基于近红外光谱技术的六大茶类快速识别

张灵枝[1,6]，黄 艳[1,2,5,6]，于英杰[3]，林 刚[4]，孙威江[*1,5,6]

（1. 福建农林大学 园艺学院，福建 福州 350002；2. 福建农林大学 安溪茶学院，福建 泉州 362400；3. 中国茶叶流通协会，北京 100801；4. 福建融韵通生态科技有限公司，福建 福州 350025；5. 福建农林大学 福建省茶产业工程技术研究中心，福建 福州 350002；6. 福建农林大学 海峡两岸特色作物安全生产省部共建协同创新中心，福建 福州 350002）

摘要：为构建高质量的六大茶类识别模型，本研究中收集了370份样品，通过采集其近红外光谱（near-infrared spectroscopy，NIRS），结合光谱预处理、特征提取以及数据挖掘分类器算法，建立六大茶类快速识别模型。结果表明：1）支持向量机（support vector machine，SVM）与随机森林（random forest，RF）分类器皆适于六大茶类快速识别模型的构建；2）SVM分类器更适于结合原始光谱（original spectrum，OS）建模，预处理易使基于该分类器建立的模型鉴别性能减弱；3）随机森林（RF）分类器更适用于预处理后光谱建模，所得模型较OS模型在识别正确率（recognition accuracy，RA）及受试者工作特征曲线下面积（area under the curve，AUC）均得到明显提升；4）特征提取中线性判别分析（linear discriminant analysis，LDA）算法表现最好，所得模型的RA较OS模型明显提升，其中最佳模型OS-LDA-SVM的RA为100.00%，AUC为1.00，识别正确率高、泛化能力强、模型性能优异，可产业化应用。综上所述，近红外光谱结合预处理、特征提取算法及分类器建立模型，进行六大茶类识别的可行性强，模型的识别正确率高、性能优异，可为茶叶贸易的茶类快速识别提供科学、准确、高效的技术支撑，为国际茶类识别模型的产业化应用奠定基础。

关键词：近红外光谱；茶类识别；支持向量机；随机森林；线性判别分析

中图分类号：O657.33；TS 272.4 **文章编号：**1673-1689（2024）01-0048-12 **DOI：**10.12441/spyswjs.20221010001

Rapid Identification of Six Major Tea Categories Based on Near-Infrared Spectroscopy

ZHANG Lingzhi[1,6]，*HUANG Yan*[1,2,5,6]，*YU Yingjie*[3]，*LIN Gang*[4]，*SUN Weijiang*[*1,5,6]

（1. College of Horticulture, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2. Anxi College of Tea Science, Fujian Agriculture and Forestry University, Quanzhou 362400, China; 3. China Tea Marketing Association, Beijing 100801, China; 4. Fujian Rongyuntong Ecological Technology Limited Company, Fuzhou 350025, China; 5. Engineering Technology and Research Center of Fujian Tea Industry, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 6. Ministerial and Provincial Joint Innovation Centre for Safety Production of Cross-Strait Crops, Fujian Agriculture and Forestry University, Fuzhou 350002, China）

收稿日期：2022-10-10　**修回日期：**2022-11-21

基金项目：国家“十三五”重点研发计划项目（2019YFD1001601）；中国白茶研究院开放课题项目（KHCZ2101A，KHCZ2104A）；福建农林大学茶产业链科技创新与服务体系建设项目（K1520005A04）；福建张天福茶叶发展基金会科技创新基金项目（FJZTF01）。

*** 通信作者：**孙威江（1964—），男，博士，教授，博士研究生导师，主要从事茶叶品质化学与标准化研究。E-mail:swj8103@126.com

Abstract: In order to construct a high-quality recognition model for the six major tea categories, this study selected 370 samples and collected their near-infrared spectroscopy(NIRS). A rapid recognition model for the six major tea categories was developed by combined these data with spectral pre-processing, feature extraction and data mining classifier algorithms. The results indicated that: 1) Support vector machine(SVM) and random forest(RF) classifiers were both suitable for constructing rapid identification models for the six tea categories. 2) The SVM classifier was more suitable for modeling with the original spectrum (OS), and pre-processing algorithms tended to weaken the discriminatory performance of the models based on this classifier. 3) The RF algorithm was more suitable for modeling with pre-processing spectra, and the resulting models had a significant improvement in recognition accuracy (RA) and area under the curve (AUC) of the receiver operating characteristic curve compared to the OS models. 4) Among the feature extraction algorithms, the linear discriminant analysis (LDA) algorithm performed the best, yielding models with significantly improved RA compared to OS models. The optimal model, OS-LDA-SVM, achieved RA of 100.00% and AUC of 1.00, demonstrating high recognition rate, strong generalization ability, excellent model performance, and potential in industrial application. In summary, NIRS combined with pre-processing, feature extraction algorithms and classifiers to build models for the identification of the six tea categories was highly feasible. The models have high recognition accuracy and excellent performance, providing scientific, accurate, and efficient technical support for the rapid identification of tea categories in the tea trade, which could lay the foundation for the industrial application of international tea category identification models.

