王怡博,庹先国,张贵宇,等. 近红外光谱结合改进的排列组合总体分析方法鉴别浓香型白酒基酒等级 [J]. 食品工业科技,2026,47(2):347−355. doi: 10.13386/j.issn1002-0306.2025020239
WANG Yibo, TUO Xianguo, ZHANG Guiyu, et al. Near-infrared Spectroscopy Combined with Improved Permutation Combination Population Analysis for Identifying Base Liquor Grades of Nongxiangxing Baijiu[J]. Science and Technology of Food Industry, 2026, 47(2): 347−355. (in Chinese with English abstract). doi: 10.13386/j.issn1002-0306.2025020239

·分析检测·

近红外光谱结合改进的排列组合总体分析方法鉴别浓香型白酒基酒等级

王怡博[1,2],庹先国[1,2],张贵宇[1,2,3,*],张　磊[1,2],向星睿[1,2],严　俊[1,2]
(1.四川轻化工大学人工智能四川省重点实验室,四川宜宾 644000;
2.四川轻化工大学自动化与信息工程学院,四川宜宾 644000;
3.酿酒生物技术及应用四川省重点实验室,四川宜宾 644000)

摘　要: 白酒基酒中含有种类繁多、关系复杂的微量成分，如何快速简便地预测其等级具有重要意义。为实现快速准确的鉴别，并考虑多组分样本物质间协调关系，降低计算复杂度，本文在近红外光谱数据基础上，提出了一种改进的排列组合总体分析（improved permutation combination population analysis，imPCPA）特征选择方法。研究基于浓香型白酒 4 种等级基酒共 687 个样品，建立了近红外光谱白酒基酒等级判别模型，采用了组合特征选择策略对光谱点进行筛选。先使用区间偏最小二乘法（interval partial least squares，IPLS）对预处理后的光谱进行无信息变量的剔除，再使用 imPCPA 对选取的区间进行波点筛选，最后建立极端梯度提升树（established extreme gradient boosting，XGBoost）分类模型。最终筛选出了 32 个特征波点，相比于改进之前的算法在计算量上减少了 80%，分类模型预测集准确率中位数达 95.65%。综上，该方法缓解了现有的感官品评方法进行酒类分析带来的主观性强、重现性差的特点，实现了多组分样本中成分间的协调作用的捕捉，特征选择结果解释性好，为白酒基酒等级的快速检测提供了参考。

关键词: 近红外光谱,白酒基酒,组合特征选择,排列组合总体分析,峰识别
中图分类号: TS262.3　　**文献标识码:** A　　**文章编号:** 1002−0306(2026)02−0347−09
DOI: 10.13386/j.issn1002-0306.2025020239
本文网刊:

Near-infrared Spectroscopy Combined with Improved Permutation Combination Population Analysis for Identifying Base Liquor Grades of Nongxiangxing Baijiu

WANG Yibo[1,2], TUO Xianguo[1,2], ZHANG Guiyu[1,2,3,*], ZHANG Lei[1,2], XIANG Xingrui[1,2], YAN Jun[1,2]
(1.Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644000, China;
2.School of Automation & Information Engineering, Sichuan University of Science & Engineering, Yibin 644000, China;
3.Liquor Making Biological Technology and Application of Key Laboratory of Sichuan Province, Yibin 644000, China)

Abstract: Base liquor of Baijiu contains a wide variety of trace components with complex interactions. How to predict its grade rapidly and simply was of great importance. To achieve rapid and accurate identification while accounting for interactions between multi-component substances and reducing computational complexity, this study developed an improved permutation combination population analysis (imPCPA) method for feature selection based on near-infrared

收稿日期: 2025−02−27
基金项目: 中国轻工业酿酒生物技术及智能制造重点实验室开放基金项目（2023-01）;四川轻化工大学科技成果转化专项项目（mes+）（HXJY01）;五粮液产学研合作项目（CXY2022ZR007）;四川轻化工大学“652”科研创新团队计划资助项目（SUSE652B005）。
作者简介: 王怡博（2000−）,男,硕士研究生,研究方向:智能酿造,E-mail:3532297440@qq.com。
*** 通信作者:** 张贵宇（1987−）,男,博士,副教授,研究方向:白酒自动化、人工智能,E-mail:gyz_118@163.com。

spectroscopy data. This study established a near-infrared spectroscopy (NIRS) model to discriminate the quality grades of base liquors in Nongxiangxing Baijiu. The model was developed using 687 samples from four quality grades. A combined feature selection strategy was applied to optimize spectral wavelength selection. The initial screening applied interval partial least squares (iPLS) to remove uninformative variables from preprocessed spectra. The improved permutation combination population analysis (imPCPA) further optimized wave point selection within the retained spectral intervals. Finally, constructed an extreme gradient boosting (XGBoost) classification model for grade prediction. The final selection identified 32 characteristic wavelength points. Compared to the original algorithm, the improved method reduced computational time by 80%. The median prediction accuracy of the classification model reached 95.65% on the prediction set. The results demonstrate that this method addresses key limitations of conventional sensory evaluation, including strong subjectivity and poor reproducibility in liquor analysis. It effectively captures synergistic interactions among multi-component substances while maintaining interpretable feature selection. The approach provides a reliable reference for rapid grade assessment of Baijiu base liquors.

Key words: near-infrared spectroscopy; Baijiu base liquor; hybrid feature selection; permutation combination population analysis (PCPA); spectral peak recognition

中国白酒是世界著名的六大蒸馏酒之一，主要是以高粱、小麦、大米等粮食作为原料，以酵母或曲类作为固体糖化发酵剂，经过固态糖化发酵，以固态蒸馏法生产[1]。白酒的酿造工艺独特，具有复杂的多组份体系，其中乙醇和水约占了总含量的 98%~99%，剩余的 1%~2% 则包含了上千种微量成分，这些成分的含量和相互作用共同决定了白酒的风味和口感[2]。基酒作为经蒸馏后还未经储存和勾兑的原酒，其质量等级对最终的成酒的质量有决定性作用[3]。目前基酒的质量评价尚无统一的标准评价体系，如何对基酒进行科学、准确的分类定级是当前急需解决的科学问题。

