基酒 FT-NIR 光谱预处理与特征波筛选方法的比较

朱雪梅[1]，庹先国[1*]，张贵宇[1,2*]，翟双[1]，罗林[1]，罗琪[1]
（1.四川轻化工大学自动化与信息工程学院，人工智能四川省重点实验室，四川宜宾 644000）
（2.西南科技大学信息工程学院，四川绵阳 621010）

摘要：为解决白酒基酒分类的问题，降低基酒的分类误差，减少基酒对摘酒师傅身体的危害，本实验选取 18 种预处理以及 3 种特征波筛选方法来较少光谱中的无关干扰信息，降低建模数据复杂度。基酒的傅里叶近红外光谱（Fourier Transform Near Infrared Spectroscopy，FT-NIR）经过光谱理化值共生距离法（SPXY）划分数据集、预处理、马氏距离（MD）异常剔除、特征波筛选、支持向量机回归（SVR）预测来完成最终的分类。研究发现：多元散射校正（Multiplicative Scatter Correction，MSC）后的训练集预测集分类准确率可以达到 100%，主成分分析（Principal Component Analysis，PCA）与特定算法结合才能实现准确分类，因此要注意与其他算法的组合，无信息变量消除法（Uninformative Variables Elimination，UVE）和竞争性自适应重加权算法（Competitive Adaptive Reweighted Sampling，CARS）都能实现高效的特征波选择，预测集的平均准确率接近 90%。实验证明，经过处理后的光谱数据最多占原数据的 47.57%，基酒近红外谱图经过预处理与特征波筛选后可以降低后期回归模型处理数据的复杂程度，提高模型的精确度。

关键词：近红外；基酒分级；多元散射校正；无信息变量消除法；竞争性自适应重加权算法
文章编号：1673-9078(2023)01-196-204 DOI: 10.13982/j.mfst.1673-9078.2023.1.0271

Comparison of FT-NIR Spectral Pretreatment and Characteristic Band Screening for Baijiu-based Liquor

ZHU Xuemei[1], TUO Xianguo[1*], ZHANG Guiyu[1,2*], ZHAI Shuang[1], LUO Lin[1], LUO Qi[1]
(1.School of Automation & Information Engineering, Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644000, China)
(2.School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China)

Abstract: To classify baijiu base wine, reduce the classification error of baijiu base wine, and reduce the harm of base wine to the body of Baijiu Based Liquor pickers, 18 pretreatment methods and three characteristic wave screening methods were selected to reduce irrelevant interference information in the spectrum and complexity of the modeling data. The Fourier-transform near infrared spectra of baijiu base wine were divided into datasets using SPXY and preprocessed, and then subjected to Mahalanobis distance anomaly elimination, eigenwave screening, and support vector machine regression prediction. After multiplicative scatter correction, the classification accuracy of training set prediction was 100%. Principal component analysis can be combined with specific algorithms to achieve accurate classification; studies are needed to combine this analysis with other algorithms. Uninformative variables elimination and competitive adaptive reweighted sampling can achieve efficient feature wavelength extraction, with an average accuracy of prediction of close to 90%. The experimental results showed that the processed spectral data accounted for up to 47.57% of the original data, the complexity of the regression model was reduced, and the accuracy of the model was improved after pretreatment and characteristic wavelength selection of the near-infrared spectrum of the base wine.

引文格式：
朱雪梅,庹先国,张贵宇,等.基酒 FT-NIR 光谱预处理与特征波筛选方法的比较[J].现代食品科技,2023,39(1):196-204
ZHU Xuemei, TUO Xianguo, ZHANG Guiyu, et al. Comparison of FT-NIR spectral pretreatment and characteristic band screening for Baijiu-based liquor [J]. Modern Food Science and Technology, 2023, 39(1): 196-204

收稿日期：2022-03-11
基金项目：四川省科技计划项目（2022YFS0554）；四川省重大科技专项项目（2018GZDZX0045）；四川省科技成果转移转化示范项目（2020ZHCG0040）
作者简介：朱雪梅（1997-），女，硕士研究生，研究方向：白酒智能酿造与应用，E-mail：19704064084@qq.com
通讯作者：庹先国（1965-），男，博士，教授，研究方向：核技术应用，E-mail：tuoxianguo@suse.edu.cn；共同通讯作者：张贵宇（1987-），男，博士生在读，讲师，研究方向：白酒自动化、人工智能，E-mail：gyz_118@163.com

Key words: near infrared; base wine classification; multivariate scattering correction; uninformative variable elimination; competitive adaptive reweighting algorithm

白酒是我国传统产业的代表之一，是我国特有的蒸馏酒[1]。白酒大都以谷物为原料，通过发酵得到高度白酒基酒[2]。基酒经过长期储存、陈化老熟、勾兑降度后包装为成品酒。由于发酵过程中不同的使用酒曲、原料、辅料以及不同的发酵工艺条件，形成了不同香气、口感的白酒[3,4]。白酒基酒作为粮食到成品酒的一个中间产物，对最终的成品酒质量有重大的影响。在生产车间，白酒基酒的分段主要依靠现场工人的经验，他们通过看花摘酒以及品尝口感来快速区分不同等级的基酒[5]。但是，刚馏出的基酒含有大量的二甲酸二甲酯、邻苯二甲酸二异丁酯、邻苯二甲酸二丁酯、邻苯二甲酸二丁酯等有害物质，这些物质对人体伤害较大，不适合长期饮用[6]。实现基酒智能分级技术将是后期白酒酿造自动化一个重要的技术点，将会推动白酒向标准化、无人化、智能化发展[7]。