Keywords: near-infrared spectroscopy (NIRS), tea category recognition, support vector machine (SVM), random forest(RF), linear discriminant analysis (LDA)

茶叶富含茶多酚、氨基酸、生物碱等多种对人体健康有益的成分，是世界上最受欢迎的三大无酒精饮料之一。根据加工工艺及发酵程度不同，茶叶可分为绿茶（未发酵）、白茶（微发酵）、黄茶（轻发酵）、乌龙茶（半发酵）、红茶（全发酵）以及黑茶（后发酵）[1-3]六大类。六大茶类识别主要通过专业人士对茶叶外形、风味进行感官审评，结果易受审评人员的身体、精神状况等影响，审评主观性和经验性强。国际层面，茶叶评价术语尚未实现规范化、统一化和标准化，化学分析的国际标准制定严重滞后，使得国外非专业人士对六大茶类的分辨更加困难[4-6]，严重制约着茶叶国际贸易和流通。因此，不少学者在化学分析技术、计算机视觉技术、光谱技术等领域开展了多茶类判别研究。Ning 等以儿茶素、咖啡碱等生化成分含量结合 Fisher 判别分析对六大茶类进行逐步判别，其最佳模型的识别正确率达到 88.30%以上[7]；Wu 等以茶多酚、咖啡碱等生化成分含量结合逐步线性回归判别分析，建立绿茶、白茶、乌龙茶以及红茶的判别函数，识别正确率达 97.80%[8]；Jiang 等基于氨基酸、儿茶素等生化成分含量结合线性判别分析，对六大茶类可实现 100.00%准确鉴别[9]；Zhang 等基于非靶向代谢组学识别绿茶、白茶以及黄茶，找出了关键分类指标，分类识别正确率可达 100.00%[10]。上述研究虽然建立了优异的判别函数和模型，但存在特征成分检测耗时长、成本高、分析复杂等问题，不易进行快速无损鉴别，难以在国际上实现产业化应用。Wu 等通过采集红茶、绿茶、乌龙茶的外形图片，利用机器学习和计算机视觉技术，结合 K 近邻（K-nearest neighbor，KNN）分类器构建判别模型，识别正确率可达 94.7%[11]。Ning 等将高光谱成像技术结合化学计量学建立的 Lib-SVM 模型对绿茶、黄茶、白茶、黑茶、乌龙茶的识别正确率可达 98.39%[12]。由于茶叶的年份差异导致的波动性较大、高光谱仪普及率低等问题，上述两项技术并未实现产业化应用。

近红外光谱作为绿色分析技术[13]，具有高效、便捷、准确性高等优点，已在食品[14-16]、炼油[17-18]、药物[19-20]等领域广泛应用。本研究中以 NIRS 结合机器

学习算法,探索六大茶类鉴定的可行性,包括以下 4 个步骤:1)获取六大茶类在 3 600~12 500 cm^{-1} 波段的近红外光谱;2) 应用最小最大归一化(minmax scaler,Minmax)、连续小波变换(continuous wavelet transform,CWT)、标准正态变换 (standard normal variate,SNV)及多元散射校正(multiplicative scatter correction,MSC)4 种预处理算法, 建模分析他们对 OS 的去噪和散射校正性能;3) 比较主成分分析(principal component analysis,PCA)、线性判别分析(linear discriminant analysis,LDA) 和连续投影算法(successive projections algorithm,SPA)3 种方法提取茶叶光谱特征的能力;4)建立基于 RF 和 SVM 分类器的茶类鉴别模型,实现对六大茶类的快速、无损识别,为近红外光谱技术在茶类识别的产业应用上奠定理论基础和科学依据。

1 材料与方法

1.1 材料与仪器

共收集 370 份六大茶类样品,包括 122 种乌龙茶、110 种绿茶、55 种红茶、34 种黑茶、29 种白茶、20 种黄茶,来自中国福建、广东、台湾等地区,以及日本、斯里兰卡等国家。所有茶叶样品均具有该茶类正常的商品外形及特有的色、香、味,无异味、无劣变、无污染、无非茶类夹杂物、无任何添加剂,满足实验材料要求,具体信息详见表 1。

MPA 型傅里叶变换近红外光谱仪:德国布鲁克光谱仪器公司;高速粉碎机:上海鼎广机械设备有限公司;CFJ-II 茶叶筛分机: 杭州大吉光电仪器有限公司。

表 1 各茶类样品信息

Table 1 Sample information of each tea category

茶类	品种	产地
乌龙茶	水仙、肉桂、大红袍、铁观音、广东奇兰、凤凰单枞、高山茶等	中国福建、广东、台湾地区
绿茶	勐海大叶种烘青、西湖龙井、阳羡雪芽、崂山绿芽、竹叶青等	中国云南、福建、浙江、江苏、安徽、山东、四川、湖南等地区
红茶	坦洋工夫、正山小种、井冈红、滇红、祁红、信阳红、印度红茶 BOPF1	中国福建、江西、云南、安徽、河南等地区,斯里兰卡、印度等国家
黑茶	藏茶、云抗 10、手筑茯砖、安化千两茶	中国四川、云南、湖南地区
白茶	福鼎大白、政和大白、福安大白、福鼎大毫	中国福建地区
黄茶	石花、蒙顶甘露、君山银针	中国四川、湖南地区