目前，基酒的等级评价已引进了多种分析方法，Ao 等[4] 利用电子舌结合气相色谱-质谱联用技术对高酒精度的基酒进行了分级，通过气相色谱-质谱联用技术分析不同等级的基酒之间化学成分的差异印证了电子舌分类结果的可靠性。He 等[5] 将程序升温顶空气相色谱-离子迁移谱应用于白酒蒸馏过程的表征，通过 3D 指纹图谱显示了不同蒸馏阶段的挥发性化合物的变化，实现了对不同蒸馏阶段白酒的分类。樊杉杉等[6] 采用 HS-SPME-Arrow 结合 GC-MS 对小曲清香型原酒挥发性成分进行定性和定量，分析不同等级原酒间挥发性成分的含量差异，对原酒样本进行了准确识别与分类。陈飞[7] 利用 GC 和 GC/O-TOFMS 对白酒基酒中微量成分的种类和含量进行了检测，并运用主成分分析、聚类分析、判别分析和神经网络建立模型对不同年份基酒样本进行分类鉴别。虽然这些方法有着较高的准确性，但都是通过对白酒中微量成分含量进行检测来间接评价白酒等级，需要对待测样品进行化学制备并检测含量，会造成样本污染，同时分析速度较慢、成本高，难以在生产中大规模应用。在生产中，基酒的分级主要依赖于人工感官品评，该方法虽方便快捷，但对专业人员的要求高，易受主观影响，缺乏严密性、标准性和重现性差[8]。而近红外光谱技术作为一种快速、无损无污染、无需样本制备、重现性好的检测技术[9]，被认为是最适合在生产中大规模应用的多组分检测方法之一，具有广阔的应用潜力。

近年来，近红外光谱技术已被应用于白酒等级鉴别。张维等[10] 将近红外光谱数据与气相色谱质谱数据相结合，通过对有效特征信息进行数据融合，实现了基酒等级的准确鉴别。Wu 等[11] 使用气相色谱质谱对基酒显著风味物质进行定量检测，再结合近红外光谱数据实现白酒品质评估。Zhong 等[12] 将集成移动窗口波段筛选的 PLS-DA 应用于白酒品牌判别分析，提出了包含五类波段的简化最优模型。宗绪岩等[13] 得到白酒基酒中典型醇和酯的光谱吸收波段在 4833~6846 cm^{-1}，通过此波段分别建立年份、等级模型。从上述研究中可以看出，近红外技术在多组分样本检测中体现出了良好的适用性，但同时这些研究多借助对典型风味物质量的检测，再结合近红外光谱实现对白酒品质的鉴别。虽然使模型的应用更加便捷，但在构建模型时仍需要进行复杂实验定量。完全使用近红外光谱实现白酒品质评价研究较少，这是由于光谱数据具有高维、小样本的特点，以及多组分样本中成分间的协调作用为提取有效的光谱特征带来了困难。

针对上述问题，本研究提出了一种改进的排列组合总体分析（improvement permutation combination population analysis, imPCPA）特征选择方法。该方法基于 Geng 等[14] 提出的 PCPA 方法改进而来，属于模型总体分析（model performance analysis, MPA）框架下的方法。MPA 类方法及其变体已在定性分析上取得了良好效果[15−18]，在理论上可以更好地衡量多组分样本中成分间的协调作用。imPCPA 方法用于实现等级分类尚属首次，为更好结合白酒样本特点，采用区间偏最小二乘法（IPLS）结合 imPCPA 实现对有效特征的精确筛选，通过改进在保持精度的情况下实现计算速度的提升。然后建立极端梯度提升树（established extreme gradient boosting, XGBoost）分类模型，对选择效果进行评价，以期为快速、合理的评价基酒等级提供一定的解决思路。

1　材料与方法

1.1　材料与仪器

酒样　采集于中国四川某酒企，样品均源自同一生产批次相同车间的不同浓香型窖池，为确保样品能够反映实际摘酒情况，样本由具有 10 年以上摘酒经验的技术人员根据酒花变化摘酒得来，共采集 687 个样本。

MATRIX-F 傅里叶变换近红外光谱仪　德国 Bruker 公司。

1.2　实验方法

1.2.1　近红外光谱采集　使用光谱仪自带的透射型光纤探头进行检测，检测前光谱仪预热 1 h，检测参数为：环境温度 20±2 ℃；空气相对湿度<80%；光谱波数范围 12500~4000 cm^{-1}，光纤长度 2.5 m，扫描频率 10 kHz，光谱分辨率 4 cm^{-1}。将基酒样品（5 mL）放入样品瓶中，每个样品扫描 32 次，自动获得平均光谱。每个样品的检测时间约为 18 s。由于波数范围在 12500~9025 cm^{-1} 的光谱几乎无化学键吸收，在 4200~4000 cm^{-1} 内的光谱受仪器和环境影响杂乱无规律，因此只选取 9025~4200 cm^{-1} 内的 1215 个光谱数据进行后续分析[10,19]。

1.2.2　酒样品评　为了避免储存对样品的影响，同步进行红外光谱采集和专业评酒师品评。所得样品由酒厂品酒小组根据 GB/T 10345-2022《白酒分析方法》、GB/T 33404-2016《白酒感官品评导则》结合酒厂质量评价标准对基酒进行综合评价。采得酒样共分四个等级，其中一级酒 139 个，二级酒 282 个，三级酒 166 个，四级酒 100 个。

1.3　数据处理

1.3.1　数据预处理与数据集划分　采集获得的近红外光谱包含有采集环境和设备引起的基线漂移和噪声，本研究使用自适应迭代重加权惩罚最小二乘法（adaptive iterative reweighted penalized least squares，airPLS）对光谱基线进行校正，使用 Savitsky-Golay 卷积平滑（SG）过滤高频噪声。预处理后使用基于联合 X 和 Y 距离的样本集划分（sample partitioning based on joint x-y distance，SPXY）划分训练集与测试集，通过在预处理的训练集上训练预测模型，根据分类准确率选择最优的预处理方法。

airPLS 是由 Zhang 等[20] 提出的一种基线校正算法。该算法使用惩罚最小二乘法拟合数据以分离背景信号，不需要任何的人为干预和先验信息（例如峰值检测等）。通过对拟合残差加权和自适应的迭代权重与基线，拟合出原始光谱的背景信号，再利用原始光谱与之相减，得到有效的光谱数据。该方法有着自适应性强、鲁棒性好、参数少的优点。