近年来，随着计算机科学、模拟技术以及分析化学的快速发展，有多种手段可以从近红外光谱里发现丰富的物质信息，近红外谱图的高效解析已经逐渐变得可行，近红外的应用前景越来越广泛[8]。同时，近红外光谱由于其快速便捷、清洁环保、低成本的特点，以及较好的检测效果备受关注[9]。如今，近红外分析技术广泛应用于石化企业[10]、饲料生产[11]、食品[12,13]、制药[14]、制烟[15,16]等各个行业[17]。近红外光（Near Infrared，NIR）是一种波长范围在 780 nm~2 526 nm 的电磁波[18]，近红外光谱区域内，物质内部的含氢官能团发生基频、合频振动，得到的物质吸收谱能较好的反映样本中的含氢基团信息。但这个谱区内存在吸收强度弱，灵敏度相对较低，吸收带宽且重叠严重等问题，这就使得最后获得的光谱与理论上的光谱有误差。此外，由于近红外光谱仪运行过程中湿度、温度等外界环境以及仪器自身噪声的影响使得获得的光谱包含大量的无关干扰信息严重影响了后续建模效果[19]。加之，基酒的微量成分的种类多，含量少，图谱间差异不明显，良好的光谱预处理以及特征波筛选就备受关注。

1 材料和方法

1.1 样本获取

本实验选择某品牌酿酒车间酒糟馏出的白酒基酒为样本。为减少个人主观因素对分段结果的影响，选取了 4 位有 10~20 年摘酒经验的师傅来分段摘取基酒，跟踪这 4 位师傅的摘酒情况，在师傅的指导下获取实验样本并贴好标签。为模拟现场摘酒环境中基酒震荡产生酒花的情况，样品通过注射器注入外接流通池，待酒花稳定后进行光谱获取，表 1 是具体分类情况。

表 1 基酒样品信息

Table 1 Baijiu sample information

标签序号	酒等级	数量	备注
1	头酒	148	质量不好，最先馏出的原浆酒，前期会掺杂上一次残留的尾酒
2	一段酒	282	质量最好，中间段流出的原浆酒，各种微量物质以及乙醇含量都较高，香味重、度数高
3	二段酒	168	质量较次，仍然含有一定量的乙醇，但是微量物质较少，香气寡淡
4	尾酒	106	质量差，各种物质含量都比较低，香气寡淡，有微溶于水的脂类物质析出使得酒内有白色絮状析出

1.2 样本光谱数据采集

本研究采用德国 Bruker 公司傅里叶变换近红外光谱仪 Matrix-F 以及配套的近红外光纤探头获取样品透射光谱，使用软件 OPUS 7.8 控制光谱仪。在实验前，先确保光谱仪在温度为（20±2）℃、空气相对湿度＜80% RH 的条件下预热近红外光谱仪 50 min 左右。在扫描样品前先检查信号，使得干涉图能量达到最大，获取背景光谱，消除水蒸气以及二氧化碳等对光谱结果的影响。使用的光纤通道长度为 5 m、流通池的光程为 2 mm、光谱扫描范围为 4 000 cm^{-1}~12 500 cm^{-1}，相位分辨率为 32 cm^{-1}，以 10 kHz 的频率、16 cm^{-1} 的分辨率累积扫描 64 次后取每个光谱点上的平均值为最终光谱。

1.3 光谱预处理以及特征波筛选方法

本实验预处理使用 The Unscrambler X 10.4 为工具，将扩展多元散射校正（Extended Multiplicative Scatter Correction，EMSC）、多元散射校正（Multiple Scatter Correction，MSC）、标准正态变换（Standard Normal Variate Transformation，SNV）、导数、归一化、平滑以及其组合方法用于原始数据的预处理。后期通过 matlab 自编程序实现特征波筛选以及基酒等级回归预测，完成基酒等级分类的目标。良好的预处理以及光谱特征波筛选有助于减少杂峰、突出检测主体，减

少后期建模数据的复杂度，增加光谱数据的代表性，降低后期建模难度，提高处理精度。

1.3.1 预处理

EMSC、MSC、SVN 都主要用于消除颗粒分布不均造成的散射效应对样品光谱的影响[20]。MSC 通过计算所有样本的平均光谱，将每条光谱与平均光谱做一元线性回归，得到线性回归方程的斜率和截距，以此对原始光谱数据进行校正，整个过程中样品的成分含量信息在整个过程中不受影响，光谱信噪比提高了，每个光谱的基线平移和偏移都在参考平均光谱后被校正[21]。EMSC 就是在 MSC 的基础上通过波长相关效应或先验信息更好地分离物理光散射效应和化学光吸收效应。SNV 相较 MSC 来说通常需要计算每条光谱的平均值，而不是所有样本的平均值，然后除以光谱集的标准偏差，这就使得 SVN 更适合于差异较大的样本。通过求导可以使得光谱更加平滑，能够得出物质吸收峰所在位置，可以明显消除基线和背景干扰，分析重叠峰，提高分辨率以及灵敏度，但同时会引入噪声，降低信噪比。归一化在可以一定程度上去除由于待测基酒含量不同所导致的数据集的方差。平滑可以消除光谱信号中叠加的随机误差从而提高信噪比，是一种最基础的去噪方法。以卷积平滑（Savitzky-Golay Smoothing）为例，它核心思想是最小二乘拟合移动窗口中的数据。

