

基于NIR和GC-MS融合技术的浓香型白酒原酒等级鉴别

张 维¹, 张贵宇^{1,2,*}, 庾先国^{1,2,*}, 付 妮¹, 李晓平¹, 庞婷婷¹, 刘科材³

(1.四川轻化工大学自动化与信息工程学院, 四川 宜宾 644000; 2.四川轻化工大学 人工智能四川省重点实验室, 四川 宜宾 644000; 3.四川轻化工大学工程实践中心, 四川 宜宾 644000)

摘 要: 以蒸馏过程中不同等级的浓香型白酒原酒为研究对象, 分别获取原酒的近红外光谱 (near infrared spectroscopy, NIR) 数据和气相色谱-质谱 (gas chromatography-mass spectrometry, GC-MS) 数据。采用5点2次卷积平滑对NIR数据进行预处理后, 利用竞争性自适应重加权算法 (competitive adaptive reweighted sampling, CARS) 筛选光谱特征波数; 结合Spearman等级相关系数、最大信息系数和随机森林变量重要性筛选GC-MS中影响原酒等级划分的关键风味成分 (key flavor components, KC)。然后利用极端梯度提升树分别建立基于NIR和GC-MS以及融合数据的原酒等级鉴别模型。结果表明, 基于CARS选择的光谱特征变量建立的模型预测准确率为89.66%, 基于特征选择后的KC建立的模型预测准确率为94.83%, 基于CARS+KC融合数据建立的模型分类准确率达到98.28%。研究表明, 将GC-MS数据和NIR数据的有效特征信息进行数据融合, 可以改善单一检测技术对不同等级原酒特征信息表征不全面的缺点, 在单一数据源的基础上提高原酒等级鉴别的准确率和稳定性, 实验结果为原酒的等级鉴别以及白酒其他的质量控制提供了新的思路和理论基础。

关键词: 浓香型白酒原酒; 近红外光谱; 气相色谱-质谱联用; 数据融合; 极端梯度提升树

Grade Identification of Raw Nongxiangxing Baijiu Based on Fused Data of Near Infrared Spectroscopy and Gas Chromatography-Mass Spectrometry

ZHANG Wei¹, ZHANG Guiyu^{1,2,*}, TUO Xianguo^{1,2,*}, FU Ni¹, LI Xiaoping¹, PANG Tingting¹, LIU Kecai³

(1. School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin 644000, China;
2. Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644000, China;
3. Engineering Practice Center, Sichuan University of Science & Engineering, Yibin 644000, China)

Abstract: Raw Nongxiangxin Baijiu of different grades were collected during the distillation process, and their near infrared spectroscopy (NIR) data and gas chromatography-mass spectrometry (GC-MS) data were acquired. After preprocessing the NIR data through 5-point 2-fold convolutional smoothing, spectral feature wavelengths were selected using the competitive adaptive reweighted sampling (CARS) algorithm; combining Spearman's rank correlation coefficient, maximum information coefficient (MIC) and random forest (RF) variable importance, the key flavor components (KC) identified by GC-MS affecting the grading of raw Baijiu were determined. Then, extreme gradient boosting tree (XGBoost) was applied to establish three grade identification models for raw Baijiu based on NIR, GC-MS and their fused data. The results showed that the prediction accuracy of the model based on the spectral feature variables selected by CARS was 89.66%, the prediction

收稿日期: 2024-04-15

基金项目: 泸州老窖研究生创新基金项目 (LJCX-2022-8); 酿酒生物技术及应用四川省重点实验室开放课题 (NJ2022-06); 五粮液产学研合作项目 (CXY2022ZR007); 中国轻工业酿酒生物技术及智能制造重点实验室项目 (2023-01); 四川轻化工大学《横向科研项目结余经费出资科技成果转化专项》项目 (HXJY01); 四川轻化工大学2023年度“652”科研创新团队资助项目 (SUSE652B005)

第一作者简介: 张维 (1996—) (ORCID: 0009-0009-3578-6772), 女, 硕士研究生, 研究方向为智能酿造。

E-mail: 1517395358@qq.com

*通信作者简介: 张贵宇 (1987—) (ORCID: 0000-0003-3890-6839), 男, 副教授, 博士, 研究方向为白酒自动化、人工智能。

E-mail: gyz_118@163.com

庾先国 (1965—) (ORCID: 0000-0003-1381-8023), 男, 教授, 博士, 研究方向为核技术。

E-mail: tuoxianguo@suse.edu.cn

accuracy of the model based on KC after feature selection was 94.83%, and the classification accuracy of the model based on the fused data of CARS + KC reached as high as 98.28%. This study shows that the fusion of effective feature information from GC-MS and NIR data can enable more accurate and stable grade identification of raw Nongxiangxin Baijiu than either analytical technique alone, which provides a new idea and theoretical basis for the grade identification and quality control of raw Baijiu.

Keywords: raw Nongxiangxin Baijiu; near infrared spectroscopy; gas chromatography-mass spectrometry; data fusion; extreme gradient boosting tree

DOI:10.7506/spkx1002-6630-20240415-119

中图分类号: TS262.3

文献标志码: A

文章编号: 1002-6630 (2024) 21-0288-09

引文格式:

张维, 张贵宇, 庾先国, 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. DOI:10.7506/spkx1002-6630-20240415-119. <http://www.spkx.net.cn>

ZHANG Wei, ZHANG Guiyu, TUO Xianguo, 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. (in Chinese with English abstract) DOI:10.7506/spkx1002-6630-20240415-119. <http://www.spkx.net.cn>

原酒又称为原浆酒, 是发酵好的酒醅经过蒸馏后得到的未经贮存和勾兑处理的半成品酒, 原酒的准确分类对分级入库贮存以及最终成品白酒的质量有重大影响^[1]。目前对于原酒的质量评价主要依靠品酒师的感官评定, 这种方式对品酒师的经验以及状态要求极高, 在评定的过程中容易受到人为主观因素的影响, 同时耗时耗力, 无法批量完成, 难以满足市场的需求。因此制定一套标准的、科学的原酒评价方法用于实现原酒智能分级是目前白酒企业研究的一个重点问题^[2]。