SG 卷积平滑是使用多项式对预设好大小的移动窗口内的数据进行最小二乘拟合，以减少平滑对有用信息的影响。通常其中多项式的次数与平滑效果成反比，平滑窗口宽度与平滑效果成正比，设定合适的多项式次数和窗口宽度能有效降低光谱中的噪声[21]。

SPXY 方法考虑了 X 和 Y 的联合分布，通过计算两者的联合距离（通常为两者的归一化欧氏距离）进行选样。通过逐步选样确保样本代表性，采用类似于优化的“最大最小距离”策略，保证整个样本空间均匀覆盖。该方法选出的训练集可以增强模型泛化能力，防止过拟合[22]。本研究中以训练集比测试集为 4∶1 的比例划分数据。

1.3.2　特征选择方法　近红外光谱存在着高维，共线性的特点，利用全光谱建立模型会使得计算量较大，还会引入无信息变量干扰模型评价，因此需要对光谱特征进行选择以提高模型性能。使用 IPLS 作为第一步，可用较小的计算负担快速地剔除无效特征，再使用 imPCPA 进行选择。imPCPA 考虑了特征间的组合作用，适合成分间协调作用复杂的基酒光谱，且相较于原 PCPA 方法计算量更小。通过两种方法的联用实现了准确度与计算量的平衡，体现了组合特征选择法的优势。

1.3.2.1　区间偏最小二乘法　IPLS 将整个光谱分为一系列等宽区间，再对每个区间建立预测模型，根据各个区间的预测误差选择需要的区间。该方法没有考虑区间间关系，同时区间大小的选择对于结果的影响也十分显著，单独使用该方法通常难以得到满意的结果。但该方法可以快速评价各区间重要性，选出无有效信息的区间，降低光谱数据的复杂度以方便进一步的特征提取。

1.3.2.2　排列组合总体分析　PCPA 算法[14] 先利用二进制矩阵采样（binary matrix sampling，BMS）生成大量子模型，计算每一个子模型的交叉验证均方根误差（root mean square error of cross-validation，RMSECV）选出最优子集组；再计算每个子集中每个波点在单独抽出后该子集 RMSECV 的变化，将抽出后和抽出前的子集 RMSECV 差值作为该波点在该子集中的重要度，将该波点在子集组中的重要度之和作为该波点的总重要度并进行排序；最后根据指数递减函数（exponential decay function，EDF）迭代的选择不同数量的波点，分别比较每次迭代选择的波点组建模效果，选出有最好模型效果的波点组。

该方法既考虑了变量间的组合作用对模型准确率的影响，也避免了仅使用频次分析计算波点重要度带来的无信息变量的污染。但该方法在使用中有着以下缺陷：使用 EDF 函数迭代选择波点，在只有 700 个波点时需要迭代近 50 次才可获得较优的效果。计算量大耗时长，同时选择的波点数量固定，不一定为最优特征组合；在 BMS 抽样方法中选定的特征出现在子集组中的概率是等同的，这增加了无信息变量混入最优子集组的概率，虽然可以通过后续的重要性度量步骤来剔除，但增加了无谓的计算量。因此，本研究针对这些问题提出了 imPCPA 方法。

1.3.2.3 改进的排列组合总体分析 与 PCPA 相比 imPCPA 作了三方面优化：采用加权引导采样（weighted bootstrapped sampling, WBS）代替 BMS 方法，该方法可以自动缩小所选择特征的数量，加快迭代速度减小计算量[23]，同时加权过程可与特征重要性度量方法相结合，快速剔除无信息变量，减少子集模型污染；通过评价最优子集组的 RMSECV 均值实现迭代的自停，减少计算量；优化了特征波点选择方法，根据重要度排序选择最优子集组采样数量个波点，再使用后向删除法进一步选优。具体步骤如下：

a.设置特征的权重（初始权重设为 1），使用 WBS 生成 N 个子集。

b.通过 PLS 结合五折交叉验证计算每个子集的 RMSECV 值。

c.根据每个子集的 RMSECV 进行排序，并选择最佳的α% 子集，并计算这些子集 RMSECV 的平均值，记为$RMSECV_{mean}$。

d.计算每个变量在每个最佳子集中的重要度，计算方法同 PCPA。

e.计算每个变量在α% 子集中的重要度之和记作该变量在当前抽样轮次中的总重要度。根据总重要度为各变量赋予新权重，如果有变量的重要度为负数则将其置 0。

f.如果$RMSECV_{mean}$的值相较于上一次的值减小，则重复步骤 a~e 继续循环，否则终止循环。

g.根据$RMSECV_{mean}$最小的一次循环计算出的变量重要度对变量进行降序排序，从高到低选取该次循环抽取样本数的变量作为基础变量组合，再使用后向删除法选择最优的波长组合。

1.3.3 建模方法 XGBoost 是由 Chen 等[24]基于梯度提升树的思想改进的算法，通过多棵分类与回归树结果的累加和得到最终的预测结果。该模型通过不断迭代产生新的树，每次迭代产生的树可拟合上一棵树预测的残差，迭代多次进而形成一个由多个弱分类器组成的强分类器。假设有n个样本，训练了m次，则第m棵树上的第i个样本的预测值可如式（1）表示：

$$\widehat{y}_i = \sum_{i=1}^{m} f_m(x_i), f_m \in F \qquad 式（1）$$

式中，x_i表示第i个样本的特征数据；$f_m(x_i)$表示第m棵树对第i个样本的预测函数；F表示预测函数的集合。此时目标函数可由式（2）表示为：

$$obj = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \qquad 式（2）$$

式中，y_i表示第i个样本的真实类别；$\sum_{i=1}^{n} l(y_i, \widehat{y}_i)$为损失函数；$\sum_{m=1}^{M} \Omega(f_m)$为惩罚项，用于控制模型复杂度防止过拟合，其计算公式见式（3）：

$$\Omega(f_m) = \gamma T + \frac{1}{2}\lambda \sum_{m=1}^{T} \omega_j^2 \qquad 式（3）$$

式中，T表示树结构中的总节点数；γ表示子叶权重；ω_j表示模型对子节点给出的分数值；λ表示子叶节点分数权重，用于防止分数过大导致过拟合。在本研究中使用的 XGB 参数如下表 1 所示。

表 1 XBG 在本研究中使用的参数设置

Table 1 Parameter settings used in XBG in this study

参数项	参数大小
学习率（Learning_rate）	0.1
最大树深度（Max_depth）	5
样本选择比例（Sub_sample）	0.8
特征选择比例（Clsample_bytree）	0.8
损失函数（Objective）	Multi:softprob
类别数目（Num_class）	4
迭代次数（Num_boosting_rounds）	100