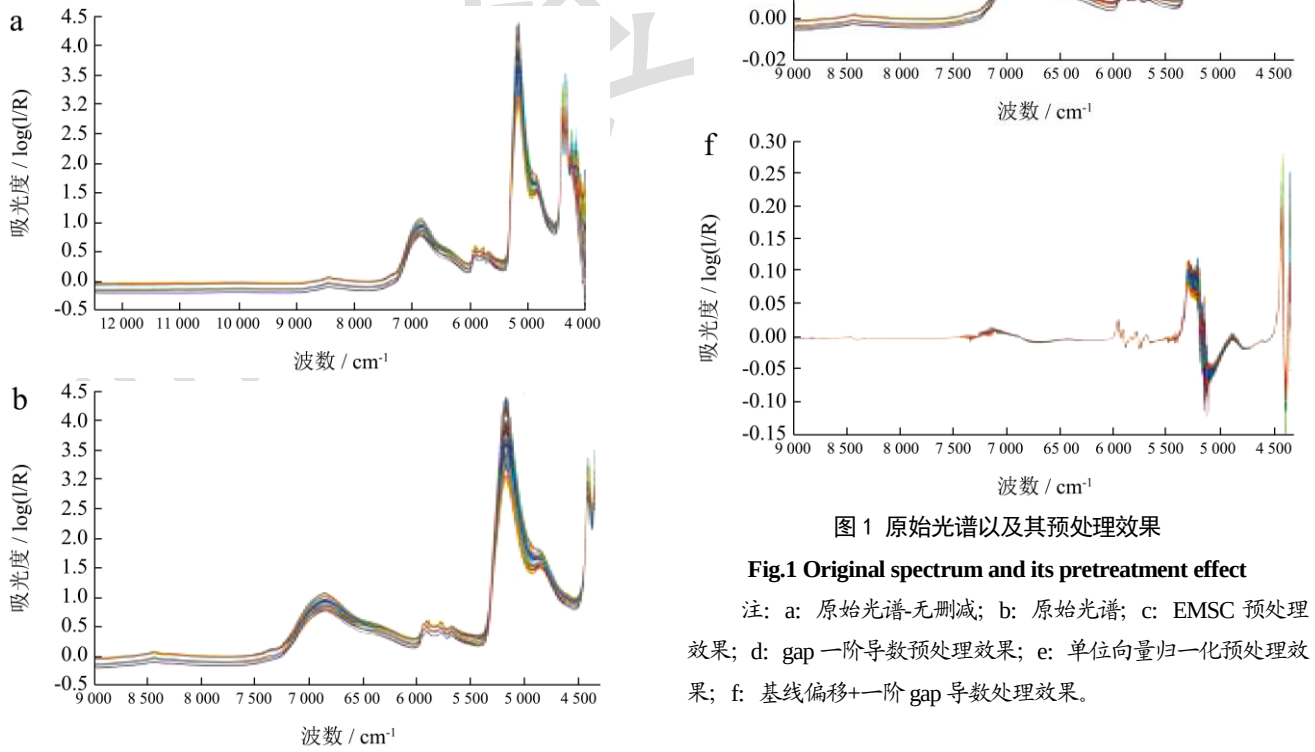

图 1 原始光谱以及其预处理效果

Fig.1 Original spectrum and its pretreatment effect

注：a: 原始光谱-无删减；b: 原始光谱；c: EMSC 预处理效果；d: gap 一阶导数预处理效果；e: 单位向量归一化预处理效果；f: 基线偏移+一阶 gap 导数处理效果。

本实验使用光谱仪获得的图 1a 的光谱图光谱范围在 34 000 cm^{-1}~12 500 cm^{-1}，但是待测基酒本身、实验仪器以及实验环境的影响，使得光谱头部 9 017.738 cm^{-1}~12 500 cm^{-1} 之间样本差异不大，会增加建模数据量，尾部 34 000 cm^{-1}~4 343.017 cm^{-1} 数据混乱无章，容易引入错误的建模规律。因此，在预处理前，剔除光谱头部以及尾部的数据，得到后期用于预处理的光谱图 1b。由图 1b 中的信息可以发现波段内有两个较明显的吸收峰分别位于 6 896 cm^{-1}、5 128 cm^{-1} 附近，以及三个不明显的波峰，位于 5 600 cm^{-1}~6 000 cm^{-1} 波段附近，除此以外，波段 8 000 cm^{-1}~9 000 cm^{-1} 之间有一个较平缓的吸收峰。结合相关近红外谱图知识可知 6 896 cm^{-1}、5 128 cm^{-1} 附近的峰是水的特征区间，因此，水含量作为基酒分类的一个较小的影响因素，这两个波段位置可以在后期特征筛选中适当减少选取的数据。图 1c~1f 是不同预处理对光谱的影响。

1.3.2 特征波筛选

特征波筛选算法中的主成分分析（Principal Component Analysis，PCA）是光谱特征波提取，竞争性自适应重加权算法（Competitive Adaptive Reweighted Sampling，CARS）和无信息变量消除法（Uninformative Variables Elimination，UVE）是光谱特征波选择。光谱特征波提取就是通过寻找光谱数据间的关系，使用新的方法表示光谱数据，这些更加简洁高效的数据就可以更好地代表以前的数据，例如：PCA、线性判别分析（Linear Discriminant Analysis，LDA）、独立成分分析（Independent Component Analysis，ICA）等。光谱特征波选择就是利用信号处理方法，在不改变原有数值的基础上将原始光谱中的决定性波数或波段筛选出来，例如：CARS、UVE、子区间最小二乘法（Interval Partial Least Squares，IPLS）等。特征波选择适用于特征峰相对集中的光谱，特征波提取主要用于特征变量分散的复杂数据。特征筛选的主要目的是降低数据维数，减少后期处理难度，剔除无用或者错误信息对最后分类结果的干扰。