近年来, 随着气相色谱仪^[3]、液相色谱仪^[4]、气相色谱-质谱 (gas chromatography-mass spectrometry, GC-MS) 联用仪、电子鼻^[5]、电子舌^[6]、红外光谱仪^[7]、核磁共振仪^[8]等分析检测仪器被广泛应用到白酒酿造领域, 白酒酿造过程开始不断向机械化、智能化转移, 使得白酒酿造智能化成为当下酒企的研究热点。利用现代化检测仪器和大数据分析建立原酒等级评价模型, 降低原酒分级过程中人为主观因素的影响, 是提高原酒等级评价客观性的一种科学手段。在众多的检测仪器中, 近红外光谱 (near infrared spectroscopy, NIR) 仪因具有快速检测和无损分析的能力在食品领域^[9-10]、农业领域^[11-12]等得到了广泛的应用, 其在白酒的检测分析中也取得了良好的效果, 广泛应用于白酒的等级划分^[13-14]、酒龄检测^[15]、掺假研究^[16]以及成分快速检测^[17-18]。GC-MS是将具有分离能力的GC仪和具有定性能力的MS仪串联起来的一种在线联用技术, 具有检出限低、灵敏度高、稳定性强、分离度好等优点, 可以在较短时间内获得待测样品的色谱和质谱数据, 进而对多组分混合物进行定性和定量分析^[19]。胡雪^[20]采用GC-MS技术对不同产地、不同等级、不同香型的白酒中的风味物质进行综合分析, 并结合主成分分析和聚类分析建立了成品白酒的评价方法, 其对

等级、产地、香型的判别准确率均达到了100%。张健等^[21]基于GC-MS技术对茅台酒特征组分进行了定性和定量, 结合主成分分析对不同来源的酱香型白酒鉴别, 并通过偏最小二乘判别分析和聚类分析筛选出了区分真假茅台酒的差异化合物。宋丹丹等^[22]采用GC-MS比较分析了六大蒸馏酒的挥发性成分, 通过主成分分析在检测得到的79种挥发性物质的基础上揭示了六大蒸馏酒的成分差异。

基于蒸馏是一个连续过程, 其原酒分段点处相邻样本的内部变化非常小, 单一的检测技术很难对质量呈连续性变化的原酒样本信息进行准确全面的表述, 而数据融合技术可以将不同检测器从不同角度获取的样本内部的特征信息结合起来, 弥补了单一检测技术对原酒质量连续变化带来的数据特征信息表征不全面的缺点, 通过增强原酒质量变化的特征信息进而在单一检测技术的基础上提高原酒等级分类的效果, 故结合GC-MS和NIR实现原酒智能分级具有较高的可行性。

鉴于此, 本研究通过GC-MS检测技术得到原酒中的挥发性风味成分含量数据, 通过NIR仪获取反映原酒内部结构的NIR数据, 基于GC-MS和NIR单独评价原酒等级分类效果, 然后比较基于单一数据源和数据融合策略建立的模型的优劣, 为数据融合可提高原酒等级分类的准确率和稳定性提供理论依据。

1 材料与方法

1.1 材料与试剂

本实验样品选自遂宁舍得酒业的浓香型白酒系列, 原酒样品采集过程由具有10 a以上摘酒经验的摘酒师傅完成, 样品采集完成后由酒企品评小组 (5名专业品酒师) 进行原酒等级评定, 具体样本分布情况见表1。

表1 原酒样本分布信息

Table 1 Distribution information of raw Baijiu samples

| 标签 | 原酒等级 | 样本数量 | 特点 |
|----|------|------|--------------------------|
| 1 | 头酒 | 92 | 乙醇含量高, 醛类物质多, 质量不佳 |
| 2 | 中段酒 | 164 | 香气浓郁, 风格突出, 酒质清澈透明 |
| 3 | 尾酒 | 130 | 乙醇含量偏低, 口感不醇正, 酒质浑浊, 有杂味 |

叔戊醇、己酸正戊酯、2-乙基丁酸(均为色谱纯)、无水乙醇(纯度99.5%) 上海麦克林生化科技有限公司; 甲醇(纯度99.9%) 上海阿达玛斯试剂有限公司; C₇~C₄₀正构烷烃(色谱纯) 北京曼哈格生物科技有限公司。

1.2 仪器与设备

Matrix-F傅里叶变换NIR仪、近红外光纤探头德国Bruker公司; 7890B GC仪、G7000D MS仪 美国Agilent公司。

1.3 方法

1.3.1 原酒NIR数据获取

原酒NIR数据的采集无需对样品进行预处理, 将傅里变换NIR仪置于温度(20±2)℃、空气相对湿度<80%的环境下预热50 min左右, 检测样品前先检查仪器信号, 使干涉图能量达到最大, 获取背景光谱用于消除水蒸气和二氧化碳等环境因素对样品光谱检测结果的影响。在室温下通过近红外光纤探头扫描样品, 采用OPUS7.8控制光谱仪并进行光谱记录, 光谱扫描范围为4 000~12 500 cm⁻¹, 相位分辨率为32 cm⁻¹, 频率为10 kHz, 分辨率为8 cm⁻¹, 累计扫描32次后取每个光谱点上的平均值为最终光谱。

1.3.2 原酒GC-MS数据获取

1.3.2.1 样品前处理

内标配制: 用超纯水配制60%的乙醇溶液作为内标化合物的溶剂, 分别准确称取2 g叔戊醇、己酸正戊酯、2-乙酸丁酯于100 mL容量瓶中, 加入配制好的乙醇溶液定容至100 mL, 放置0~4℃低温冰箱中保存备用。

样品准备: 使用微量移液枪量取1 mL原酒样品溶液于测样瓶中, 加入10 μL配制好的内标溶液, 做好标签记录, 混合均匀后送至GC-MS仪中检测分析。

1.3.2.2 GC-MS参数设置

GC采用自动进样, 色谱柱为Agilent DB-WAX(60 m×0.25 mm, 0.25 μm)。进样量为1 μL, 分流比为40:1, 进样口温度为250℃; 载气为高纯氦气(He), 流速为1 mL/min; 初始柱温为60℃, 保持5 min, 以10℃/min升温至250℃, 并保持2 min; 电子电离源, 电离能70 eV; 离子源温度230℃, 四极杆温度150℃; 全扫描方式; 扫描范围m/z 30~540。

1.3.2.3 定性定量分析

确定待测样品中复杂成分种类和具体含量即为定性定量分析^[23-24]。本实验采用内标法, 利用自动识别的色谱峰与美国国家标准技术研究所(National Institute of

Standards and Technology, NIST) 12质谱库检索对比结合保留指数(retention index, RI)对原酒中的风味成分进行定性分析; 参考GB/T 10345—2022《白酒分析方法》中峰面积法对各风味成分进行定量分析。