1.2 实验方法

1.2.1 样品制备 茶叶粉碎后,置于茶叶筛分机中过 80 目筛,取筛下茶粉 5 g,密封编号,放于 4 ℃冰箱中备用。

1.2.2 光谱采集 使用 MPA 型傅里叶变换近红外光谱仪采集样品光谱信息。仪器工作时温度控制在 25 ℃,相对湿度<70%。光谱采集工作流参数:波数为 3 600~12 500 cm^{-1},光谱扫描次数为 64 次,分辨率为 8.0 cm^{-1}。为确保近红外光谱检测数据的可靠性,每个样品扫描 3 次,取平均光谱作为原始光谱数据进行后续分析。

1.2.3 光谱预处理 采用 Minmax 算法增强数据;选用 CWT 算法校正基线漂移并消除高频噪声;使用 SNV 及 MSC 算法校正散射, 消除因茶粉粒径的不均匀、光程不恒定等因素所带来的影响。

1.2.4 特征提取 为提升模型性能、运算效率,选用 PCA、LDA 及 SPA 方法进行光谱数据的特征提取,降低数据维数。

1.2.5 模型构建与评价 数据挖掘分类器广泛应用于 NIRS 数据的分析与利用, 不存在始终保持最优效果的分类器,因此使用多种分类器建模更利于优质模型的构建[21]。本文中基于 RF、SVM 两种分类器,结合不同预处理、特征提取方法,优化模型参数,探究六大茶类最佳识别模型构建流程。

为确保所建模型的适用性, 将数据按照 3:1 的比例划分为训练集和验证集两个子集,样本数分别为 277 和 93 个,其中训练集用于模型训练,验证集用于测试模型稳健性。使用 RA、AUC 以及混淆矩阵作为模型精度及性能的评价指标。

1.2.6 数据处理软件 数据处理软件包括 MATLAB 2016a、Origin 2019b、Excel 及 Python。

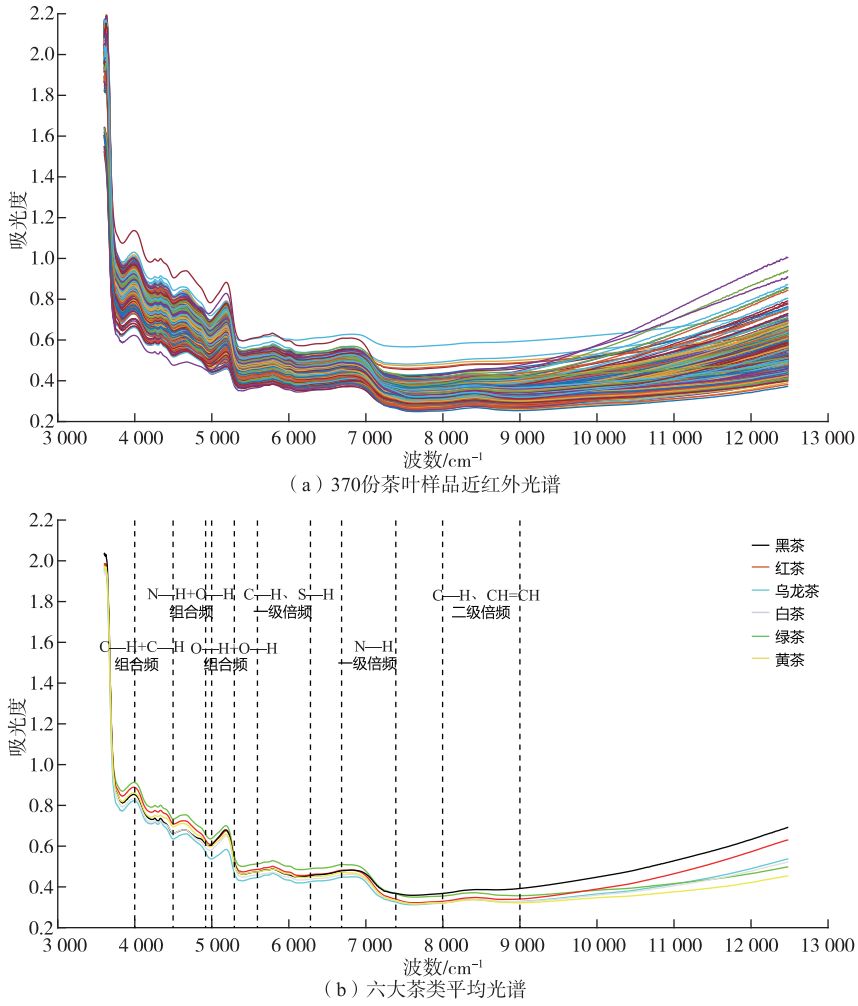
2 结果与分析

2.1 光谱分析

370 份茶叶样品在 3 600~12 500 cm^{-1} 波段的近红外光谱如图 1(a)所示,各波段吸光度变化趋势趋于一致。随着波数的增加,吸光度总体呈现下降趋势,变化范围处于 0.249~2.196。