1.3.2.1 竞争性自适应重加权算法（CARS）

CARS[22]是一种应用于筛选、剔除冗余光谱数据的方法，这个方法将每个波长变量当作一个独立个体，对波长进行逐步淘汰。本实验中，进行 1 000 次蒙特卡洛迭代采样与偏最小二乘回归，以此来获取 CARS 的最佳主成分数。筛选前期，变量随着采样次数的增加，筛选出的变量数快速减少，后期，即使采样次数增加，波长数量也很难减少了，最后，通过最小的交叉验证均方根误差（RMSECV）值来确定保留的波数。下图是以 CARS 对基酒光谱进行处理的一个图，图 2a~2c 中最上面是筛选波的数量随着采样次数变化的曲线图，中间是 RMSECV 值随 ARS 运行次数变化的曲线，下面是变量回归系数路径随 ARS 运行次数的变化趋势图。图 2d 是没有经过预处理，只使用了马氏距离（Mahalanobis Distance，MD）剔除异常的 CARS 选取的数据结果，图中加粗的蓝色为选出的特征波。

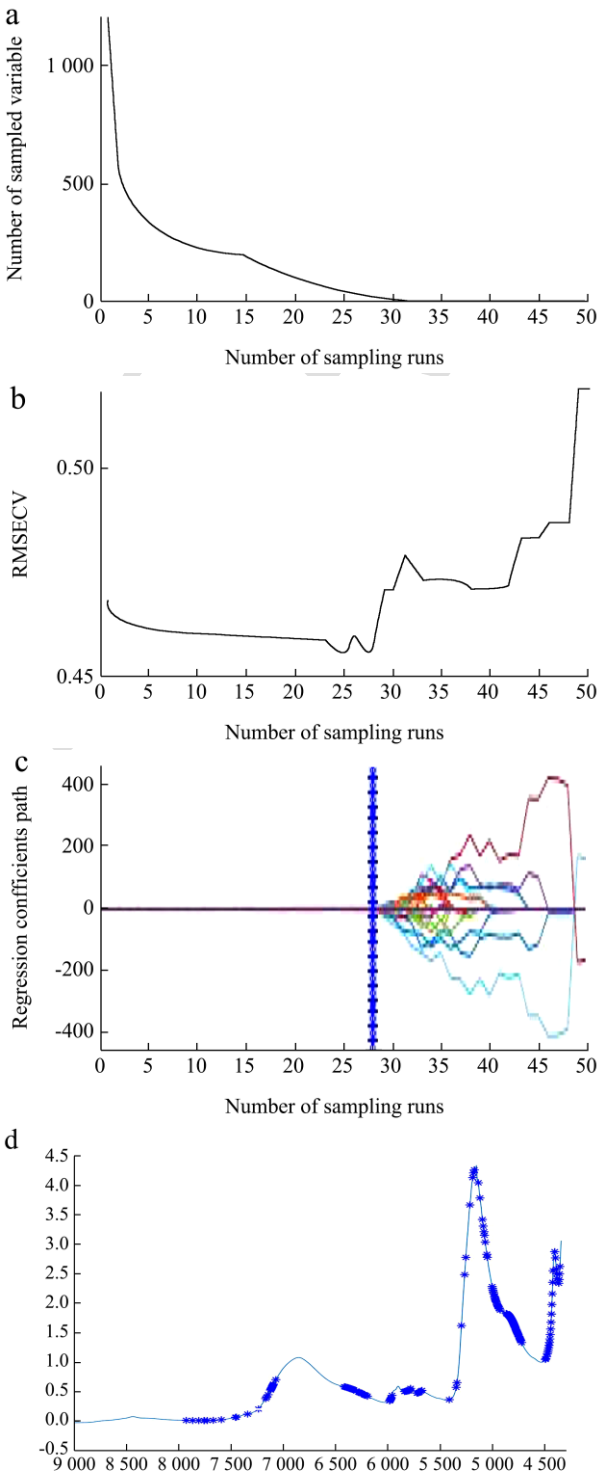

图 2 CARS 选取及结果

Fig.2 Selection and results of CARS

注：a~c：CARS 提取基酒特征波；d：CARS 提取结果。

1.3.2.2 无信息变量消除法（UVE）

UVE 是基于偏最小二乘回归系数构建的波长选择算法，以回归系数作为波长选择的最重要的衡量指标，剔除包含无效信息的光谱数据[23]。图 3a 是 UVE 处理的一个图，图中右边深红色为添加的 1 215 个随机光谱噪声数据，两条水平轴为变量选择的阈值，两轴之间的黑色部分是被剔除的无用数据，绿色部分为稳定的光谱数据值，是筛选出来的超过阈值的波数，这些数据将用于后期建模。图 3b 是没有经过预处理，只使用了马氏距离剔除异常的 UVE 选取的数据结果，图中加粗的蓝色为选出的特征波。