1.3.3 数据融合

数据融合是将通过不同分析技术获取的特征信息进行融合, 利用不同技术之间的协同作用获得质量更高、更加全面的数据信息, 使其能够更加充分地解析检测目标^[25]。根据不同层次的数据源可将数据融合分为数据级融合、特征级融合和决策级融合3种融合方式^[26]。其中, 数据级融合是将通过不同检测技术获取的全部数据信息直接串连起来进行变量分析及建模; 特征级融合是分别对不同检测器获取的数据进行特征提取后再将这些数据组合到一起进行建模分析; 决策级融合是利用每种数据源独立地构建模型, 然后将模型决策边界进行融合, 进行融合时需要对其分别赋予不同的权重, 其计算量大, 需要很高的计算资源和处理能力, 故本研究选择数据级融合和特征级融合两种方法进行建模分析。

1.3.4 极端梯度提升树(extreme gradient boosting tree, XGBoost)及超参数优化

1.3.4.1 XGBoost

XGBoost是基于梯度提升树(gradient boosting decision trees, GBDT)的思想进行改进的算法, 通过累加多棵分类与回归树(classification and regression tree, CART)结果得到最终的预测结果^[27]。和GBDT相似, XGBoost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^[28]。不同的是, XGBoost考虑了树的复杂度, 在损失函数中加入了正则项用于控制模型的复杂程度, 防止过拟合现象; 另外, GBDT使用的损失函数负梯度作为标签, 而XGBoost通过求目标函数极值点和二阶泰勒展开逼近的方式得到树的结构, 进一步考虑了梯度变化的趋势, 提高了拟合速度和精确度。

假设有n个样本数据, 已经训练了m次, 则第m棵树上第i个样本的模型预测值 \hat{y}_i 为:

$$\hat{y}_i = \sum_{j=1}^m f_j(x_i), f_j \in F \quad (1)$$

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; F为CART构成的集合。

则目标函数(obj)为:

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

式中: y_i 表示第i个样本的真实类别; $l(y_i, \hat{y}_i)$ 为损失函数, 表示预测值 \hat{y}_i 和真实值 y_i 之间的差异; $\Omega(f_m)$ 为正则项, 用于控制模型的复杂度, 防止模型过拟合; M表示树的数量。

1.3.4.2 网格搜索超参数优化

超参数的设置对XGBoost模型的预测性能有重要影响,本研究采用网格搜索确定通过不同数据源建立模型时的最优参数。网格搜索是目前应用最为广泛的超参数搜索算法,其通过查找搜索范围内的所有点确定最优参数,超参数含义及其搜索范围见表2。

表2 XGBoost超参数含义及其搜索范围

Table 2 Meaning of hyperparameters in XGBoost and their search ranges

| 超参数名称 | 含义 | 搜索范围 |
|------------------|----------|----------|
| n_estimators | 决策树数量 | [50,200] |
| learning_rate | 学习率 | [0,1] |
| max_depth | 最大树深度 | [2,10] |
| colsample_bytree | 特征随机采样比例 | [0,1] |
| min_child_weight | 最小叶子节点权重 | [0,10] |

2 结果与分析

2.1 NIR数据处理

2.1.1 光谱数据预处理

傅里叶变换NIR仪扫描得到的原始NIR范围为12 500~4 000 cm⁻¹,其中波数范围在12 500~9 025 cm⁻¹内的光谱几乎无化学键的吸收,波数范围在4 300~4 000 cm⁻¹内的光谱受仪器和环境的影响杂乱无规律,因此这两段内的光谱对于后续的数据处理没有价值,故对这两部分波段进行删除处理,截取9 025~4 300 cm⁻¹内的1 226个光谱数据进行后续分析处理,截取后的NIR如图1所示。

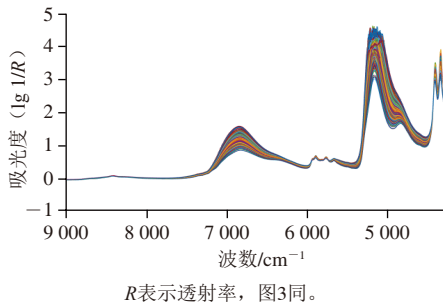


图1 原酒9 025~4 300 cm⁻¹范围内的NIR图

Fig. 1 NIR spectra of raw Baijiu in the wavenumber range of 9 025~4 300 cm⁻¹

因光谱数据采集过程中受环境、仪器和操作的影响容易造成光谱存在基线漂移、散射效应等降低变量解释能力的干扰信息,相关研究证明适当的预处理方式可以削弱或者消除光谱采集过程中因环境或仪器带来的误差影响,提高数据信噪比,增强光谱数据的表征能力^[29-30]。本研究通过5折交叉验证评估NIR原始数据和经过不同预处理后建立的模型预测效果,经过模型验证后,确定5点2次卷积平滑处理原酒NIR光谱建立的模型预测效果较好。其中,原始数据建立的模型5折交叉验证准确率为

84.21%,经过5点2次卷积平滑预处理后的光谱数据建立的模型5折交叉验证准确率达到87.30%,在原始光谱的基础上提高了3.09%,故后续研究在5点2次卷积平滑预处理的基础上进行。

2.1.2 特征选择

由于删减后的光谱变量维数仍然较高,且还含有大量与原酒等级变化无关的冗余信息,需要对其进行降维处理。本研究采用竞争性自适应重加权算法(competitive adaptive reweighted sampling, CARS)对光谱数据进行特征选择,CARS是一种结合蒙特卡洛采样和偏最小二乘回归(partial least squares regression, PLSR)系数进行特征选择的方法^[31]。CARS作特征选择的具体步骤如下:

设原酒样本构成的光谱矩阵为 $X_{n \times m}$,样本类别矩阵为 $Y_{n \times 1}$,其中 n 为样本个数, m 为光谱波长个数。

1) 蒙特卡洛模型采样

利用蒙特卡洛采样法选取校正集,用选取的校正集建立偏PLSR模型。模型表示为:

$$Y = Xb + e \quad (3)$$

式中: b 为 m 维的模型回归系数; e 为误差向量。

2) 指数衰减波长筛选

利用衰减指数法确定去除波长的个数,第一次采样时所有光谱变量均参与建模, N 次迭代中确定的变量逐次递减,其中第 i 次采样确定的变量个数比例 r_i 计算如下:

$$r_i = ae^{-ki} \quad (4)$$

式(4)的约束条件为: $r_1 = n$, $r_N = 2/n$ 。则指数递减参数 a 和 k 的计算公式如下:

$$a = \left(\frac{n}{2}\right)^{1/(N-1)} \quad (5)$$

$$k = \frac{\ln \frac{n}{2}}{N-1} \quad (6)$$

3) 自适应重加权采样

采用自适应重加权采样技术基于步骤2)保留的特征波数子集进行特征筛选,通过 w_i 表示每个光谱波数变量的权重,其计算公式如下:

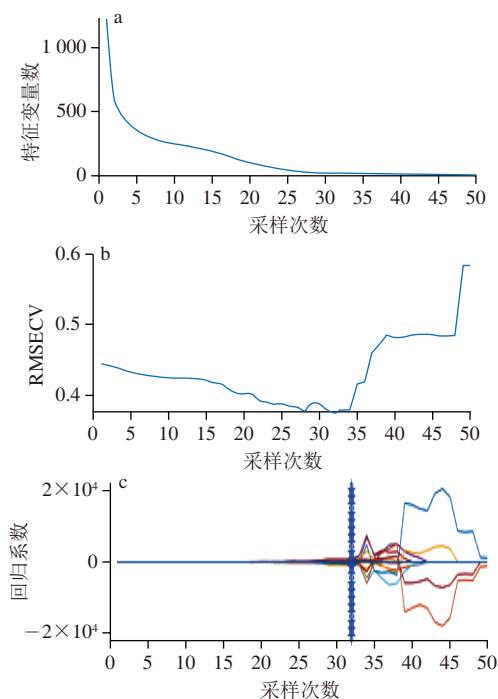
$$w_i = \frac{|b_i|}{\sum_{i=1}^n |b_i|} \quad (7)$$

式中: $|b_i|$ 表示第 i 个波数变量的回归系数绝对值,其值越大,第 i 个波数变量对模型的贡献程度越大。

4) 循环迭代

按照设置的循环迭代次数 N 进行采样建立PLSR模型,取模型交叉验证均方根误差(root mean square error of cross validation, RMSECV)最小时PLSR模型采用的光谱波数集合作为CARS选择的特征子集。

CARS选择过程如图2所示。



a.选择的波数数量; b. RMSECV图; c. 变量回归系数。

图2 CARS特征波长选择过程

Fig. 2 Selection process of characteristic wavelengths by CARS

从图2a可以看出,随着采样次数的增加,选择的变量数呈现出先快速减少后缓慢减少的趋势,表明这是一个粗选到精选的过程;图2b是经过十折交叉验证所得PLSR模型的RMSECV随自适应重加权采样运行次数的变化趋势,可以看出在1~32次采样过程中,RMSECV的整体变化趋势减小,表明剔除的特征波数与样本属性无关,从33次采样开始RMSECV开始随着采样次数逐渐递增,表明可能剔除了与样本类别变化相关的关键变量;图2c是变量回归系数随采样次数的变化趋势图,其中的蓝色“*”组成的竖线对应图2b中的RMSECV最小值点,此处剩余的特征波数即为CARS选定的特征波数。其对应的特征选择结果如图3所示,共筛选出46个特征波数。将筛选得到的46个特征波数作为输入建立原酒等级鉴别模型,采用5折交叉验证评估其所建模型的预测精度,得到其5折交叉验证准确率为88.59%,在预处理的基础上提高了1.29%,故保留的46个特征波数有效。

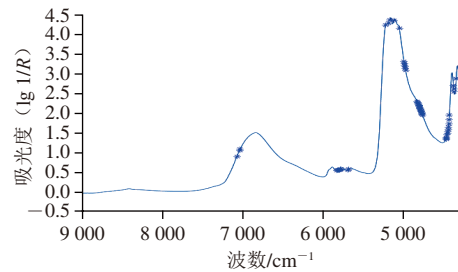


图3 CARS特征选择结果

Fig. 3 Results of feature selection by CARS

2.2 GC-MS数据处理

2.2.1 挥发性风味成分GC-MS分析

通过GC-MS结合内标法对原酒样品进行定性定量分析可知,386个浓香型白酒原酒样本中一共检出86种物质。剔除部分在样本中检出率较低且无规律的物质,保留58种物质进行分析,具体物质种类和含量分布信息见表3。

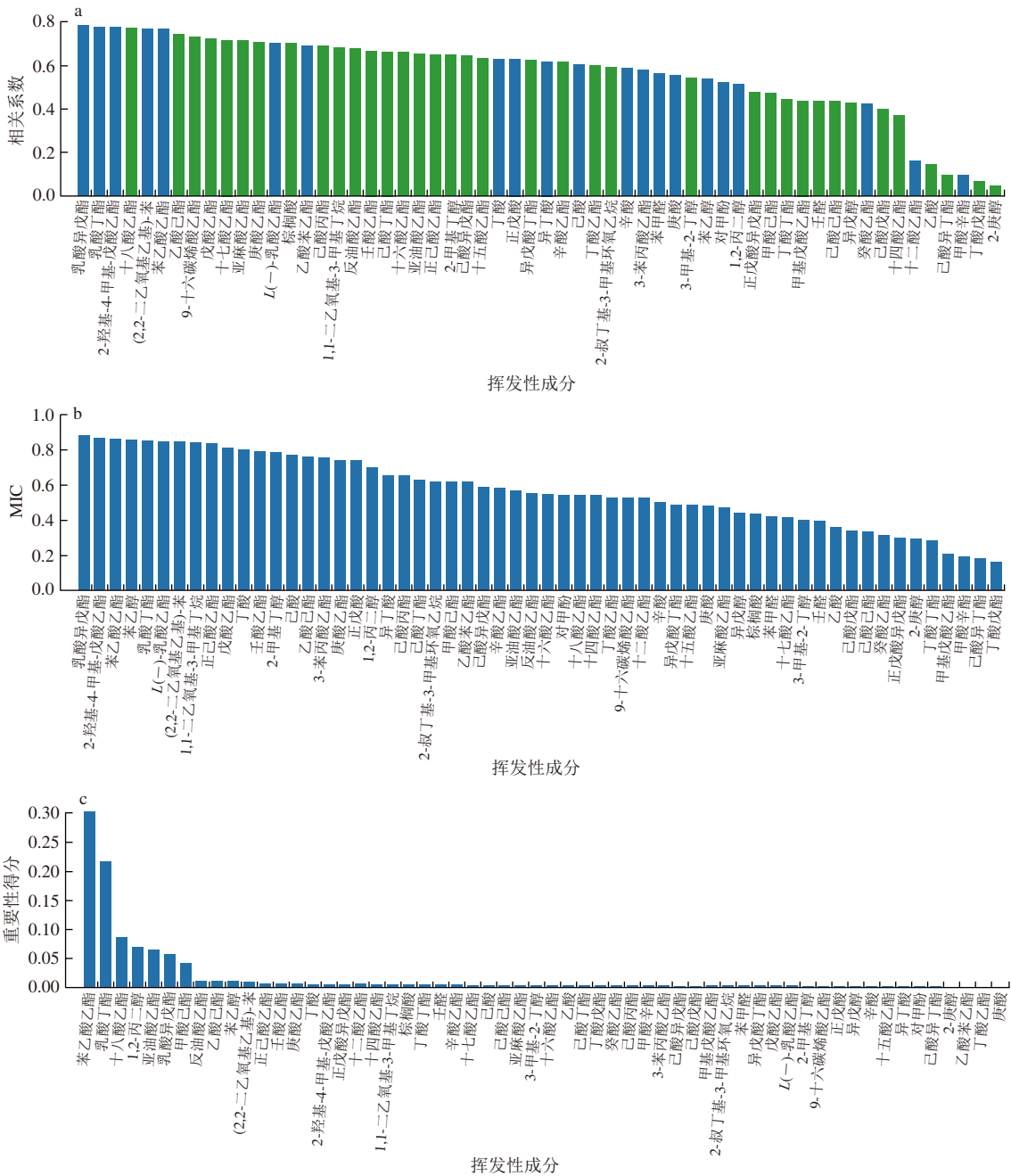
表3 原酒中挥发性风味成分名称与含量分布信息
Table 3 Concentration ranges of volatile flavor components in raw Baijiu