1.4 模型评价指标与分析软件

为评估模型性能，本研究采用预测集分类准确率作为主要评价指标。实验采用 VS code 2024、The Unscrambler X 10.4 和 Origin 2024 进行数据处理，其中原始光谱的预处理在 The Unscrambler X 10.4 软件中进行，训练集和测试集的划分、特征波长筛选、建立 XGBoost 分类模型在 VS code 2024 中使用 Python 3.9.0 进行，图表的绘制在 Origin 2024 软件中进行。

2 结果与分析

2.1 近红外光谱预处理结果

预处理主要是针对近红外光谱存在的基线漂移和毛刺明显的情况，先使用 airPLS 进行基线校正，再

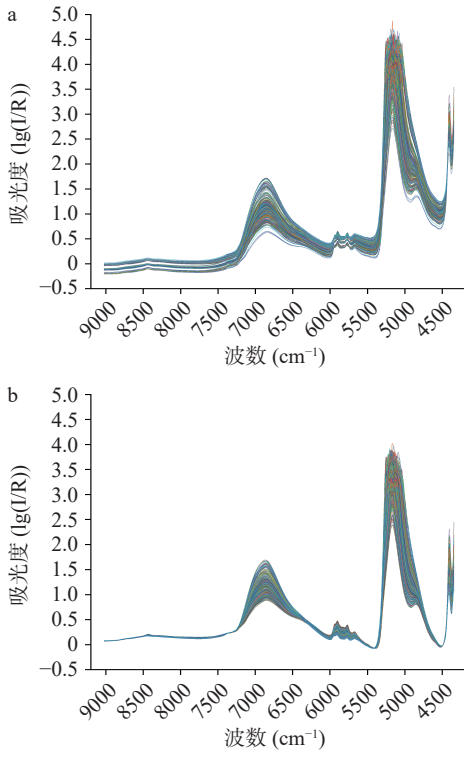


图 1 airPLS 结合 SG 卷积平滑预处理前后光谱对比

Fig.1 Spectral comparison before and after airPLS combined with SG convolution smoothing pretreatment

注：a. 原始光谱；b. airPLS+SG（11-2）处理后的光谱。

使用 SG 卷积平滑对光谱进行预处理，通过比较构建的 XGB 模型的准确度来确定最佳的预处理方式。airPLS 结合 SG 卷积平滑处理前后近红外光谱图如图 1 所示，建立模型的预测结果如表 2 所示，其中 airPLS 的惩罚项λ设为 10^7，SG(n-m)为 n 点 m 次卷积平滑。

表 2　不同光谱预处理方法的 XGBoost 模型预测结果
Table 2　XGBoost model prediction results of different spectral preprocessing methods

预处理	训练集准确率(%)	测试集准确率(%)
无(原始光谱)	93.26	91.24
airPLS	95.26	92.75
airPLS-SG(3-2)	95.62	93.47
airPLS-SG(5-2)	95.81	94.20
airPLS-SG(7-2)	96.17	94.20
airPLS-SG(9-2)	96.53	94.92
airPLS-SG(11-2)	97.99	95.65

图 1a 中可以看出原始近红外光谱有着较为明显的基线漂移现象，在图 1b 中明显减弱。对使用不同参数预处理后的光谱数据建立 XGB 分类模型，预测集准确率如表 2 所示。不同数据(原始光谱、airPLS、airPLS-SG(3-2)、airPLS-SG(5-2)、airPLS-SG(7-2)、airPLS-SG(9-2)、airPLS-SG(11-2))的准确率分别为：91.24%、92.75%、93.47%、94.20%、94.20%、94.92% 和 95.62%。其中 airPLS 结合 SG(11-2)的分类效果最好，在原始数据的基础上提升了 4.38%，因此采用该方法处理后的光谱数据作为后续算法的输入数据。

2.2　IPLS 无效特征剔除

将 airPLS 结合 SG(11-2)处理后的基酒近红外全光谱 1215 个波点均匀地划分为 45 个波段，每个波段包含有 27 个波点。对每个区间进行编号，通过 5 折交叉验证得出每个区间的 RMSECV 并进行排序，通过从低到高以 5 为间隔选取区间进行组合，检验每个组合的 XGB 模型预测效果，不同区间选择个数的预测效果如表 3。

表 3　不同光谱预处理方法的 XGBoost 模型预测结果
Table 3　XGBoost model prediction results of different spectral preprocessing methods

指标	区间数					
	40	35	30	25	20	15
训练集准确率(%)	97.99	97.44	97.26	97.26	95.81	93.80
测试集准确率(%)	95.65	95.65	95.65	95.65	93.47	90.57

对比选择不同区间数时的模型预测精度，发现在保留多于 25 个区间的情况下时，模型准确率几乎保持 95.65% 不变，这表明在原始光谱中确有许多无信息变量。当区间数量减少到 15 个时准确率开始快速下降，这表明该数据用于等级分类的特征在光谱的分布较为分散，只选择几个区间难以准确表征数据特点，需要进行进一步筛选。为尽可能防止误删有效信息，不再进一步优化选择区间个数。在后续步骤中均采用保留 25 个区间的光谱数据作为输入数据，区间位置如图 2 所示。

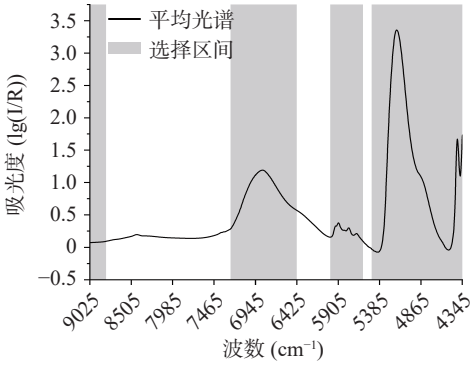


图 2　IPLS 选择区间分布
Fig.2　IPLS selection interval distribution

2.3　imPCPA 参数选择

在 imPCPA 中有两个参数会显著影响特征提取的性能，需要对它们的取值进行优化。

N，这个参数代表了每次迭代中 WBS 的采样次数。在 VCPA 及其变体中[14,25−26]讨论了 500、1000、2000 等值，区间组合优化(interval combination optimization, ICO)中[23]讨论了 100、500、1000 等值。参考前述研究，在本研究中考虑到计算成本，将 N 设置为 1000。