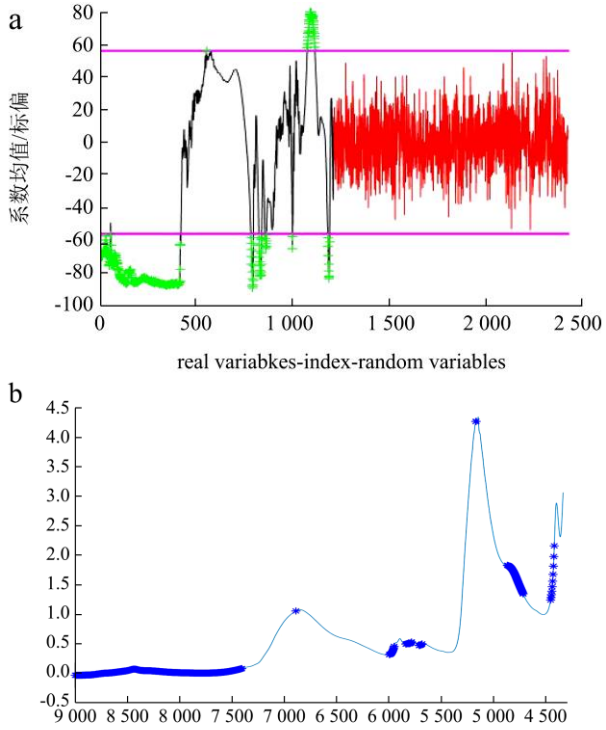

图 3 UVE 选取及结果 UVE

Fig.3 UVE selection and results

注：a: CARS 提取基酒特征波；b: CARS 提取结果。

1.3.2.3 主成分分析（PCA）

PCA 是利用线性变化简化数据集的一种算法，这个算法的原理是将原矩阵变量组合成一组新的相互正交的几个综合变量，然后根据精度需要从前到后选择信息互不重复的新变量去反映原来的变量[24]。一般来讲，新变量的综合指标是通过方差来表示，方差越大，包含的原光谱信息就越多。PCA 不同于前面的两种特征选择算法，它在保留原始特征的基础上将高维数据降为低维数据。图 4 是未经过预处理的光谱 PCA 降维结果，在帕累托图中可以看出，前两个维度就提取出了大于 90%的原光谱特征，在保留数据特征的基础上，很好的压缩了数据量。但是为了避免选取太少的波数从而出现过拟合，在本实验中规定 PCA 的贡献大于 1 的成分是选取的成分，但是由于 PCA 能够通过很少的主成分极大限度地表征被测样品的性质以及物质组成，因而导致后期建模过程出现过拟合，因而，额外规定 PCA 选取主成分至少为 20 个。

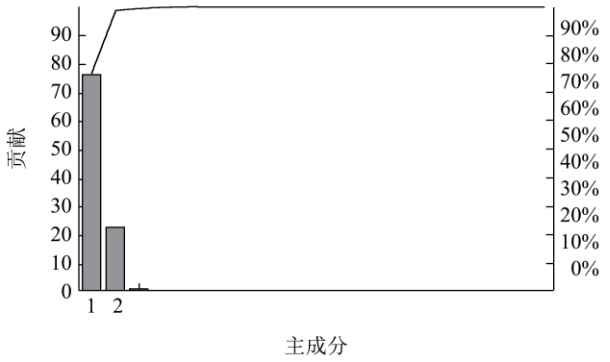

图 4 PCA 提取特征波

Fig.4 PCA extraction of characteristic bands

2 分类模型的建立及分析

2.1 模型的建立与结果比较

为验证预处理方法：扩展多元散射校正（EMSC）、多元散射校正（MSC）、标准正态变换（SVN）、一阶 gap 导数（gap）、一阶 SG 导数（SG）、卷积 3 点平滑（SG+[3]）、卷积 5 点平滑（SG+[5]）、卷积 7 点平滑（SG+[7]）、高斯滤波、移动均值滤波、中值滤波、单位向量归一化、面积归一化、均值归一化、基线偏移、基线偏移+gap 一阶导数（基线+1D）、基线偏移+线性基线校正（偏移+线性校正）、基线偏移+卷积平滑（偏移+卷积）的有效性，所有数据样本通过光谱理化值共生距离法（SPXY）给数据集分类，然后使用预处理校正样本数据以及 MD 剔除异常数据，随后分别选取 PCA、UVE、CARS 去筛选光谱特征波，最后通过支持向量回归（SVR）预测基酒种类。在筛选特征变量的过程中，使用了决定系数（R^2）、交叉验证均方根误差（*RMSECV*）来衡量 CARS 算法，校正集标准偏差（*RMSEC*）、相关系数（*R*）来衡量 UVE 算法，使用预测集集标准偏差（RMSEP）、预测集相关系数来（R_p）衡量 PCA 算法，表 3 是具体测试结果。

