$$x_{i,1st}=\frac{x_{i+g}-x_i}{g} \tag{1}$$

其中 x_i 为第 i 个样品的光谱峰值的纵坐标，g 为步长，试验中是离散点求导，采用步长为 1。

再对一阶导数后的光谱用移动平均法进行平滑处理，数学表达式为

$$x_i=\frac{x_{i-n}+x_{i-n+1}+\cdots+x_i+\cdots+x_{i+n-1}+x_{i+n}}{2n} \tag{2}$$

式(2)中，$2n+1$ 为窗口大小、试验中 n 取 2；i 从第 3 点开始，对 x_{i-2}，x_{i-1}，x_i，x_{i+1}，x_{i+2} 五点求平均，然后赋值给 x_i，之后移动窗口，使 i 点遍历整个光谱到 3 098 点结束，即完成了移动平均法的平滑处理。

通过一阶导数消除了原始光谱曲线的平移和漂移，但同时曲线噪声被放大，原有多处波峰消失，并改变了拉曼光谱的形状。从图 3 中可知，采用常规的一阶导数＋平移平滑的预处理方法，需要再结合平移平滑对每个样本数据进行校正，消除数据中的噪音，突出显示光谱特征。

2.3.2　二阶导数＋平移平滑的预处理方法

二阶导数的数学表达式为

$$x_{i,2nd}=\frac{x_i-2x_i+x_{i+g}}{g^2} \tag{3}$$

其中 x_i 为第 i 个样品的光谱峰值的纵坐标，g 为步长，采用步长为 1，再对二阶导数后的光谱用移动平均法进行平滑处理。

常规的二阶导数＋平移平滑的预处理方法，在一阶导数基础上进行二阶导数并结合平滑滤波处理，如图 4 所示，因为二阶导数是对一阶导数处理后曲线再求拉曼强度的变化率，导致结果曲线峰值变小，特征谱峰不明显甚至消失。

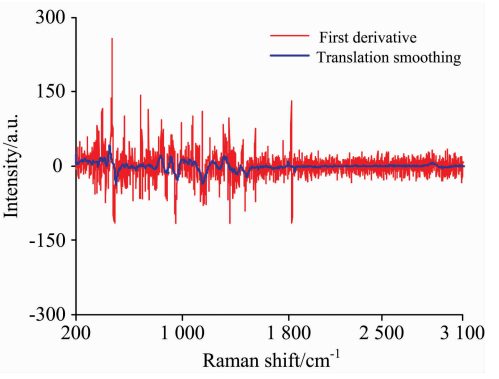

图 3　一阶导数＋平移平滑的预处理方法

Fig. 3　Pre-processing method of first derivative＋translation smoothing

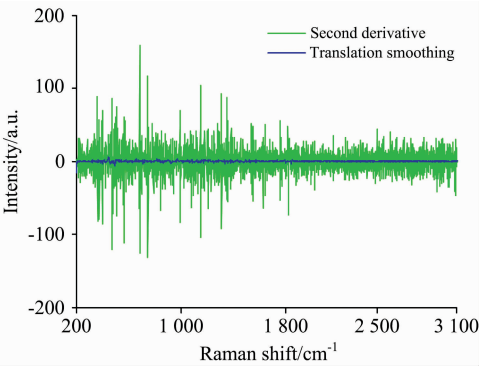

图 4　二阶导数＋平移平滑的预处理方法

Fig. 4　Pre-processing method of second derivative＋translation smoothing

2.3.3　小波变换＋去除基线的预处理方法

小波变换改善了傅里叶变换不能进行局部分析的缺陷，将信号用母小波函数 $\psi(t)$经过不同的平移和压缩分解成一系列小波，因为小波变换可以精细的对时域和频域的细节进行放大，使其具有很好的自适应性，但母小波函数不具有唯一性又使得分析时需要不断尝试，往往依靠经验和不断试验才能达到去噪和去除基线的目的。母小波函数的数学公式为[11]

$$\psi_{a,b(t)}=\frac{1}{\sqrt{|a|}}\psi\left[\frac{t-b}{a}\right],\ a,b\in R,\ a\neq 0 \tag{4}$$