α，最佳子模型的比例。较大的取值将会导致重要变量的重要度被稀释，而无用变量出现在子集中的概率将增加，影响模型效果并增加计算量；较小的取值则会无法全面的评估变量组合对模型的影响。在本研究中分别讨论了α取 2%、5%、10%、20% 四种情况，通过比较在四种参数下进行 20 次特征提取的最小$RMSECV_{mean}$的分布选择最佳参数，其结果如图 3 所示，当α取 2% 时$RMSECV_{mean}$值的中位数最小为 0.5085，同时重复试验无异常值结果出现，故后续研究中设置α=2%。

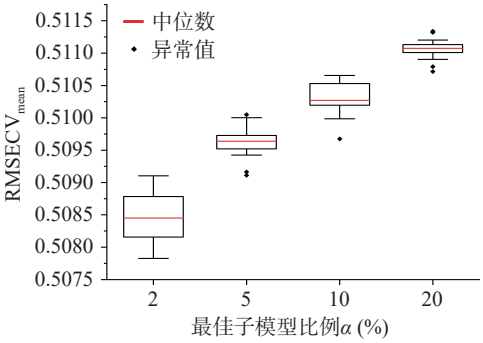


图 3　参数α% 对模型的影响
Fig.3　Effect of parameter α% on the model

2.4　特征选择结果分析

在特征选择方法对比实验中，为保证输入数据的一致性与可比性，统一采用通过 IPLS 方法筛选出

的 25 个光谱区间作为特征输入，分别应用不同特征选择策略进行建模和性能评估。除使用原始 PCPA 算法与改进后的算法作对比，还选用了较为常用的 CARS 和无变量信息消除法（uninformative variables elimination, UVE）以及同属 MPA 框架下的迭代保留信息变量（iteration retention information variable, IRIV）、变量组合种群分析（variable combination population analysis, VCPA）四种方法作对比。鉴于单次实验结果可能受上述算法的随机性影响，为保证评估的稳定性与客观性，本研究采用 20 次重复实验，并以分类准确率的中位数作为统计指标。相比平均值，中位数更能抵抗偶发异常结果的干扰，真实反映模型在不同等级基酒判别任务中的典型性能。各算法特征选择后建模效果如图 4 所示，选择变量数量及特征提取过程平均单次运行时间如表 4 所示。

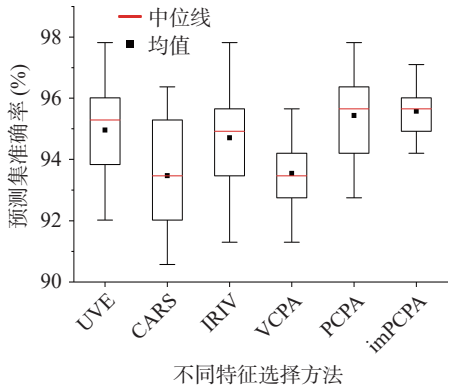


图 4 不同特征选择方法重复预测结果

Fig.4 Repeated prediction results of different feature selection methods

表 4 不同特征选择方法整体模型效果

Table 4 Overall model effect of different feature selection methods

特征选择方法	选择特征数	最优预测效果(%)	最优效果最小特征数	最接近中位数预测效果(%)	中位数效果最小变量数	平均运行时间(s)
UVE	71~334	97.82	71	95.65	110	40
CARS	9~49	96.37	22	93.47	12	38
IRIV	28~42	97.82	34	94.92	31	7535
VCPA	19~71	95.65	48	93.47	21	62
PCPA	35~65	97.82	45	95.65	40	2646
imPCPA	28~42	97.10	38	95.65	32	528

由图 4 和表 4 可以看出，除 CARS 和 VCPA 两种方法外，其余方法的预测结果中位数（UVE：95.65%、IRIV：94.92%、PCPA：95.65%、imPCPA：95.65%）均与特征选择前的预测效果（95.65%）十分接近，这表明这四种方法可以较好的保留样本特征，而 CARS 追求最小化特征数，在一定程度上剔除了有效信息，同时没有考虑特征间的协同作用，导致虽然有着最快的运行速度和最小的特征选择量，但在预测结果上准确率中位数 93.47% 是所有方法中最低的。VCPA 虽然同为 MPA 框架下的特征选择算法，但它采用的是统计特征在最优子集中的出现频次作为特征的对应重要性，这种方法虽然极大地加快了运算效率，但也导致了无信息特征和干扰特征的保留，以至于准确率中位数 93.47% 明显小于 IPLS 提取后的 95.65%。UVE 虽有较好预测效果（97.82%）和很快的运行速度（40 s），也是选择特征数最多和波动最大的。该方法通过叠加噪声以剔除和噪声相当的变量来实现特征筛选，结果受构建的噪声影响较大，提取的特征数量很不稳定。IRIV 和 PCPA 方法都表现出了较稳定的提取特征数量和预测效果，但在时间复杂度上两种算法都很高，虽然 PCPA 通过改变重要性度量方法实现了缩减计算量，提升预测性能，但运行时间仍然达到了千秒级别。相比之下 imPCPA 进行了抽样方法的改良，相比于原方法使用 BMS 抽样结合 EDF 函数的方式，需要先完成 EDF 函数迭代，再进行最优特征组合选择。改进方法使用 WBS 抽样结合重要性度量方法，使得迭代次数大幅减少，同时在迭代中寻优自停，进一步缩小计算量，相较于原方法的 2646 s 缩小了约 80% 只需 528 s 就可完成计算。虽然相比于原方法构建更少的特征组合导致最大准确率有所下降，但整体上准确率保持了较好效果，准确率中位数与 IPLS 提取后的光谱预测结果持平。同时也减轻了随机采样带来的随机性，预测集结果的分布也更加集中。在特征数量上取特征平均数 35，只使用了约原始光谱 1215 个特征的 3%，IPLS 光谱的 5%，很好的实现了特征选择的目的。