六大茶类的平均光谱在大多波段范围内趋势一致(见图 1(b)),吸收光谱趋于平行,各茶类对应谱图于 3 750 cm^{-1} 处初步分开,于 3 750~9 000 cm^{-1} 处波动最为明显。受 C—H+C—H 组合频伸缩、变形振动影响,4 000 cm^{-1} 处出现明显波峰;参照相关文献[22-25],4 500~4 950 cm^{-1} 的谱带变化可归因于 N—H+O—H 的组合频;5 000~5 300 cm^{-1} 的谱带变化主要受 O—H+O—H 的组合频振动影响;5 600~6 300 cm^{-1} 的谱带变化受 C—H 和 S—H 的一级倍频振动影响;6 700~7 400 cm^{-1} 的谱带变化主要受 N—H 的一级倍频振动影响;8 000~9 000 cm^{-1} 中出现的波峰可能与 CH═CH 及 C—H 的二级倍频振动有关。

3 750~9 000 cm^{-1} 中各平均光谱能基本分开,说明六大茶类样品吸光度在该波段增减性不同,即光谱信息与茶类间具有相关性。首段谱带中各平均光谱间交叉重叠现象频发,说明该波段体现的与产地相关的有效信息较少,信噪比低;末端谱带波动同原始光谱一致,趋于平缓,无明显波峰波谷,特征信息不显,因此结合 Meng 等结论[23],本文中的模型构建使用 3 750~9 000 cm^{-1} 波段的光谱数据。



(a) 370份茶叶样品近红外光谱
(b) 六大茶类平均光谱

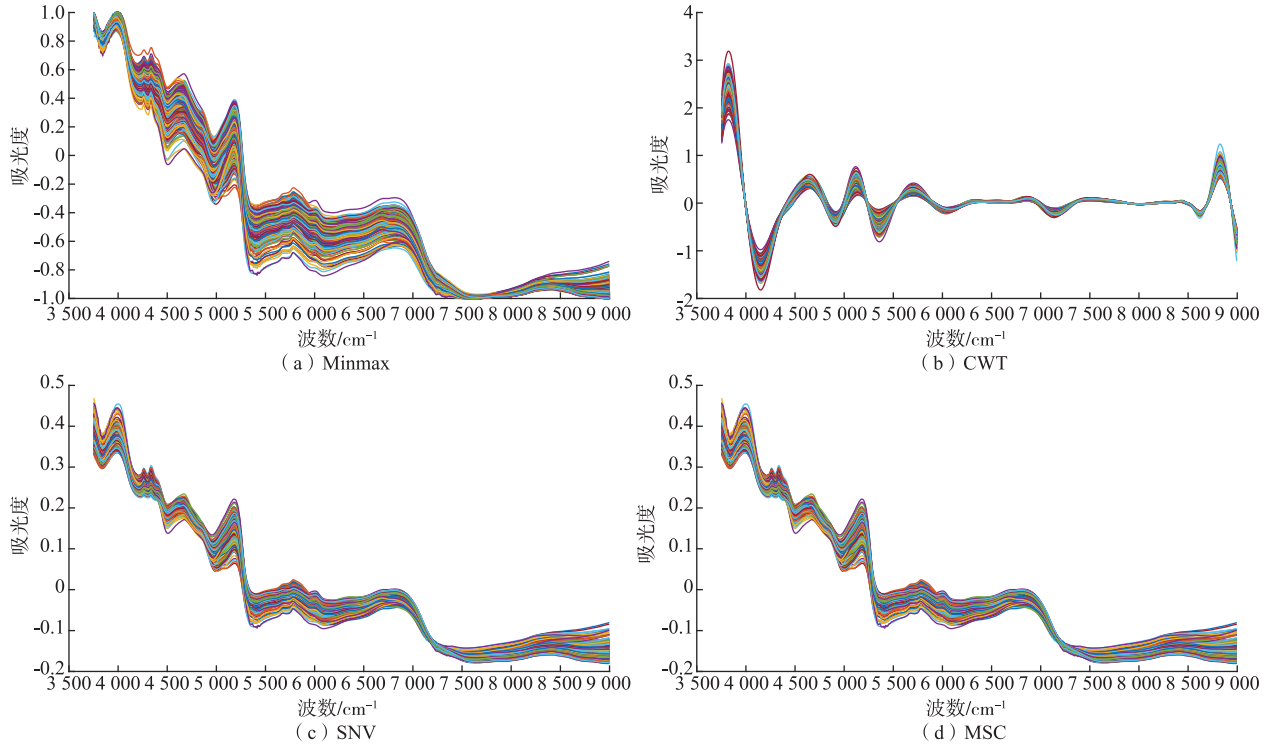
图 1 370 份茶叶样品近红外光谱和六大茶类平均光谱图

Fig. 1 Near-infrared spectra of 370 tea samples and average spectra of six major tea types