其中 a 为压缩因子，b 为平移因子。大米光谱属于离散光谱经过多次对比分析选择效果最佳的信号进行处理，选取小波高通滤波采用 db9 小波基函数对原始光谱棱角 8 级分解，滤掉低频背景信号，选择硬阈值去噪，如图 5 所示，经小波变换处理后的光谱基线得到了校正，但基线仍有一定程度的漂移现象，主要产生在波段[1 800，3 100]这段背景噪声较大的区间。

2.3.4　分段多项式拟合＋去除基线的预处理方法

在相近产地大米鉴别中，因大米内部物质成分相似度极高，必要的预处理可以去除噪声，增强特征峰的强度，上述的三种预处理方法对荧光背景进行去除后，存在不能保持原

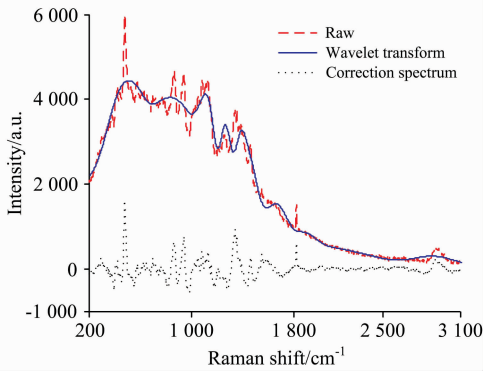

图 5 小波变换十去除基线的预处理方法

Fig. 5 Pre-processing method of wavelet transform+baseline removal

有波峰的形状或基线漂移去除的不彻底的现象，为了改善以上缺点，提出一种分段多项式拟合＋去除基线的预处理方法，这种预处理方法能保证拟合曲线恰到好处的通过原始波形下方，改进了传统的多项式拟合方法，对光谱区间进行分段，校正后的波形与原始波形最大限度的保持相似性。

在区间[200，3 100]共 2 901 个光谱数据点，设定窗口半宽为 15 cm^{-1}，从第 16 点开始，每 31 个点的平均值 $\bar{y}_i$ 赋值给中心点 i，比较 $\bar{y}_i$ 与 y_i 的大小，记录二者中较小值作为新的 y_i，之后移动 i 点到 3 086 点完成整条曲线的 y_i 的选取，初始 15 个点和最后 15 个点用原始光谱数据赋值，至此第一次迭代结束，随着迭代次数的增多，光谱峰值高度逐渐降低，滤波后基线都完全在原始光谱下方，使光谱的校正值都为正数，得到的光谱分段区间最小值，然后将这些最小值提取出来，将每个区间的最小值 y_i 用分段多点拟合方法[15]进行赋值，形成分段式多项式拟合方法基线，具体步骤如下：

(1)窗口半宽为 w，各测点 i 对应值为 y_i，在$(w+1, n-w)$区间取 y_i 的平均值，记为式(5)

$$\bar{y}_i = \frac{1}{2w+1}\sum_{j=-w}^{w} y_{i+j},\ i \in (w+1, n-w) \tag{5}$$

(2)将 y_i 和 $\bar{y}_i$ 进行比较，取两者中较小值作为新的 y_i 代替原值进行迭代，直到满足精度要求。其中 k 为迭代次数，记为

$$y_{i,k+1} = \min\{y_{i,k}, \bar{y}_{i,k}\},\ i \in (w+1, n-w) \tag{6}$$

(3)将迭代后的 y_i 值连接成线，找出曲线所有区间的最小值，记为

$$y_i < y_{i-1} \text{ 且 } y_i < y_{i+1},\ i \in (w+1, n-w) \tag{7}$$

(4)记录每个位移点的拟合值，在每个光谱位移点上得到一个初始 y_i 值，根据拟合点的个数 n，以步长 1 cm^{-1} 顺次移动光谱，历遍整个光谱在每个位移点上进行 n 次拟合后，取这些拟合值的平均值 $\bar{y}_i$，将 $\bar{y}_i$ 赋值给 y_i；

(5)将每个区间的 y_i 连接起来，形成分段多点拟合方法基线，如图 6(a)所示；

(6)在相同的拉曼位移上，用原始光谱曲线的数值对应减掉用分段多点拟合法的 y_i 数值，形成去除基线后的光谱。

采用上述方法对区间[200，1 700]的波段分别进行 3 点 2 次拟合、4 点 2 次拟合、5 点 2 次拟合，观察到分段多点拟合法中选取3点2次拟合曲线结果使更多的点通过原始光