为进一步分析不同特征选择方法在不同等级基酒上的分类效果，以及样本不平衡可能带来的影响，以全面反应特征选择方法的性能。选取了不同特征选择方法多次试验中预测结果最接近中位数的特征子集，选择其中特征点最少的子集建立的分类模型进行分析，各方法对应的各等级基酒分类准确率如表 5 所示，对应混淆矩阵如图 5 所示。整体上，所有方法都对四级酒有着极高的识别准确率，从混淆矩阵中可以看出没有出现将其他类别大量错误预测为四级酒的现象，这可以说明四级酒的高识别率并非模型对小样本类别过拟合导致的。其高识别率可能是因为四级酒作为摘酒过程中的尾酒，有着较为明显的酒体特征，特征空间边界明显，易于识别。imPCPA 方法对其余三个等级的基酒分类准确率都保持在 93% 以上，各类别指标相对均衡，横向对比其他方法分类效

表 5 不同特征选择方法对各等级基酒的分类效果

Table 5 Classification effect of different feature selection methods on each grade of base liquor

特征选择方法	一级酒准确率(%)	二级酒准确率(%)	三级酒准确率(%)	四级酒准确率(%)
UVE	93.1	92.59	100	100
CARS	93.1	94.44	90.91	95.45
IRIV	93.1	94.44	93.94	100
VCPA	89.66	92.59	93.94	100
PCPA	96.55	94.44	96.97	95.45
imPCPA	93.1	94.44	96.97	100

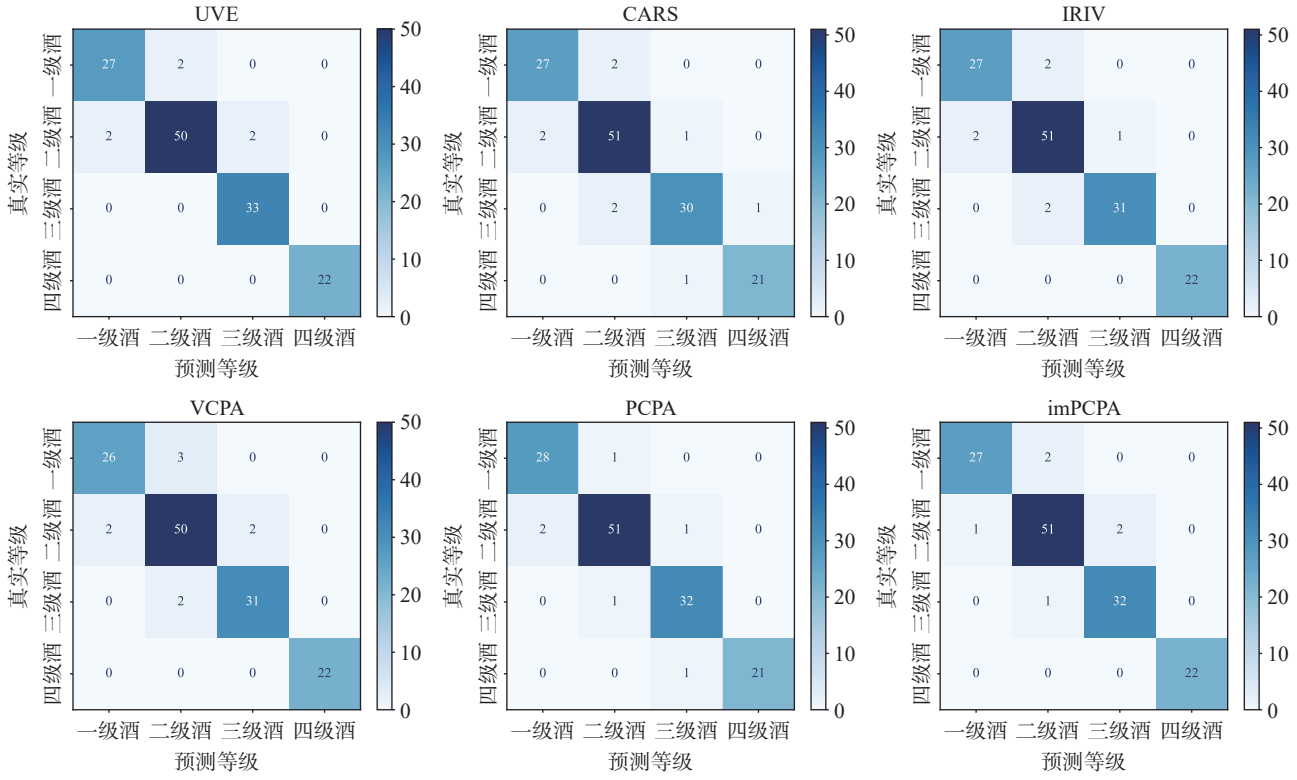


图 5　不同特征选择方法的预测模型混淆矩阵

Fig.5　Prediction model confusion matrix of different feature selection methods

果均处于高位水平，整体保持较高的判别能力，可认为分类效果良好。从混淆矩阵可以看出错误分类多集中在对二级酒的误判，表明二级酒在特征空间上与一、三级酒存在部分重合，有待在后续工作中进一步针对性的优化特征选择与分类算法。

尽管各类别样本数量存在一定程度的不均衡，但从分类模型混淆矩阵可以看出，模型对小样本类别（四级酒）仍保持较高的识别准确率（95% 以上），且未出现大类别样本大量误判为小类别的现象。各类别的召回率大都超过 90%，整体准确率维持在 93% 以上，说明样本不均衡对模型性能的影响较小。

选取不同特征选择方法多次试验中预测结果最接近中位数的特征子集，选择其中特征点最少的子集对其在光谱上的分布进行分析，其结果如图 6 所示。六种方法所选择的特征有着几处较为明显的重合，分别为 6000~5700 cm^{-1}，该波段被认为是甲基的一个重要吸收区间，与白酒中的正戊醇、乙醇相关[27−28]；5400~5300 cm^{-1}，该波段则被认为是醛类物质的 C-H 键吸收区[29]；4800~4700 cm^{-1}，该波段则是 C=O 键的吸收区间，与乙醇、甲酯类和羧酸类物质的表征相关[30−31]。这几种物质在白酒基酒的品质评价中都有着重要作用，可以明显看到 UVE 在这三个波段连续的选择了大量的波点，其中既有方法本身的特点，也从侧面印证了这几个区间的重要性。IRIV、VCPA、PCPA 和 imPCPA 这四种基于 MPA 框架的方法还集中选择了 7240~7140 cm^{-1}、6550~6450 cm^{-1} 两个区间，前者是甲基的组合频吸收峰[32]，后者是聚