2.2 光谱预处理

为校正光谱采集过程中因环境、光程不恒定及样品粒径差异等因素所带来的误差，从数据增强、基线校正、散射校正等角度，使用 Minmax、CWT、SNV 及 MSC 算法对 OS 进行预处理。CWT 中小波参数(小波基、分解尺度)的选择至关重要，直接决定了后续模型的优劣，经比对分析后，本研究中选择应用最为广泛的 db(daubechies)族中的 db4 小波基，分解尺度定为 100[26]。

从光谱变化情况(见图 2)可知，4 种处理方式都使光谱形态发生了较大改变。Minmax(见图 2(a))将光谱吸光度凝练到-1.0~1.0，增强了数据，消除了数据量纲及取值范围的影响，后续可使所建模型收敛速度加快，提高模型性能。应用 CWT 进行光谱预处理如图 2(b)所示，其形态变换程度为 4 种预处理方式中最大，基线漂移、背景干扰、噪音现象等得到明显消除，谱峰更清晰，差异信息段更明显。受茶粉颗粒大小不均、产生的散射影响，采用 SNV 与 MSC 进行预处理(见图 2(c)和图 2(d))，处理后光谱中的散射干扰得到明显消除，特征信息更加突出。相较于 OS，预处理可有效消除光谱中因光散射、基线漂移等造成的信号干扰，但处理后的光谱图仍无法直观分辨茶类间的差异，这可能是由于不同茶类在内含物的组成与含量上具有较多的相似性。为了进一步评估各处理对建模结果的影响，将各处理所得的光谱吸光度分别作为模型的输入变量，后续基于模型评价指标判断预处理效果。



图中为经预处理后的 3 750~9 000 cm^{-1} 谱图。

图 2 光谱预处理效果

Fig. 2 Effect of spectral pretreatment

2.3 光谱数据降维

近红外光谱的连续波数中存在大量的冗余信息，其与特征信息之间存在很强的相关性[27]。通过选取特征向量或波数降低数据维数，可保留原始数据中的主要特征信息，减少后续处理的计算任务。本研究中采用 PCA、LDA 和 SPA 方法进行数据降维。

2.3.1 PCA PCA 是一种常用于降低大数据集维数的无监督特征提取方式，能从大量数据中提取出特征，转换为仍包含绝大部分有效信息却拥有较小维数的数据集，最大程度保留原始数据信息，因而 PCA 是一种最优、最为常用的方法[28]。使用 PCA 方法对 3 750~9 000 cm^{-1} 中 OS 及经 Minmax、CWT、SNV 和 MSC 预处理后的 4 种光谱进行降维，截取

前 15 个主成分特征值与累积贡献度。结果如表 2 所示,以特征值大于 1、累积贡献度大于 80%为原则,筛选模型输入主成分个数。OS 以及 Minmax、CWT、SNV 和 MSC 预处理后光谱分别筛出 6、11、13、12、12 个主成分,累积贡献度分别达到 99.89%、99.71%、99.67%、99.82%、99.82%,符合原则。基于筛选的主成分构建模型。

表 2 PCA 特征值及累积贡献度

Table 2 PCA feature values and cumulative contribution

主成分个数	OS		Minmax 预处理光谱		CWT 预处理光谱		SNV 预处理光谱		MSC 预处理光谱	
	特征值	累积贡献度/%	特征值	累积贡献度/%	特征值	累积贡献度/%	特征值	累积贡献度/%	特征值	累积贡献度/%
1	1 092.430 94	80.15	900.45	66.06	584.892 04	42.91	920.424 91	67.53	920.424 91	67.53
2	246.898 59	98.26	192.57	80.19	331.804 91	67.26	193.571 98	81.73	193.571 98	81.73
3	10.289 27	99.02	118.34	88.87	152.378 51	78.44	121.729 43	90.66	121.729 43	90.66
4	7.672 56	99.58	58.45	93.16	126.068 36	87.68	69.751 09	95.78	69.751 09	95.78
5	3.055 28	99.81	40.67	96.15	87.458 64	94.10	19.443 17	97.21	19.443 17	97.21
6	1.139 92	99.89	19.91	97.61	22.544 93	95.76	15.843 92	98.37	15.843 92	98.37
7	0.660 02	99.94	13.43	98.59	20.176 24	97.24	9.099 92	99.04	9.099 92	99.04
8	0.259 52	99.96	6.87	99.10	9.860 38	97.96	3.895 10	99.32	3.895 10	99.32
9	0.241 39	99.97	3.55	99.36	8.920 74	98.61	2.330 94	99.49	2.330 94	99.49
10	0.099 47	99.98	3.17	99.59	6.502 02	99.09	1.870 47	99.63	1.870 47	99.63
11	0.075 09	99.99	1.65	99.71	3.202 10	99.33	1.518 50	99.74	1.518 50	99.74
12	0.055 61	99.99	0.98	99.78	2.838 48	99.53	1.081 55	99.82	1.081 55	99.82
13	0.029 84	99.99	0.91	99.85	1.802 64	99.67	0.637 79	99.87	0.637 79	99.87
14	0.022 34	99.99	0.48	99.89	0.958 05	99.74	0.417 10	99.90	0.417 10	99.90
15	0.016 44	100.00	0.31	99.91	0.898 56	99.80	0.289 51	99.92	0.289 51	99.92