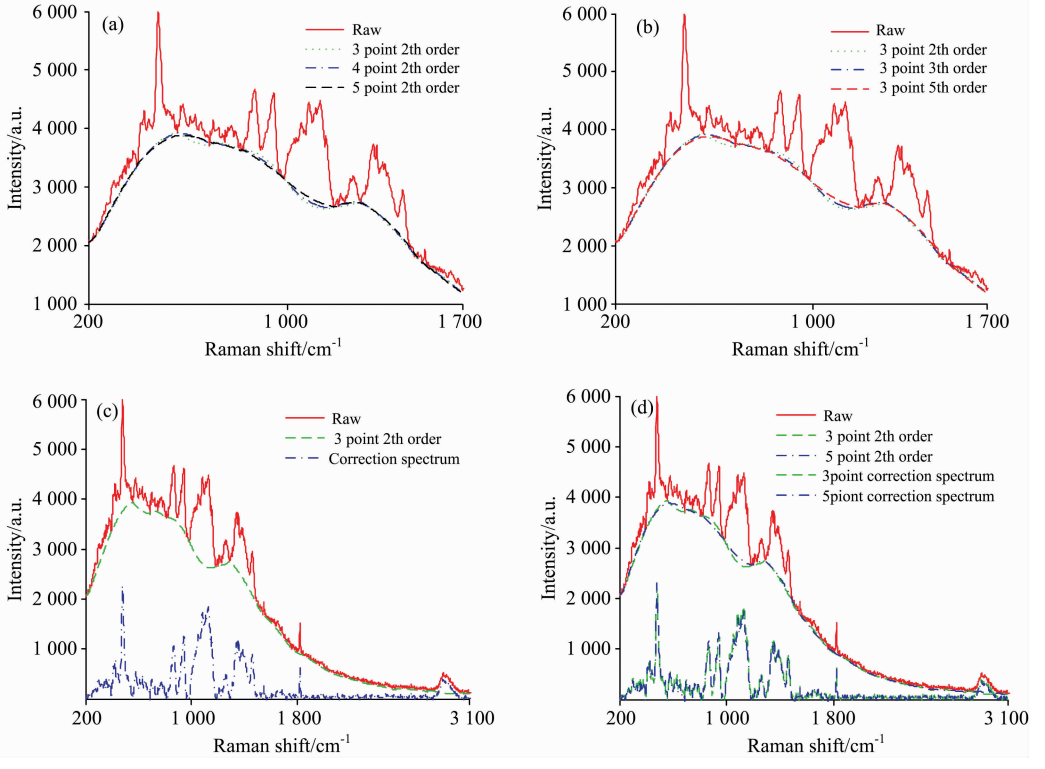

图 6 分段式多项式拟合十去除基线的预处理方法

Fig. 6 Pre-processing method of piecewise polynomial fitting+baseline removal

谱的波谷点，选择的拟合点数越多，拟合曲线的偏离值就越大，如图 6(a)中 5 点 2 次拟合曲线，拟合点通过波谷点的个数远远少于 3 点拟合曲线，究其原因是拟合的点数越多导致在这些拟合值的平均值 $\bar{y}_i$ 就和 y_i 的距离越大，如 6(a)中 5 点 2 次基线与 3 点 2 次基线比较，5 点 2 次的基线波峰和波谷的高度都相对较小，不如 3 点 2 次的波峰和波谷变化明显，基线变化的灵敏度不高，反应相对滞后，如图 6(d)所示，5 点 2 次校正光谱与 3 点 2 次校正光谱相比，使得原始光谱减掉对应的 5 点 2 次基线后校正光谱引入了更多的噪声。可见，点数越多，导致偏移值越大，因此本文采用 3 点拟合曲线去除基线。

再讨论拟合的阶数对波形的影响，如图 6(b)中分别进行 3 点 2 次拟合、3 点 3 次拟合、3 点 5 次拟合，可见拟合的次方越大，会使拟合曲线震荡的越剧烈，如图 6(b)中[200，600]区间，阶数越高偏移越大，而在 600 cm^{-1} 以后，几乎没有影响，分析原因可能是[200，600]区间分峰的大小和波形所致，如图 6(c)所示为采用 3 点 2 次拟合去除基线后的光谱，更好的保持了原有的特征峰面积和特定值，为实现光谱定量分析打下理论基础。