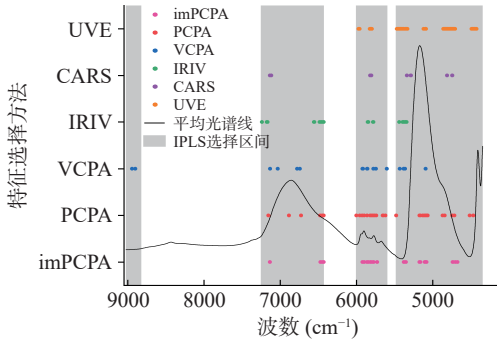


图 6　不同特征选择方法选择波长分布

Fig.6　Different feature selection methods select the wavelength distribution

合醇的 O-H 键吸收峰[33]。这两个区间都不只反应某一类物质，而是多类物质的共同表征区间，对这两个区间的选择从一定程度上也反映了 MPA框架方法对于物质间相互关系的考虑。imPCPA 还选择了 5180~5050 cm^{-1} 这一区间，在该区间醛和酮的羰基二级倍频都在此处表征，乙酸乙酯的羰基二级倍频吸收峰也在此处[32]，该物质作为白酒基酒的重要风味物质很大程度上影响了基酒的等级[34]。

从整体上来看，imPCPA 相对于其他几种特征选择方法更有优势。在预测准确率上，该方法的准确率中位数有着最好的表现，与之有着同样表现的 PCPA 则在计算复杂度上远大于 imPCPA。作为准确率和计算速度都表现优秀的 UVE 则有着最不稳定的特征提取数量，平均数量也是最多的，不能很好地选出最有代表性的特征点。速度最快的 CARS 则预测效

果整体偏低，不能很好地的评价白酒基酒等级。从选择的特征点对应的物质表征也可以看出，imPCPA 所选特征与物质联系紧密，也反映了对于物质间联系的考虑。因此可以认为 imPCPA 在保证准确率的情况下大幅减小了原算法的计算复杂度，很好地实现了对于多组分样本中成分间的协调作用为提取，以及对于高维光谱数据的降维。

3 结论

针对白酒基酒近红外光谱高维、小样本的特点，以及成分间的协调作用为提取有效的光谱特征带来的困难，本文探索了一种改进的 PCPA 特征选择方法，结合 IPLS 建立混合特征选择方法，建立了基于近红外光谱的白酒基酒等级检测模型，并与其他 5 种特征选择方法进行比较。结果表明，基于改进的 PCPA 方法在预测准确率上取得了良好的稳定性和最高的中位数准确率 95.65%，同时相比原算法计算复杂度降低了 80%。准确率接近中位数的最小特征数为 32，占全光谱的 2.6%。通过近红外光谱吸收谱带对特征波长进行分析，验证了特征波长与物质间的联系，同时也表明了方法对于物质间相关性的考虑。研究结果表明，该方法能够有效判别白酒基酒等级，具有较好的解释性，为白酒生产的在线检测提供了思路。

© The Author(s) 2026. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

参考文献

[1] QIAN Wei, LU Zhenming, CHAI Lijuan, et al. Cooperation within the microbial consortia of fermented grains and pit mud drives organic acid synthesis in strong-flavor Baijiu production[J]. Food Research International, 2021, 147: 110449.

[2] HONG Jiaxin, ZHAO Dongrui, SUN Baoguo. Research progress on the profile of trace components in Baijiu[J]. Food Reviews International, 2023, 39(3): 1666−1693.

[3] 韩强, 刘鑫, 周永帅, 等. 勾调自动化在白酒生产中的应用研究进展[J]. 食品与机械, 2021, 37(12): 215−220,231. [HAN Q, LIU X, ZHOU Y S, et al. Research progress on the application of blending automation in the production of Baijiu[J]. Food & Machinery, 2021, 37(12): 215−220,231.]

[4] AO Ling, GUO Kai, DAI Xinran, et al. Quick classification of strong-aroma types of base Baijiu using potentiometric and voltammetric electronic tongue combined with chemometric techniques[J]. Frontiers in Nutrition, 2022, 9: 977929.

[5] HE Fei, DUAN Jiawen, ZHAO Jiwen, et al. Different distillation stages Baijiu classification by temperature-programmed headspace-gas chromatography-ion mobility spectrometry and gas chromatography-olfactometry-mass spectrometry combined with chemometric strategies[J]. Food Chemistry, 2021, 365: 130430.

[6] 樊杉杉, 唐洁, 乐细选, 等. 基于 HS-SPME-Arrow-GC-MS 和化学计量学的小曲清香型原酒等级判别[J]. 食品与发酵工业, 2021, 47(13): 254−260. [FAN S S, TANG J, YUE X X, et al. Grade discrimination of Xiaoqu light aroma Baijiu based on HS-SPME-Arrow-GC-MS and chemometrics[J]. Food and Fermentation Industries, 2021, 47(13): 254−260.]

[7] 陈飞. 基于微量组分的白酒基酒分类鉴别与模式识别研究[D]. 重庆: 重庆大学, 2019. [CHEN F. Study on classification identification and pattern recognition of base liquors based on trace components[D]. Chongqing: Chongqing University, 2019.]

[8] LI Huanhuan, ZHANG Xin, GAO Xiaojuan, et al. Comparison of the aroma-active compounds and sensory characteristics of different grades of light-flavor Baijiu[J]. Foods, 2023, 12(6): 1238.

[9] LUO Wei, TIAN Peng, FAN Guozhu, et al. Non-destructive determination of four tea polyphenols in fresh tea using visible and near-infrared spectroscopy[J]. Infrared Physics & Technology, 2022, 123: 104037.

[10] 张维, 张贵宇, 庹先国, 等. 基于 NIR 和 GC-MS 融合技术的浓香型白酒原酒等级鉴别[J]. 食品科学, 2024, 45(21): 288−296. [ZHANG W, ZHANG G Y, TUO X G, et al. Grade identification of raw Nongxiangxing Baijiu based on fused data of near infrared spectroscopy and gas chromatography-mass spectrometry[J]. Food Science, 2024, 45(21): 288−296.]

[11] WU Jianhang, PENG Houbo, LI Li, et al. FT-IR combined with chemometrics in the quality evaluation of Nongxiangxing Baijiu[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 284: 121790.

[12] ZHONG Jie, CHEN Jiemei, YAO Lijun, et al. Discriminant analysis of liquor brands based on moving-window waveband screening using near-infrared spectroscopy[J]. American Journal of Analytical Chemistry, 2018, 9(3): 124−133.