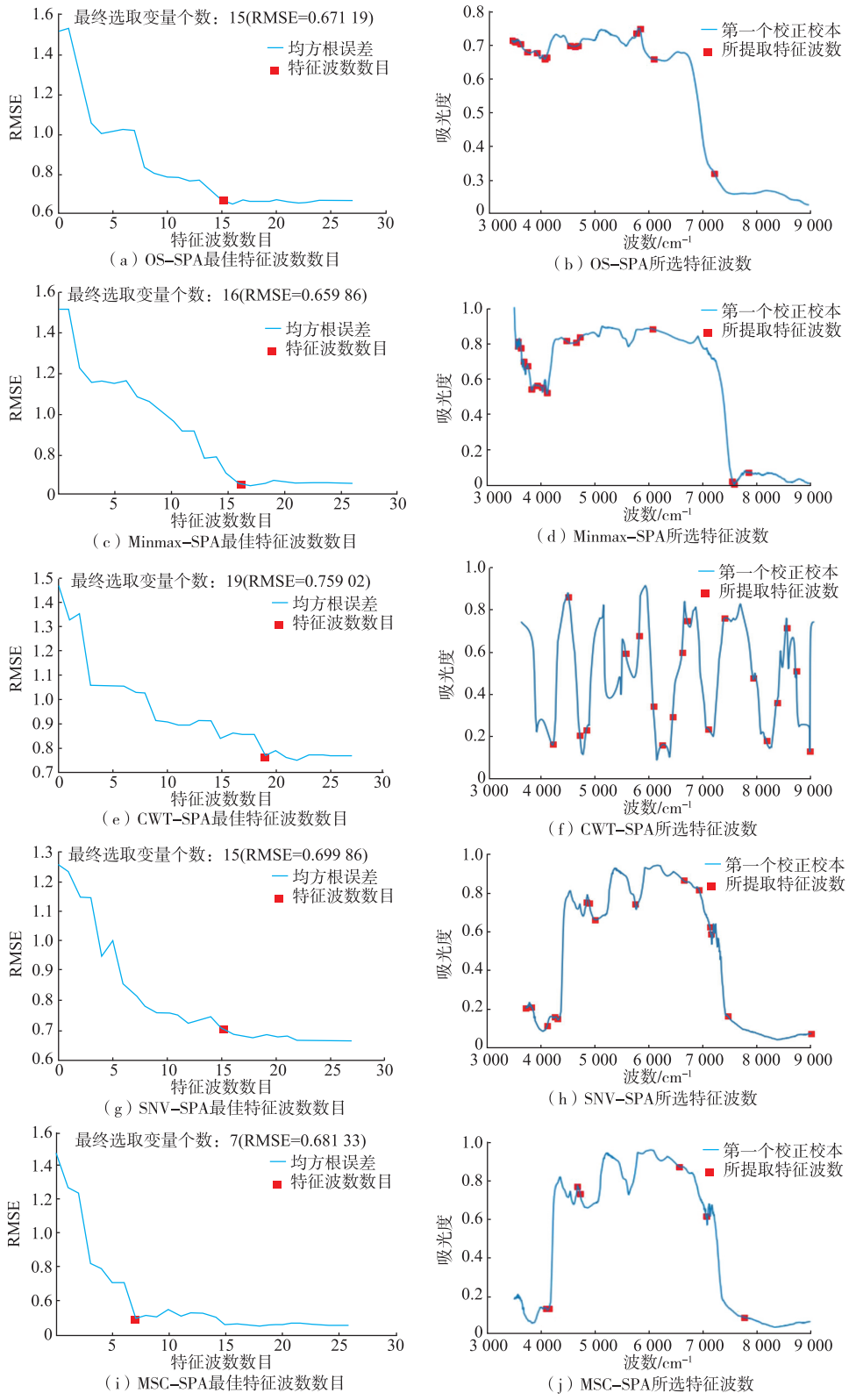
2.3.2 LDA LDA 是一种有监督的特征提取方法[29],在茶叶领域常作为分类器使用。而利用 LDA 进行光谱特征提取、降低数据维数,并结合分类器建立茶类识别模型的研究,目前尚未见相关报道。LDA 最多可使数据矩阵降至类别数减 1 的维数,降低维数的同时不过多丢失原始信息,LDA 降维后所得维数将被用于后续建模。

2.3.3 SPA SPA 是一种前向循环特征提取方法,其可以通过连续投影的方式从原始光谱矩阵中提取有效预测响应变量的信息,最大限度减少光谱变量之间的共线性效应,达到所选响应变量预测能力的最大化。其主要通过将波数投影于其他波数,比较投影向量大小,波数间投影向量大者为待选波数,最终投影向量最大且与特征集内波数共线性最小的波数选入特征集合[30]。特征波数的数目由校准集内部完全交叉验证的均方根误差(root mean square error,RMSE)确定,与最小 RMSE 对应的特征波数数目和特征波数为最佳值[31-32]。