2.4　基于偏最小二乘法的不同预处理方法分类结果分析

为了对比上述不同预处理方法的优劣，每份样本中随机选取 33 个作为训练集样本、其余 17 个作为测试集样本。采用偏最小二乘法进行建模分析。并采用相关系数(r)、均方误差(MSE)、均方根误差(RMSE)来评价预处理的效果，其中 r 越大、MSE 和 RMSE 越小说明样本的预处理效果越好。

表 1 是对不同预处理方法所作的统计结果，从表中可见，在训练集中一阶导数＋平移平滑的预处理方法相关系数值最大、均方误差和均方根误差最小，3 点 2 次拟合＋去除基线的预处理方法相关系数值稍差，但与一阶导数＋平移平滑差距不明显，小波变换＋去除基线的预处理方法相关系数值最小、均方误差和均方根误差最大；在测试集中采用 3 点 2 次拟合＋去除基线的预处理方法的相关系数值最大、均方误差和均方根误差最小，3 点 3 次拟合＋去除基线的预处理方法稍差，二阶导数＋平移平滑的预处理方法最差。经过综合比较，采用 3 点 2 次拟合＋去除基线的预处理方法在训练集和测试集中都是比较理想的预处理方法。

表 1　不同预处理方法的相关系数 CC、均方误差 MSE、均方根误差 RMSE

Tabel 1　Correlation coefficient，Mean square error，Root mean square error of different pretreatment methods

序号	预处理方法	训练集			测试集		
		CC	MSE	RMSE	CC	MSE	RMSE
1	一阶导数＋平移平滑	0.993 8	0.002 7	0.052 2	0.935 7	0.027 7	0.166 4
2	二阶导数＋平移平滑	0.943 4	0.024 4	0.156 3	0.839 8	0.065 5	0.255 9
3	小波变换＋去除基线	0.885 6	0.047 9	0. 218 6	0.884 6	0.048 4	0.220 0
4	3 点 2 次拟合＋去除基线	0.990 0	0.004 4	0.066 5	0.960 6	0.017 2	0.131 1
5	4 点 2 次拟合＋去除基线	0.994 6	0.002 4	0.048 9	0.887 0	0.050 5	0.224 7
6	5 点 2 次拟合＋去除基线	0.954 8	0.019 6	0.140 1	0.690 6	0.154 5	0.393 0
7	3 点 3 次拟合＋去除基线	0.981 7	0.008 1	0.089 9	0.929 7	0.030 2	0.173 7
8	3 点 4 次拟合＋去除基线	0.977 6	0.009 8	0.099 2	0.923 9	0.032 6	0.180 5
9	3 点 5 次拟合＋去除基线	0.976 3	0.010 4	0.102 0	0.896 0	0.044 2	0.210 3

表 2　17 个测试样本中不同预处理方法的识别个数和识别率

Table 2　The number and recognition rate of different pre-processing methods in 17 test sample

序号	预处理方法	A		B		C		总识别率/%
		识别数	识别率/%	识别数	识别率/%	识别数	识别率/%	
1	一阶导数＋平移平滑	17	100	14	82.4	14	82.4	88.2
2	二阶导数＋墙移平滑	16	94.1	14	82.4	14	82.4	86.2
3	小波变换＋平移平滑	17	100	15	88.2	17	100	96.1
4	3 点 2 次拟合＋去除基线	17	100	17	100	17	100	100
5	4 点 2 次拟合＋去除基线	17	100	14	82.4	15	88.2	90.2
6	5 点 2 次拟合＋去除基线	15	88.2	7	41.1	5	29.4	52.9
7	3 点 3 次拟合＋去除基线	16	94.1	16	94.1	17	100	96.1
8	3 点 4 次拟合＋去除基线	16	94.1	17	100	17	100	98.0
9	3 点 5 次拟合＋去除基线	16	94.1	17	100	16	94.1	96.1