| 序号 | 挥发性风味成分 | 质量浓度范围/(mg/L) |
|----|-----------------|----------------|
| 1 | 十二酸乙酯 | 0~19.38 |
| 2 | 壬酸乙酯 | 0~17.92 |
| 3 | 9-十六碳烯酸乙酯 | 0~111.94 |
| 4 | (2,2-二乙氧基乙基)-苯 | 0~14.17 |
| 5 | 癸酸乙酯 | 2.31~27.79 |
| 6 | 乳酸丁酯 | 0~14.7 |
| 7 | 乙酸己酯 | 0~54.41 |
| 8 | 正戊酸 | 0~65.79 |
| 9 | 乳酸异戊酯 | 0~21.92 |
| 10 | 己酸丙酯 | 0~95.77 |
| 11 | 十四酸乙酯 | 0~204.84 |
| 12 | 己酸异戊酯 | 4.47~74.04 |
| 13 | 戊酸乙酯 | 0~246.06 |
| 14 | 2-甲基丁醇 | 0~65.73 |
| 15 | 1,1-二乙氧基-3-甲基丁烷 | 0~165.3 |
| 16 | 反油酸乙酯 | 7.58~855.02 |
| 17 | 庚酸乙酯 | 17~274.63 |
| 18 | 己酸己酯 | 20.92~284.69 |
| 19 | 2-羟基-4-甲基-戊酸乙酯 | 0~92.01 |
| 20 | 丁酸 | 21.5~215.24 |
| 21 | 十六酸乙酯 | 2.53~1 454.88 |
| 22 | 辛酸乙酯 | 0~404.03 |
| 23 | 己酸 | 57.26~466.78 |
| 24 | 异戊醇 | 0~186.55 |
| 25 | 乙酸 | 63.04~333.04 |
| 26 | 正己酸乙酯 | 33.98~3 684.09 |
| 27 | L(-)-乳酸乙酯 | 156.09~720.82 |
| 28 | 十七酸乙酯 | 0~16.43 |
| 29 | 2-叔丁基-3-甲基环氧乙烷 | 0~204.9 |
| 30 | 甲酸己酯 | 0~171.12 |
| 31 | 十八酸乙酯 | 0~99.91 |
| 32 | 亚麻酸乙酯 | 0~199.55 |
| 33 | 己酸丁酯 | 0~189.14 |
| 34 | 亚油酸乙酯 | 0~1 298.21 |
| 35 | 棕榈酸 | 0~179.98 |
| 36 | 2-庚醇 | 0~5.68 |
| 37 | 己酸异丁酯 | 0~11.23 |
| 38 | 己酸戊酯 | 0~42.87 |
| 39 | 3-苯丙酸乙酯 | 0~73.5 |
| 40 | 苯乙酸乙酯 | 0~36.4 |
| 41 | 3-甲基-2-丁醇 | 0~4.69 |
| 42 | 丁酸丁酯 | 0~33.51 |
| 43 | 异戊酸丁酯 | 0~8.06 |
| 44 | 丁酸戊酯 | 0~11.37 |
| 45 | 对甲酚 | 0~20.67 |
| 46 | 异丁酸 | 0~21.42 |
| 47 | 苯乙醇 | 0~17.73 |
| 48 | 十五酸乙酯 | 0~37.96 |
| 49 | 正戊酸异戊酯 | 0~2.92 |
| 50 | 壬醛 | 0~3.29 |
| 51 | 辛酸 | 0~57.66 |
| 52 | 1,2-丙二醇 | 0~41.66 |
| 53 | 庚酸 | 0~49.42 |
| 54 | 甲基戊酸乙酯 | 0~8.03 |
| 55 | 苯甲醛 | 0~4.74 |
| 56 | 甲酸辛酯 | 0~9.51 |
| 57 | 丁酸乙酯 | 0~24.71 |
| 58 | 乙酸苯乙酯 | 0~4.61 |

由表3可知, 58 种物质中酯类物质数量最多, 有 38 种。另外, 酸类物质有 8 种、醛类物质有 2 种、醇类物质有 6 种、其他物质有 4 种。蒸馏过程中物质的含量变化并不是非增即减, 部分噪音物质的含量变化并不明显或者没有规律, 故在建模之前有必要进行特征选择。

2.2.2 关键风味成分 (key flavor components, KC) 选择
采用 Spearman 等级相关系数、最大信息系数 (maximum information coefficient, MIC) 和随机森林 (random forest, RF) 变量重要性排序对影响原酒等级