由 SPA 方法提取的特征向量如图 3 所示。SPA 中 4 种预处理后的光谱信息的 RMSE 迭代下降曲线分别如图 3(a)、图 3(c)、图 3(e)、图 3(g)和图 3(i)所示。从图中可以看出,当选择特定数量的波数时,RMSE 达到最小值;而后 RMSE 虽仍波动下降,但降幅很小且导致所选波数增加,没有必要为了追求微小的 RMSE 而增加维数。因此,最终从 OS 及经 Minmax、CWT、SNV 和 MSC 预处理光谱中获得的特征波数数目分别为 15、16、19、15、7 个(如图 3 所示,特征波数具体信息如表 3)。

2.4 模型参数优化

SVM 是近年来茶叶中应用最广、效果最好的机器学习方法之一。它是一种利用核函数将 n 维输入向量映射到 K 维特征空间($K>n$),从而通过高维特征空间进行分类的算法[33]。为提高模型质量,本文中所有 SVM 模型皆基于高斯(radial basis function,RBF)核函数,该核函数可降低训练过程的计算复杂度,在一般平滑假设下具有良好性能;与此同时,惩罚参数及 gamma 参数最优值的确定也至关重要,SVM 模型精度取决于这两个参数的组合。根据初步



（a）OS-SPA最佳特征波数数目　（b）OS-SPA所选特征波数
（c）Minmax-SPA最佳特征波数数目　（d）Minmax-SPA所选特征波数
（e）CWT-SPA最佳特征波数数目　（f）CWT-SPA所选特征波数
（g）SNV-SPA最佳特征波数数目　（h）SNV-SPA所选特征波数
（i）MSC-SPA最佳特征波数数目　（j）MSC-SPA所选特征波数

SPA 提取特征波数包括 OS-SPA、Minmax-SPA、CWT-SPA、SNV-SPA、MSC-SPA 选择的最佳特征波数数目以及 OS-SPA、Minmax-SPA、CWT-SPA、SNV-SPA、MSC-SPA 提取的特征波数。

图 3　连续投影算法(SPA)提取特征波数

Fig. 3　Extraction of feature wavenumbers by successive projections algorithm (SPA)

表 3　通过 SPA 筛选出的特征波数

Table 3　Feature wavenumbers selected by SPA

预处理方式	特征波数/cm^{-1}	参数个数
未处理(OS)	3 749.17、3 791.60、3 803.18、3 899.60、4 011.46、4 196.61、4 323.89、4 358.61、4 779.04、4 860.04、4 914.04、5 959.34、6 013.34、6 256.34、7 336.35	15
Minmx	3 803.18、3 814.75、3 864.89、3 907.32、3 992.18、4 053.89、4 154.18、4 235.18、4 323.89、4 671.04、4 852.33、4 921.76、6 217.77、7 633.35、7 656.49、7 922.64	16
CWT	4 320.04、4 593.90、4 798.33、4 914.04、5 619.91、5 862.91、6 117.48、6 275.62、6 468.48、6 634.34、6 726.91、7 112.63、7 405.78、7 911.07、8 154.07、8 343.07、8 528.22、8 694.07、8 940.93	19
SNV	3 749.17、3 837.89、4 138.75、4 277.61、4 323.89、4 860.04、4 914.04、5 022.04、5 762.62、6 653.63、6 942.92、7 139.63、7 155.06、7 471.35、9 002.65	15
MSC	4 320.04、4 362.47、4 867.76、4 917.90、6 672.91、7 151.20、7 833.92	7

注:包括 3 750、9 000 cm^{-1} 附近波数。

试算结果,将惩罚参数取值定为 1×10^{3}、1×10^{4}、1×10^{5}、1×10^{6},gamma 参数取 1×10^{0}、1×10^{-1}、1×10^{-2}、1×10^{-3}。将 3 750~9 000 cm^{-1} 的 OS 数据作为输入量,结合模型的识别正确率进行参数优化。结果如表 4 所示,当惩罚参数为 1×10^{6}、gamma 参数为 1×10^{-2} 时模型具最佳识别正确率,后续模型皆基于此参数构建。

表 4　SVM 模型参数优化

Table 4　Parameter optimization of SVM model

惩罚参数	gamma 参数	识别正确率/%
1×10^{3}	1×10^{0}	75.27
1×10^{3}	1×10^{-1}	88.17
1×10^{3}	1×10^{-2}	87.10
1×10^{3}	1×10^{-3}	75.27
1×10^{4}	1×10^{0}	75.27
1×10^{4}	1×10^{-1}	87.10
1×10^{4}	1×10^{-2}	87.10
1×10^{4}	1×10^{-3}	87.10
1×10^{5}	1×10^{0}	75.27
1×10^{5}	1×10^{-1}	87.10
1×10^{5}	1×10^{-2}	91.40
1×10^{5}	1×10^{-3}	89.25
1×10^{6}	1×10^{0}	75.27
1×10^{6}	1×10^{-1}	87.10
1×10^{6}	1×10^{-2}	92.47
1×10^{6}	1×10^{-3}	89.25