注：方法①—⑨在训练集的识别率均为 100%。

为了进一步验证不同预处理效果的差异，对 3 个产地样品共 150 份大米采用 PLS 进行建模分析，在训练集中，采用表 1 中的 9 种预处理方法对 A，B 和 C 三种大米的正确判别率均为 100%。在测试集中如表 2 所示：采用 3 点 2 次拟合＋去除基线预处理方法对 A，B 和 C 三产地大米总识别率为 100%，采用 5 点 2 次拟合＋去除基线预处理方法对 A，B 和 C 三产地大米总识别率为 52.9%，其他分段多项式拟合介于二者之间；采用一阶导数＋平移平滑、二阶导数＋平移平滑

和小波变换的预处理方法总识别率分别为 88.2%，86.2%和 96.1%；从中发现，采用一阶导数＋平移平滑的方法稍好于二阶导数＋去除基线的方法，这是因为二阶导数的噪声使更多特征峰不能突显出来，导数处理不如小波变换和 3 点 2 次拟合＋去除基线的效果，但小波变换过程需要通过先验知识确定的参数过多，没有通用规律可循，分段式多项式拟合中的 3 点 2 次拟合＋去除基线的预处理方法优势明显，与表 1 中 r，MSE 和 RMSE 的结果吻合，总体识别率高，鉴别效果稳定。

采用 3 点 2 次拟合＋去除基线的预处理方法进行建模，并分别将 A，B 和 C 三产地大米样本赋值 1，2 和 3，结果在 1±0.5(不含 1.5)鉴别为 A 大米、结果在 2±0.5(不含 2.5)鉴别为 B 大米、结果在 3±0.5(不含 3.5)鉴别为 C 大米，结果如图 7 所示。A 大米的测试值主要集中在 0.69～1.02、B 大米样本的测试值主要集中在 1.54～2.01、C 大米的测试值主要集中在 2.75～3.01，均具有明显的聚类趋势。说明该模型预测结果具有较好的精度，可以很好的实现三种近地大米的产地鉴别。

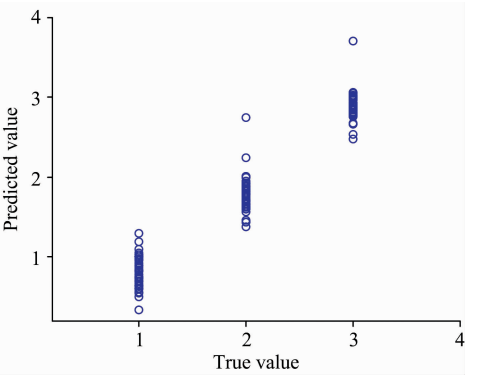

图 7 真值与预测值关系图

Fig. 7 Relationship between true value and predicted value

3 结 论

拉曼光谱技术结合不同预处理方法对相近三个产地的大米进行鉴别，分别采用一阶导数＋平移平滑、二阶导数＋平移平滑、小波变换＋去除基线的方法进行光谱预处理，因为这些方法存在不能保持原有波峰的形状或基线漂移的现象，提出一种分段多项式拟合＋去除基线的预处理方法，通过偏最小二乘法 PLS 对 150 个样本三个产地大米建立拉曼模型，实验结果表明经过分段多项式拟合＋去除基线中的 3 点 2 次多项式的预处理后建立的模型精度最高，在训练集和测试集中三个产地的识别率均为 100%，聚类效果好。通过 3 点 2 次多项式＋去除基线的预处理为相近产地大米鉴别分析提供了一种有效方法，同时为近地域其他农作物鉴别提供技术参考。