[13] 宗绪岩, 彭厚博, 吴键航, 等. 化学计量学结合 NIR 对浓香型白酒年份、等级的研究[J]. 包装与食品机械, 2022, 40(2): 87−94. [ZONG X Y, PENG H B, WU J H, et al. Study on the year and grade of Luzhou-flavor liquor by chemometrics and NIR[J]. Packaging and Food Machinery, 2022, 40(2): 87−94.]

[14] GENG Jingxuan, YANG Chunhua, LUO Qiwu, et al. iPCPA: Interval permutation combination population analysis for spectral wavelength selection[J]. Analytica Chimica Acta, 2021, 1171: 338635.

[15] 马玲, 马倩, 李亚娇, 等. 基于高光谱成像技术的不同时期番茄植株的快速无损判别研究[J]. 分析测试学报, 2023, 42(2): 210−215. [MA L, MA Q, LI Y J, et al. Rapid and nondestructive identification of tomato plants in different periods based on hyperspectral imaging technique[J]. Journal of Instrumental Analysis, 2023, 42(2): 210−215.]

[16] 李杨, 李翠玲, 王秀, 等. 高光谱成像的黄瓜病虫害识别和特征波长提取方法[J]. 光谱学与光谱分析, 2024, 44(2): 301−309. [LI Y, LI C L, WANG X, et al. Identification of cucumber disease and insect pest based on hyperspectral imaging[J]. Spectroscopy and Spectral Analysis, 2024, 44(2): 301−309.]

[17] 崔程, 刘翠玲, 孙晓荣, 等. 基于近红外高光谱成像技术的花生冻伤检测方法研究[J]. 食品工业科技, 2024, 45(6): 226−233. [CUI C, LIU C L, SUN X R, et al. Peanut frostbite detection method based on near infrared hyperspectral imaging technology[J]. Science and Technology of Food Industry, 2024, 45(6): 226−233.]

[18] 王静, 马玲, 马思艳, 等. 荧光高光谱成像技术对甜瓜叶片过氧化氢酶活性的无损检测[J]. 光谱学与光谱分析, 2024, 44(12): 3455−3462. [WANG J, MA L, MA S Y, et al. Nondestructive detection of catalase activity in melon leaves by fluorescence hyperspectral imagery[J]. Spectroscopy and Spectral Analysis, 2024, 44(12): 3455−3462.]

[19] ZHANG Guiyu, TUO Xianguo, PENG Yingjie, et al. A rapid

nondestructive detection method for liquor quality analysis using NIR spectroscopy and pattern recognition[J]. Applied Sciences, 2024, 14(11): 4392.
[20] ZHANG Zhimin, CHEN Shan, LIANG Yizeng. Baseline correction using adaptive iteratively reweighted penalized least squares [J]. The Analyst, 2010, 135(5): 1138.
[21] 刘玉娟, 刘颜达, 宋莹, 等. 近红外光谱的甲醇汽油定量检测与分析方法[J]. 光谱学与光谱分析, 2023, 43(5): 1489−1494. [LIU Y J, LIU Y D, SONG Y, et al. Near infrared spectroscopic quantitative detection and analysis method of methanol gasoline[J]. Spectroscopy and Spectral Analysis, 2023, 43(5): 1489−1494.]
[22] GALVAO R, ARAUJO M, JOSE G, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736−740.
[23] SONG Xiangzhong, HUANG Yue, YAN Hong, et al. A novel algorithm for spectral interval combination optimization[J]. Analytica Chimica Acta, 2016, 948: 19−29.
[24] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[25] YUN Yonghuan, WANG Weiting, DENG Baichuan, et al. Using variable combination population analysis for variable selection in multivariate calibration[J]. Analytica Chimica Acta, 2015, 862: 14−23.
[26] YUN Yonghuan, BIN Jun, LIU Dongli, et al. A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration[J]. Analytica Chimica Acta, 2019, 1058: 58−69.
[27] 刘建学, 杨国迪, 韩四海, 等. 白酒基酒中典型醇的近红外预测模型构建[J]. 食品科学, 2018, 39(2): 281−286. [LIU J X, YANG G D, HAN S H, et al. Prediction model for typical alcohols in base liquor based on near infrared spectroscopy[J]. Food Science, 2018, 39(2): 281−286.]
[28] MARTELO-VIDAL M J, VÁZQUEZ M. Evaluation of ultraviolet, visible, and near infrared spectroscopy for the analysis of wine compounds[J]. Czech Journal of Food Sciences, 2014, 32(1): 37−47.
[29] SRIVASTAVA P K, ULLAS G, RAI S B. Overtone spectroscopy of benzaldehyde[J]. Pramana, 1994, 43(3): 231−236.
[30] TIPPARAT P. Determination of ethanol in liquor by near-infrared spectrophotometry with flow injection[J]. Talanta, 2001, 53(6): 1199−1204.
[31] 赵玉霞, 王茹, 张世芝, 等. 基于紫外-近红外融合光谱对“互助”青稞酒判别分析研究[J]. 食品与发酵工业, 2025, 51(4): 301−308. [ZHAO Y X, WANG R, ZHANG S Z, et al. Qualitative identification of Chinese Huzhu Qingke liquor based on ultraviolet spectroscopy combined with near-infrared spectroscopy[J]. Food and Fermentation Industries, 2025, 51(4): 301−308.]
[32] WORKMAN J J, WEYER L. 近红外光谱解析实用指南(褚小立, 许育鹏, 田高友, 译)[M]. 北京: 化学工业出版社, 2009: 21−23. [WORKMAN J J, WEYER L. Practical guide to near infrared spectroscopy analysis (Chu Xiaoli, Xu Yupeng, Tian Gaoyou, Trans.)[M]. Beijing: Chemical Industry Press, 2009: 21−23.]
[33] ENGLAND L, SCHIÖBERG D, LUCK W A P. The meaning of the very different O–H bandshapes in infrared fundamental and overtone regions of H-bonded methanol[J]. Journal of Molecular Structure, 1986, 143: 325−328.
[34] 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. [GAO C, ZHANG Y F, XIN Y, et al. Quantitative analysis of total esters in Baijiu base liquor by near-infrared spectroscopy combined with band selection[J]. China Brewing, 2021, 40(4): 155−158.]