表 3 具体测试结果

Table 3 Specific test results

分类	序号	预处理	R^2/%	*RMSECV*/%	剩余波长个数	分类准确率/%	
						训练集	预测集
CARS 处理结果	1	原始数据	76.10	45.54	101	92.25	87.78
	2	EMSC	99.04	8.35	89	100.00	100.00
	3	MSC	99.04	8.40	115	100.00	100.00
	4	SVN	75.24	46.18	101	82.00	76.49
	5	gap	77.24	41.62	24	100.00	97.49
	6	SG	76.36	42.41	16	95.75	89.12
	7	SGolay+[3]	74.62	44.16	68	91.50	82.57
	8	SGolay+[5]	75.28	43.55	60	92.75	90.00
	9	SGolay+[7]	76.76	42.26	68	97.50	96.68
	10	高斯滤波	72.51	45.96	78	93.25	90.87
	11	移动均值滤波	75.98	45.74	46	84.50	81.91
	12	中值滤波	76.30	45.43	78	87.00	80.81
	13	单位向量归一化	75.87	46.31	89	97.00	92.34
	14	面积归一化	73.21	45.81	31	82.50	82.50
	15	均值归一化	73.75	45.35	192	86.25	82.16
	16	基线偏移	71.99	45.80	46	81.25	81.25
	17	基线+1D	79.32	42.21	31	98.50	95.15
	18	基线偏移+线性基线校正	74.16	38.92	53	82.00	81.27
	19	基线+卷积	76.88	45.47	27	94.50	92.80
UVE 处理结果	1	原始数据	47.54	86.00	526	96.50	82.46
	2	EMSC	7.99	99.47	1033	100.00	99.75
	3	MSC	7.32	99.48	1031	100.00	100.00
	4	SVN	38.37	85.35	162	79.50	82.46
	5	gap	36.71	84.35	1031	98.75	87.03
	6	SG	36.94	84.38	1061	98.50	85.82
	7	SGolay+[3]	47.70	85.95	564	92.00	92.00
	8	SGolay+[5]	46.44	86.74	405	93.75	90.77
	9	SGolay+[7]	46.48	86.71	343	97.32	95.54
	10	高斯滤波	46.39	87.72	279	90.50	82.99
	11	移动均值滤波	46.43	84.82	477	96.00	85.89
	12	中值滤波	48.58	83.24	665	95.25	92.53
	13	单位向量归一化	50.44	84.49	559	81.75	79.44
	14	面积归一化	50.63	82.03	567	83.25	82.08
	15	均值归一化	48.24	83.84	218	83.25	82.04
	16	基线偏移	47.33	83.73	376	90.50	83.47
	17	基线+1D	36.41	86.59	1023	98.50	87.03
	18	基线偏移+线性基线校正	64.03	72.89	279	79.75	90.84
	19	基线+卷积	53.12	85.71	376	96.00	90.25

续表 3

分类	序号	预处理	R^2/%	*RMSECV*/%	剩余波长个数	分类准确率/%	
						训练集	预测集
PCA 处理结果	1	原始数据	23.90	86.91	20	86.75	74.17
	2	EMSC	4.11	99.51	20	100.00	100.00
	3	MSC	4.50	99.44	20	100.00	100.00
	4	SVN	28.23	81.02	20	89.75	79.75
	5	gap	25.42	80.69	42	98.25	81.17
	6	SG	24.71	83.57	41	98.25	81.17
	7	SGolay+[3]	26.40	87.64	20	89.00	74.27
	8	SGolay+[5]	24.22	89.79	20	84.75	74.54
	9	SGolay+[7]	28.54	45.05	20	80.75	73.44
	10	高斯滤波	26.56	86.80	20	86.75	74.91
	11	移动均值滤波	27.37	78.93	20	84.75	84.06
	12	中值滤波	28.84	86.16	20	82.75	71.96
	13	单位向量归一化	29.69	84.83	20	79.00	73.29
	14	面积归一化	26.92	80.54	20	95.75	77.70
	15	均值归一化	26.41	84.36	20	79.75	73.88
	16	基线偏移	46.43	24.18	20	90.25	76.70
	17	基线+1D	23.72	86.59	20	98.25	81.17
	18	基线偏移+线性基线校正	43.08	62.65	20	92.25	82.56
	19	基线+卷积	27.93	85.71	20	83.75	77.42

2.2 分类效果分析

2.2.1 评价指标

上述预处理与特征波筛选算法结合后的分类效果中，通过计算 18 种预处理方法、3 种特征波筛选方法的平均准确率以及准确率的绝对偏差平均值来评价算法整体，具体数据如下表所示：由这些数据可以看出，原始数据经过三种预处理算法后的平均正确率为 81.47%，除了 SVN、面积归一化、均值归一化、基线偏移以外的预处理方法均能提高分类正确率，其中 EMSC、MSC 算法预处理处理后的数据经过 3 种特征波算法筛选后分类效果接近 100%；CARS、UVE、PCA 结合本实验的所有预处理方法后预测集的平均分类准确率分别达到 88.48%、88.02%、79.59%，准确率的绝对偏差平均值分别为 6.28%、5.00%、5.60%。由此可以看出，EMSC、MSC、gap、SG+[3]、SG+[5]、SG+[7]、高斯滤波、移动均值滤波、基线+1D、基线+线性偏移、基线+卷积、CARS 和 UVE 均能提高后期回归模型的分类准确率。