References

[1] YANG Wan-jiang, LIU Qi(杨万江，刘 琦). Research of Agricultural Modernization(农业现代化研究), 2019, 40(1): 44.
[2] YANG Yong(杨 永). Journal of Zhejiang Agricultural Sciences(浙江农业科学), 2018, 59(7): 1082.
[3] HUANG Jia-rong, WU Bo-di, ZHAN Qiu-qiang(黄嘉荣，伍博迪，詹求强). Acta Laser Biology Sinica(激光生物学报), 2015, 24(3): 237.
[4] SUN Juan, ZHANG Hui, WANG Li, et al(孙 娟，张 晖，王 立，等). Food & Machinery(食品与机械), 2016, 32(1): 41.
[5] ZHAO Ying, LI Ming, WANG Xiao-long, et al(赵 迎，李 明，王小龙). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(5): 1468.
[6] SHA Min, GUI Dong-dong, ZHANG Zheng-yong, et al(沙 敏，桂冬冬，张正勇，等). Journal of the Chinese Cereals and Oils Association(中国粮油学报). 2020, 35(1): 168.
[7] Hu H, Bai J, Xia G, et al. Photonic Sensors, 2018, 8(4): 332.
[8] Cai Y Y, Yang C H, Xu D G, et al. Journal of Raman Spectroscopy, 2019, 50(3): 454.
[9] Monago-Marana O, Eskildsen C E, Afseth N K, et al. Food Chemistry, 2019, 274: 187.
[10] ZHU Zi-ying(朱自莹). Application of Raman Spectroscopy in Chemistry(拉曼光谱在化学中的应用). Changchun: Northeast University Press, 1998. 295.
[11] CHU Xiao-li(褚小立). Molecular Spectroscopy Analytical Technology Combined with Chemometrics and Its Appllcations(化学计量学方法与分子光谱分析技术). Beijing: Chemical Industry Publishing House(北京：化学工业出版社), 2011. 49.
[12] Leon-Bejarano F, Mendez MO, Ramirez-Elias MG, et al. Applied Spectroscopy, 2019, 73(12): 1436.

Identification of Rice From Similar Areas With Different Pretreatment Methods of Raman Spectrum

WANG Ya-xuan[1], TAN Feng[2*], XIN Yuan-ming[2], LI Huan[2], ZHAO Xiao-yu[2], LU Bao-xin[3]

1. College of Civil Engineering and Water Conservancy, Helongjiang Bayi Agricultural University, Daqing 163319, China
2. College of Electrical and Information, Heilongjiang Bayi Agricultural University, Daqing 163319, China
3. Food College, Heilongjiang Bayi Agricultural University, Daqing 163319, China

Abstract It is difficult for consumers to distinguish regional rice brands formed by geographical factors instead of rice of similar producing areas. Based on Raman spectroscopy, the experiment compared three common pretreatment methods including first derivative + translational smoothing, second derivative + translational smoothing, wavelet transform + baseline removal. In addition, an improved piecewise polynomial fitting + baseline removal was proposed, and a total of four pretreatment methods respectively were combined with partial least square method to identify rice of similar origin, and an optimal pretreatment method for identifying rice of similar origin was proposed. Firstly, 150 rice spectral samples with a Raman displacement of 200～3 100 cm^{-1} were collected by Raman spectrometer from three similar producing areas in an county, Heilongjiang province. Then, the original Raman spectra were preprocessed by first derivative + translational smoothing, second derivative + translational smoothing, wavelet transform+baseline removal, and piecewise polynomial fitting + baseline removal. A total of 99 samples were selected from 33 samples from each origin for training, and a partial least square analysis model based on partial least square method was established for the unknown 51 samples. In the training set, the preprocessing method of first derivative+translation smoothing had the maximum correlation coefficient value, the minimum mean square error and the minimum root mean square error. Andwavelet transform+baseline removal hadthe minimum correlation coefficient value, the maximum mean square error and the maximum root mean square error. In the test set, the preprocessing method of 3 points and 2 times fitting+baseline removal had the maximum value of correlation coefficient, the minimum mean square error and the minimum root mean square error, and the preprocessing method of the second derivative+translation smoothing was the worst. Finally, the PLS modeling results showed that, in the training set, the correct discrimination rate of rice from three producing areas was 100% by using four kinds and nine pretreatment methods. In the test set, the total recognition rate of rice from three producing areas was 100% by using 3-point 2-time fitting+baseline removal, and 52.9% by using 5-point 2-time fitting + removing baseline. Other piecewise polynomial fitting was between the two. The total recognition rates of the first derivative + translation smoothing, the second derivative + translation smoothing and the wavelet transform were 88.2%, 86.2% and 96.1%, respectively. It is found that the preprocessing method of 3-point 2-time fitting + removing baseline in piecewise polynomial fitting has obvious advantages and is consistent with the results of the correlation coefficient, mean square error and root mean square error, with high overall recognition rate and stable identification effect.

Keywords Raman spectrum; Baseline removal; Rice; Pretreatment method

(Received Dec. 25, 2019; accepted Apr. 22, 2020)

* Corresponding author