变化的 KC 进行选择。其中, Spearman 等级相关系数是一种非参数统计方法, 适用于任何形态的数据分布, 它基于两列成对等级数之间的等级差计算相关性, 只要变量之间的观测值可以转换为成对等级资料, 不论数据分布形态或样本容量大小如何, 都可以进行相关性分析^[32-33]; MIC 是由 Reshef 等^[34]在 2011 年提出的一种用于衡量两个变量关系的非参数统计方法, MIC 不仅可以衡量变量之间的线性关系, 还可以衡量两个变量之间的非线性关系, 它通过两个变量之间的联合概率密度衡量两个变量之间



a. Spearman 等级相关系数; b. MIC 值; c. RF 变量重要性。

图 4 3 种方法得到的挥发性风味成分的重要性排序

Fig. 4 Ranking of importance of volatile flavor components obtained by three methods

的相关程度。RF是以决策树为基学习器的集成学习算法。RF主要用于分类和回归任务，也适用于数据降维问题；RF用于数据降维主要是通过评估每个特征在RF的每棵树上面的贡献率，然后取平均值得到^[35]。因为RF具有双重随机性，仅根据特征变量在决策树中出现的频率衡量特征重要性不够可靠，为了更准确地反映特征的重要性，本研究选择基于袋外数据（out-of-bag，OOB）的平均精度下降来计算均方根误差的平均值作为特征的重要性。3种方法得到的风味成分的重要性排序如图4所示。

分析上述3种方法所得风味成分的相关系数以及重要性得分可知，它们在判断与原酒等级划分相关性较大的风味成分时，得出的结果有所差异，若直接设定阈值从3种方法中筛选KC，其结果代表性不高，为更好地筛选具有代表性的KC。本研究取3种方法所得前 n 个特征的交集建立等级鉴别模型，采用5折交叉验证评估模型的结果，其交叉验证结果如图5所示。

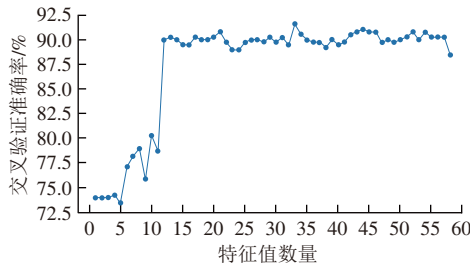


图5 5折交叉验证评估有效特征化合物数量

Fig. 5 Evaluation of the number of effective characteristic compounds by 5-fold cross-validation

由图5可知，当模型输入通过交集获取的前33个风味成分时，其5折交叉验证准确率最高，故KC为3种方法的前33个挥发性成分的共有风味成分。对3种方法的前33个风味成分进行Venn分析，如图6所示。

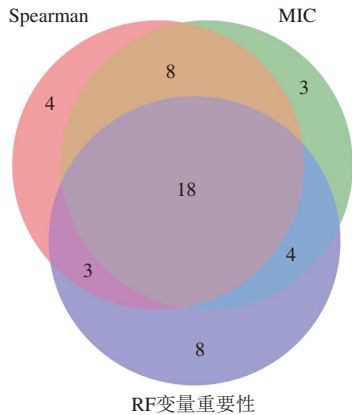


图6 前33个特征化合物的Venn图

Fig. 6 Venn diagram of the first 33 characteristic compounds

由图6可知，3种方法在前33个风味成分中的共有风味成分有18个，则这18个风味成分即为筛选的KC，筛选的18个KC的名称如表4所示。

表4 KC名称

Table 4 Key flavor components

| 序号 | 风味成分 | 序号 | 风味成分 |
|----|-----------------|----|-------|
| 1 | 乳酸丁酯 | 10 | 正己酸乙酯 |
| 2 | 乳酸异戊酯 | 11 | 丁酸 |
| 3 | 苯乙酸乙酯 | 12 | 亚油酸乙酯 |
| 4 | (2,2-二乙氧基乙基)-苯 | 13 | 反油酸乙酯 |
| 5 | 乙酸己酯 | 14 | 十六酸乙酯 |
| 6 | 2-羟基-4-甲基-戊酸乙酯 | 15 | 辛酸乙酯 |
| 7 | 庚酸乙酯 | 16 | 己酸 |
| 8 | 壬酸乙酯 | 17 | 十八酸乙酯 |
| 9 | 1,1-二乙氧基-3-甲基丁烷 | 18 | 己酸丁酯 |

2.3 浓香型原酒等级分类模型建立

2.3.1 基于单一数据源建模

本研究采用随机分层抽样的方式按照7:3的比例将数据集划分为训练集和测试集进行模型的训练与验证。在机器学习领域，模型建立好后需要通过一系列模型评价指标从多个维度量化模型的效果，为了对模型有一个更加全面的评估，本研究采用每段酒的具体分类准确率对建立的原酒等级分类模型进行评价。分别采用处理前后的NIR数据和GC-MS数据建立XGBoost分类模型，其模型结果见表5。

表5 基于单一数据源建立的XGBoost分类模型预测结果

Table 5 Prediction results of XGBoost classification models based on single data sources

| 数据源 | 特征变量数 | 头酒分类准确率/% | 中段酒分类准确率/% | 尾酒分类准确率/% | 准确率/% | F1分数/% |
|-------|-------|-----------|------------|-----------|-------|--------|
| NIR | 处理前 | 1 226 | 71.43 | 83.67 | 100 | 86.21 |
| | 处理后 | 46 | 75.00 | 89.80 | 100 | 89.66 |
| GC-MS | 处理前 | 58 | 85.71 | 95.92 | 97.44 | 93.97 |
| | 处理后 | 18 | 92.86 | 95.92 | 94.87 | 94.83 |

由表5可知，经过CARS选择的特征波数建立的模型预测准确率和F1分数分别为89.66%、89.57%，其值在原始数据集的基础上分别提高了3.45%、3.48%，通过KC建立的模型其预测准确率和F1分数为94.83%、94.89%，在原始数据集的基础上分别提高了0.86%和0.96%，说明经筛选后的特征变量能有效地剔除冗余特征，保留对原酒等级划分有效的信息；另外，可以看出基于NIR数据建立的模型对尾酒的预测效果更好，基于GC-MS数据建立的模型对头酒和中段酒的预测效果更好，这一结果表明，原酒的质量变化是一个非常复杂的过程，单一检测技术对不同等级原酒的特征信息表征并不全面。