2.2.2 预处理结果对比分析

通过图 5 可以看出 EMSC 和 MSC 的分类效果稳定且分类准确率明显要高于其他预处理方法，这两个算法经过三个不同的波长筛选算法后训练集和预测集都能达到近 100%的正确率，由于 EMSC 较 MES 更复杂，所以同样准确率的情况下更加倾向于使用 MES 处理光谱。在理论上有同样处理效果的 SVN 的平均准确率只有接近 80%，这说明白酒基酒光谱差异不大，需要通过所有样本的平均光谱来校正误差。除了两个明显提高准确率的预处理算法外，还有部分预处理算法与适当的特征波筛选算法结合后可以较好地提高精度。Gap 导数、SGolay+[7]、基线+1D、基线+卷积、单位向量归一化、高斯滤波的预处理结合 CARS 后预测集的准确率分别为 97.49%、96.68%、96.68%、92.80%、92.34%、90.87%；SGolay+[7]、SGolay+[5]、SGolay+[3]、中值滤波、基线偏移+线性基线校正、基线+卷积的预处理结合 UVE 预测集的准确率分别为 95.54%、90.77%、92.00%、92.53%、90.84%、90.25%，均能满足酒企基酒分级需要。除 EMSC、MSC 外的预处理方法结合 PCA 后不能使得预测集的准确率高于 90%，因此目前选取的预处理算法不太适合与 PCA 结合。此外，由图 5 可知，不是所有预处理都能提高基酒分类准确率，以归一化为例，均值归一化后的数据，经过特征提取后的数据比原始数据都有所降低，而且经过单位向量归一化、面积归一化后有部分数据最后的分类准确率出现了下降。分析可知，基酒光谱经过处理后可能会使得特征光谱区被当成噪声处理后失去代表性，降低最后的分类准确率。

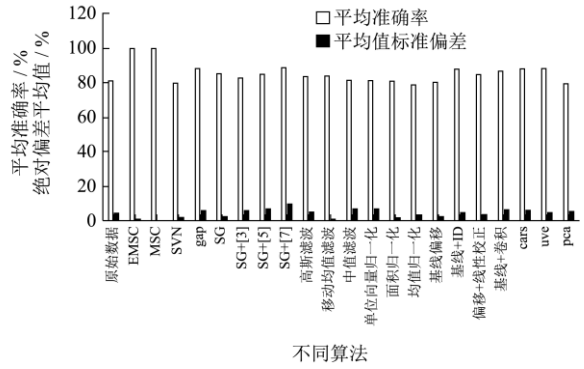

图 5 不同算法的平均值与绝对偏差平均值

Fig.5 Mean and absolute deviation mean of different algorithms

2.2.3 特征波筛选结果对比分析

在 CARS、UVE、PCA 三个特征波筛选算法中 CARS 和 UVE 能实现较好的特征波提取。三个特征波筛选算法选择后剩余的平均波数为分别为 69、578、22，通过选取的数据点可以看出，后期分类耗时 PCA＜CARS＜UVE。CRAS 最后的选取结果是以 RMSECV 最小为依据，忽略了选取的波数越多，RMSECV 更容易偏大，导致最后的选取结果中可能包含波数过少，不能完全表征光谱信息的情况，因此筛选结果差异较大，筛选稳定性不好。CARS 通过先通过指数衰减的两个阶段去粗筛和精筛变量，然后通过 ARS 以竞争的方式去剔除权重较小的数据，两次的筛选可以很大程度地剔除无用数据，使得后期分类速度与分类正确率都较高。UVE 的剔除后的剩余数据量占全光谱数据的 47.57%，剩余波长的数量相对其他两种算法来说最多，导致后期分类时间较长，但是也因此具有较好的、稳定的分类准确率。对于 PCA 来说，通过相互正交的几组特征向量来表示光谱特征，使得光谱特征被极大保留，后续数据量不大，在不同的预处理方法下分类结果比较稳定，但是 PCA 主要是针对线性数据降维，对于非线性的基酒光谱数据来说，这个过程中损失了部分重要特征，因此分类的过程中需要其他算法相互配合才能达到较好的分类准确率，例如结合预处理算法 EMSC、MSC，此时 PCA 降维后基酒分类准确率较高的同时还极大程度减少了输入 SVR 算法的数据量，极大提高了分类速度。

2.2.4 结果对比

目前，已经有很多学者结合近红外光谱技术对各种酒的质量检测以及品质分级技术进行了研究。以高畅[25]为例，他选取白酒基酒为研究对象，使用间隔偏最小二乘法（BiPLS）筛选特征波，偏最小二乘法（PLS）建立分析模型，得到决定系数为 93.37%，RMSEC 及 RMSEP 值分别为 1.72%、1.77%的总酯定量模型。朱宏霞[26]使用 4 种归一化、两种导数单独或组合处理的方法做预处理，利用主成分回归法（PCR）和 PLS 建立定标模型，最终得到最佳效果验证集决定系数 0.99，验证集预测偏差 0.13。陈林[27]利用近红外光谱与气象色谱结合标定基酒甲酸含量，校正集决定系数 99.25%，验证集决定系数 97.6%。在本实验中，通过对比 18 种预处理方式，选出最优预处理 EMSC、MSC，结合 CARS 后得到最高决定系数 99.04%，以及最低交叉验证均方根误差 8.35%；结合 UVE 后得到高相关系数 99.48%，最低 RMSEC 为 7.32%；结合 PCA 后得到最高 R_P 为 99.51%，最低 RMSECP 为 4.11%。由上述数据可以看出，本实验评估后选取的预处理与特征波筛选算法组合后的模型具有良好的性能。

3 结论

使用不同预处理方法以及特征波选取方法来将原始的高维光谱降成低维数据可以在很大程度上避免维度灾难，使得后面对数据进行回归处理时更加高效稳定，且具有一定得泛化性能。实验证明上述大部分预处理方法以及 2 种特征筛选方法可以提高白酒基酒分类准确率，实现白酒基酒分类的目的。实验表明，EMSC 和 MSC 可以很好地处理白酒基酒酒花散射对最后结果的影响，提高分类准确率，但是 MSC 算法更加简单。Gap 导数、Sgolay+[7]、基线+1D、基线+卷积、单位向量归一化、高斯滤波、Sgolay+[7]的预处理结合 CARS，SGolay+[7]、SGolay+[5]、SGolay+[3]、中值滤波、基线偏移+线性基线校正、基线+卷积的预处理结合 UVE 能使得预测集分类准确率于 90%，可以应用于实际的白酒分级过程，只有 EMSC、MSC 与 PCA 结合可以使得预测集实现较高分类。由于 PCA 只有结合特定算法可以实现快速高效分类，因此建立的模型泛化性较低，需要注意算法的搭配，可以在不断调试优化后投入实际应用，CRAS 的分类速度较快，可以考虑以后用于酒厂实时白酒摘酒，UVE 的分类效果稳定，可以考虑用于基酒入库时的基酒分类。

参考文献

[1] 江伟,韦杰,李宝生,等.不同原料酿造单粮白酒风味物质特异性分析[J].食品科学,2020,627(14):234-238

[2] 程伟,陈雪峰,陈兴杰,等.HS-SPME-GC-MS 结合感官评价分析金种子馥合香白酒的风味成分[J].食品与发酵工业,2022, 48(3):250-256,265

[3] 李利利,马宇,黄永光,等.酱香白酒机械化酿造不同基酒风味化合物解析[J].食品科学,2021,42(18):199-206

[4] 张群.浓香型白酒典型特征风味来源及其风味调控技术研

究[J].食品与生物技术学报,2020,39(4):112
[5] 余锴鑫.基于图像分类算法的自动化摘酒方法研究[D].杭州:浙江大学,2019
[6] 张旋,韩韬,颜廷才,等.浓香型原酒中塑化剂的吸附剂筛选[J].食品科学,2017,38(5):92-97
[7] 杨静娴,任小洪.基于图像处理的白酒酒花轮廓检测[J].食品与机械,2019,35(12):52-55,145
[8] 褚小立,陈瀑,李敬岩,等.近红外光谱分析技术的最新进展与展望[J].分析测试学报,2020,39(10):1181-1188
[9] 张进,胡芸,周罗雄,等.近红外光谱分析中的化学计量学算法研究新进展[J].分析测试学报,2020,39(10):1196-1203
[10] 褚小立,陈瀑,许育鹏,等.化学计量学方法在石油分析中的研究与应用进展[J].石油学报(石油加工),2017,33(6):1029-1038
[11] 李守学,陈玉艳,贾铮,等.饲料添加剂 L-赖氨酸硫酸盐中 L-赖氨酸含量近红外速测方法研究[J].动物营养学报,2017,29(10):3710-3717
[12] 郭阳,史勇,郭俊先,等.近红外光谱技术结合反向区间偏最小二乘算法-连续投影算法预测哈密瓜可溶性固形物含量[J].食品与发酵工业,2022,48(2):248-253
[13] 苗钧魁,张雅婷,刘小芳,等.近红外光谱技术在南极磷虾粉水分、脂肪和蛋白质含量快速检测中的应用[J].食品与发酵工业,2022,48(4):243-249
[14] 刘伟,何勇,吴斌,等.过程分析技术(PAT)在原料药生产中的应用[J].分析测试学报,2020,39(10):1239-1246
[15] 李跑,马雁军,马莉,等.基于近红外漫反射光谱和化学计量学方法的晒红烟常规化学指标的快速测定[J].湖南农业大学学报(自然科学版),2018,44(3):251-255
[16] 张辞海,胡芸,刘娜,等.烤烟烟碱近红外定量模型的适用性[J].烟草科技,2019,52(1):53-59
[17] 褚小立,史云颖,陈瀑,等.近五年我国近红外光谱分析技术研究与应用进展[J].分析测试学报,2019,38(5):603-611
[18] 胡耀强,郭敏,叶秀深,等.近红外光谱法间接测定白酒酒精度[J].光谱学与光谱分析,2022,42(2):410-414
[19] 孙彦华,范永涛.近红外光谱分析中温度影响的修正[J].光谱学与光谱分析,2020,40(6):1690-1695
[20] 第五鹏瑶,卞希慧,王姿方,等.光谱预处理方法选择研究[J].光谱学与光谱分析,2019,39(9):2800-2806
[21] 王动民,纪俊敏,高洪智.多元散射校正预处理波段对近红外光谱定标模型的影响[J].光谱学与光谱分析,2014,34(9):2387-2390
[22] Li H, Liang Y, Xu Q, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. Analytica Chimica Acta, 2009: 648(1): 77-84
[23] Centner V, Massartdl, De Noordoe, et al. Elimination of uninformative variables for multivariate calibration [J]. Analyti- cal Chemistry, 1996, 68(21): 3851-3858
[24] 许伟栋,赵忠盖.基于PCA-SVM算法的马铃薯形状分选[J].控制工程,2020,27(2):246-253
[25] 高畅,张宇飞,辛颖,等.近红外光谱技术结合波段筛选用于白酒基酒总酯定量分析[J].中国酿造,2021,40(4):155-158
[26] 朱宏霞,邓德文,郑校先.傅立叶变换近红外透射法测定黄酒酒精度[J].中国酿造,2008,12:80-82
[27] 陈林,庹先国,张贵宇,等.白酒基酒中甲酸的近红外预测模型构建[J].酿酒科技,2019,4:30-35