2.3.2 基于融合数据建模

白酒是一个多组分混合物，蒸馏过程中原酒的品质变化具有连续缓慢变化的特点，因此，仅通过单一检测器获取的数据实现原酒等级鉴别是不够的，故本研究采用数据融合策略结合GC-MS和NIR两种检测数据进行联合建模。首先，将检测得到的GC-MS数据和NIR数据直接拼接起来组成数据级融合数据，然后将经CARS提取的NIR特征变量和经Spearman等级相关系数、MIC和RF变量重要性筛选的GC-MS关键成分拼接起来组成特征级融

合数据,最后通过两种融合数据分别结合XGBoost建立原酒等级鉴别模型,其模型预测结果如表6所示。

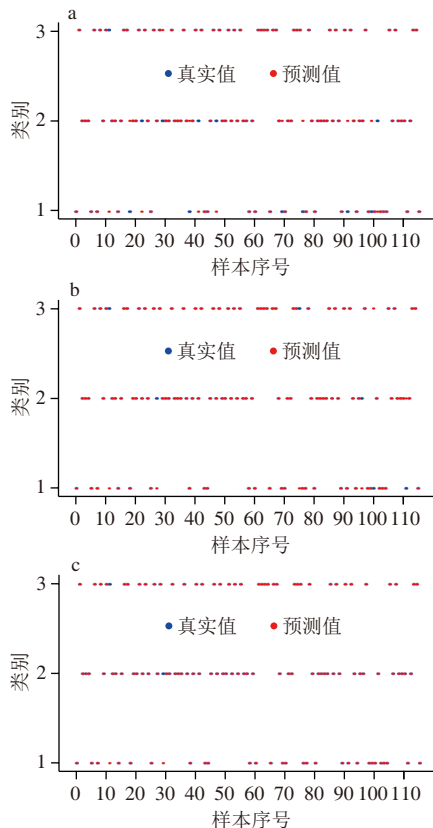
表6 基于数据融合的原酒等级分类模型预测结果
Table 6 Prediction results of raw Baijiu grade classification model based on fused data

| 数据 | 特征变量数 | 头酒分类准确率/% | 中段酒分类准确率/% | 尾酒分类准确率/% | 准确率/% | F1分数/% |
|---------|-------|-----------|------------|-----------|-------|--------|
| 数据集融合数据 | 1 284 | 89.29 | 95.92 | 97.44 | 94.83 | 94.84 |
| CARS+KC | 64 | 100 | 97.96 | 97.44 | 98.28 | 98.30 |

注: CARS+KC.特征级融合数据。

由表6可知,通过特征级融合数据建立的模型预测效果优于数据级融合数据,其原因可能是数据级融合直接融合了全部NIR数据和GC-MS数据,这些数据中包含了大量与原酒等级变化无关的噪声信息,导致了建模效果较差,而特征级融合滤除了原始数据集中的噪声信息,融合后的数据集特征信息得到了增强,因此预测效果较好。与表5对比分析可知,基于数据级融合数据建立的模型预测效果在单一数据源的基础上没有明显提升,而基于特征级融合数据建立的分模型准确率达到98.28%,对比基于特征选择后的NIR数据和GC-MS数据建立的模型,其预测准确率分别提高了8.62%和3.45%。

结合表5、6可得基于3种数据源建立的最佳鉴别模型的具体判别结果,具体如图7所示。



a. CARS-XGBoost; b. KC-XGBoost; c. CARS+KC-XGBoost。

图7 基于3种数据源建立的最佳分类模型具体分类结果

Fig. 7 Classification results of the best classification models based on three data sources

由图7可知,基于CARS+KC-XGBoost建立的模型仅中段酒和尾酒各有1个分类错误,其3个类别的原酒分类准确率分别为100%、97.96%和97.44%,分类效果较好;另外,可以看出准确率的提高主要体现在头酒的分类结果上,弥补了基于单一数据源在个别等级上分类准确率低的情况,并提高了原酒等级分类的整体分类准确率。说明来自GC-MS和NIR两种数据的互补可很好地表征不同等级原酒的特征信息,提高原酒等级分类的准确率和稳定性。

3 结论

本研究以蒸馏过程中不同等级的浓香型白酒原酒为研究对象,分别获取了原酒的NIR数据和GC-MS数据,基于两种数据研究了数据融合在单一检测技术基础上对原酒等级鉴别的影响,实验主要结论如下:1)基于特征选择后的NIR数据和GC-MS数据建立的模型分类准确率分别为89.66%和94.83%,其中基于NIR数据建立的模型对尾酒的分类效果较好,基于GC-MS数据建立的模型对头酒和中段酒的分类效果较好;2)采用数据融合策略将两种数据进行特征级融合建立的模型分类准确率达到98.28%,相较于单一数据源建立的模型,其分类效果有明显提升。本研究结果为原酒的等级分类以及白酒其他的质量控制提供了新的思路 and 理论依据。

参考文献:

- [1] 周轩. 浓香型白酒基酒挥发性成分分析及等级识别研究[D]. 镇江: 江苏大学, 2019.
- [2] LI H H, ZHANG X, GAO X J, et al. Comparison of the aroma-active compounds and sensory characteristics of different grades of light-flavor Baijiu[J]. Foods, 2023, 12(6): 1238. DOI:10.3390/foods12061238.
- [3] 杨博, 郭启鹏, 苏正, 等. 现代检测技术在白酒真实性鉴别中的应用研究进展[J]. 食品工业, 2024, 45(2): 192-196.
- [4] 骆茂香, 邱树毅, 徐兴江, 等. 白酒中非挥发性风味成分检测分析研究进展[J]. 中国酿造, 2023, 42(9): 19-25. DOI:10.11882/j.issn.0254-5071.2023.09.004.
- [5] MOHD ALI M, HASHIM N, ABD AZIZ S, et al. Principles and recent advances in electronic nose for quality inspection of agricultural and food products[J]. Trends in Food Science & Technology, 2020, 99: 1-10. DOI:10.1016/j.tifs.2020.02.028.
- [6] KIRSANOV D, CORREA D S, GAAL G, et al. Electronic tongues for inedible media[J]. Sensors, 2019, 19(23): 5113. DOI:10.3390/s19235113.
- [7] ZHANG W W, KASUN L C, WANG Q J, et al. A review of machine learning for near-infrared spectroscopy[J]. Sensors, 2022, 22(24): 9764. DOI:10.3390/s22249764.
- [8] QU Q, JIN L. Application of nuclear magnetic resonance in food analysis[J]. Food Science and Technology, 2022, 42: e43622. DOI:10.1590/fst.43622.
- [9] 李娟, 梁漱玉. 近红外快速无损检测食用油品质的研究进展[J]. 食品与机械, 2016, 32(11): 225-228. DOI:10.13652/j.issn.1003-5788.2016.11.051.

- [10] 王明, 刘新. 近红外技术在液态食品成分检测中的应用研究进展[J]. 激光杂志, 2018, 39(10): 9-13. DOI:10.14016/j.cnki.jgzz.2018.10.009.
- [11] 郭东升, 张志勇, 武志明, 等. 基于近红外光谱的大豆水分和粗脂肪含量的快速检测[J]. 食品安全质量检测学报, 2020, 11(20): 7378-7384. DOI:10.19812/j.cnki.jfsq.11-5956/ts.2020.20.036.
- [12] 王永, 杨国耀, 乔俊峰, 等. 便携式近红外光谱仪及其在农业中的应用现状[J]. 江苏农业科学, 2022, 50(7): 10-17. DOI:10.15889/j.issn.1002-1302.2022.07.002.
- [13] 翟双, 张贵宇, 庾先国, 等. 近红外光谱结合二维卷积在白葡萄酒判别中的应用[J]. 食品科技, 2022, 47(9): 250-256. DOI:10.13684/j.cnki.spkj.2022.09.019.
- [14] 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. DOI:10.3969/j.issn.1005-1295.2022.02.016.
- [15] 周涛, 张志勇, 韩宁, 等. 基于二维相关近红外光谱的白酒酒龄鉴别[J]. 食品与机械, 2022, 38(12): 56-59; 98. DOI:10.13652/j.spjx.1003.5788.2022.80606.
- [16] 李展鸿. 基于便携式光谱仪的白酒掺假检测研究[D]. 无锡: 江南大学, 2021. DOI:10.27169/d.cnki.gwqgu.2021.000276.
- [17] 张卫卫, 刘建学, 韩四海, 等. 白酒基酒中醛类物质的傅里叶变换近红外光谱检测[J]. 食品科学, 2016, 37(6): 111-115. DOI:10.7506/spkx1002-6630-201606019.
- [18] 熊雅婷, 李宗朋, 王健, 等. 近红外光谱波段优化在白酒酒酯成分分析中的应用[J]. 光谱学与光谱分析, 2016, 36(1): 84. DOI:10.3964/j.issn.1000-0593(2016)01-0084-07.
- [19] 黄锬钊, 贺子豪, 王玉荣, 等. 基于智能感官和GC-MS技术分析市售酱香型白酒的品质[J]. 中国酿造, 2023, 42(12): 232-236. DOI:10.11882/j.issn.0254-5071.2023.12.036.
- [20] 胡雪. 基于质谱结合化学计量学对白葡萄酒产地、香型和等级判别分析[D]. 自贡: 四川轻化工大学, 2021. DOI:10.27703/d.cnki.gsclg.2021.000268.
- [21] 张健, 尹宝华, 廉哲, 等. 基于GC-MS技术的化学计量学方法鉴别酱香型白酒初探[J]. 刑事技术, 2024, 49(1): 85-92. DOI:10.16467/j.1008-3650.2023.0037.
- [22] 宋丹丹, 何鹏辉, 陈娟, 等. 气相色谱-质谱联用检测分析六大蒸馏酒中挥发性成分差异[J]. 中国酿造, 2020, 39(6): 190-195. DOI:10.11882/j.issn.0254-5071.2020.06.036.
- [23] 任玉兰, 田密, 李春彦, 等. 白酒中微量组分的气相色谱分析[J]. 中国酿造, 2011, 30(7): 177-179. DOI:10.3969/j.issn.0254-5071.2011.07.050.
- [24] YAN Y, CHEN S, NIE Y, et al. Quantitative analysis of pyrazines and their perceptual interactions in soy sauce aroma type Baijiu[J]. Foods, 2021, 10(2): 441. DOI:10.3390/foods10020441.
- [25] MORO M K, DE CASTRO E V R, ROMÃO W, et al. Data fusion applied in near and mid infrared spectroscopy for crude oil classification[J]. Fuel, 2023, 340: 127580. DOI:10.1016/j.fuel.2023.127580.
- [26] ZHOU Q Y, DAI Z H, SONG F H, et al. Monitoring black tea fermentation quality by intelligent sensors: comparison of image, e-nose and data fusion[J]. Food Bioscience, 2023, 52: 102454. DOI:10.1016/j.fbio.2023.102454.
- [27] ZHENG H T, YUAN J B, CHEN L. Short-term load forecasting using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation[J]. Energies, 2017, 10(8): 1168. DOI:10.3390/en10081168.
- [28] GÜNDOĞDU S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique[J]. Multimedia Tools and Applications, 2023, 82: 34163-34181. DOI:10.1007/s11042-023-15165-8.
- [29] 王雨莹, 戴宇佳, 王悦悦, 等. 基于近红外光谱技术的香榧蛋白质和脂肪含量无损检测方法研究[J]. 食品工业科技, 2024, 45(18): 250-257. DOI:10.13386/j.issn1002-0306.2023100276.
- [30] 朱雪梅, 庾先国, 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. DOI:10.13982/j.mfst.1673-9078.2023.1.0271.
- [31] 程惠珠, 杨婉琪, 李福生, 等. 面向XRF的竞争性自适应重加权算法和粒子群优化的支持向量机定量分析研究[J]. 光谱学与光谱分析, 2023, 43(12): 3742-3746. DOI:10.3964/j.issn.1000-0593(2023)12-3742-05.
- [32] 刘树鑫, 周柱, 刘洋, 等. 基于振动信号的交流接触器触头系统退化阶段划分[J]. 高电压技术, 2023, 49(12): 4971-4981. DOI:10.13336/j.1003-6520.hve.20221773.
- [33] 兰文宝, 车畅, 陶成云. 基于斯皮尔曼等级相关的单演谱成分选择及其在SAR目标识别中的应用[J]. 电波科学学报, 2020, 35(3): 414-421. DOI:10.13443/j.cjors.2019063001.
- [34] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334: 1518-1524. DOI:10.1126/science.1205438.
- [35] HEIDARI M, MOATTAR M H, GHAFARI H. Forward propagation dropout in deep neural networks using Jensen-Shannon and random forest feature importance ranking[J]. Neural Networks, 2023, 165: 238-247. DOI:10.1016/j.neunet.2023.05.044.