RF 是一种有监督的集成分类算法,主要是为解决单一决策树可能出现的很大误差和过拟合的问题,在分类问题中表现优异,具有成为各情况下效果最优分类器的巨大潜力[34]。该模型由许多的决策树组成,但每一棵决策树之间没有关联,得到森林之后对新样本进行判断或预测时,将由森林中的每一棵决策树分别进行判断,分辨该样本属于哪类,比对出选择数最多的类别,从而对样本类别做出判断,因此该模型中树木数量的选择极为重要。OS 数据试算后,选定树木数量为 1~100 进行参数优化,结果如图 4 所示,树木数量为 70 时,RF 模型具最佳识别正确率,后续模型皆基于此参数构建。

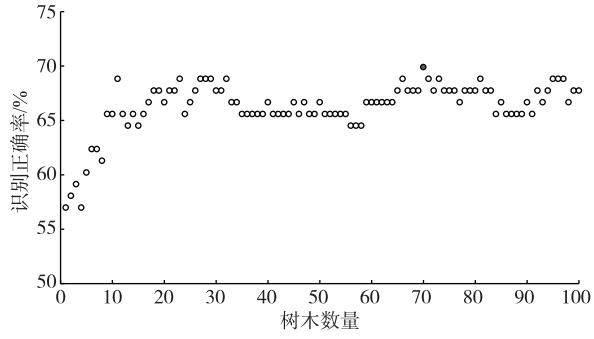


图 4　RF 模型参数优化

Fig. 4　Parameter optimization of RF model

2.5　模型性能评价

RA 常被用于模型预测能力的评估,AUC 则常被用于模型泛化能力的评价,取值区间为 0.5~1.0,值的大小与模型质量呈正相关。因此,采用 RA、AUC 联合评估模型性能。此外,为直观呈现所建模型对各茶类识别性能的优劣,引入混淆矩阵对模型预测结果进行评价。

由表 5 可知,各茶类 NIRS 数据结合不同预处理、特征提取方法及数据挖掘分类器最终获得 40

个茶类识别模型，识别正确率介于59.14%~100.00%,AUC处于0.70~1.00，大多数模型识别正确率高于70%,模型性能良好。OS结合RF、SVM所得模型识别正确率分别为69.89%及92.47%,光谱经预处理、特征提取后,建立的RF模型绝大多数识别正确率、AUC显著提升,模型精度、泛化能力均得到改善，最佳模型OS-LDA-RF的识别正确率可达94.62%,AUC可达0.96;SVM结合预处理后光谱建模效果欠佳,多出现识别准度下降的问题,不同特征提取方法中效果最佳的为LDA,结合不同预处理后的光谱数据皆优化了模型性能，基于SVM建立的茶类识别模型中最佳的是OS-LDA-SVM,识别正确率为100.00%,AUC为1.00。总体而言,不同分类器结合不同预处理、特征提取方法所取得的效果也不尽相同,在茶类识别模型构建中,RF适合与不同化学计量学方法结合,多数预处理、特征提取方法对RF模型性能优化效果显著;SVM适合基于去除头尾信息匮乏波段的原始光谱结合特征提取方法进行建模,所得模型特征数可得到简化,提升运算速度及模型质量。

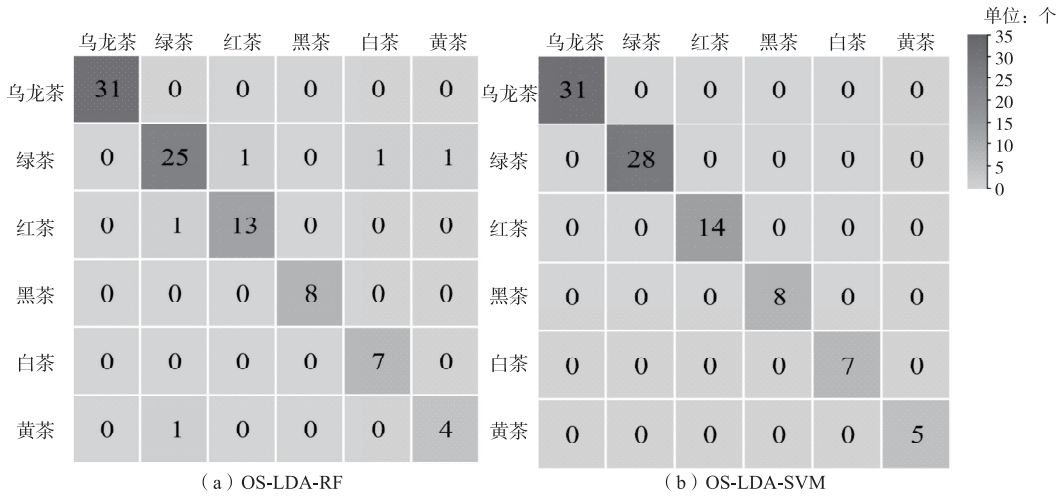
表5　基于RF及SVM构建的茶类识别模型性能评估

Table 5　Performance evaluation of tea recognition model based on RF and SVM

序号	分类模型	特征数	识别正确率/%	AUC	序号	分类模型	特征数	识别正确率/%	AUC
1	OS-RF	1 363	69.89	0.77	21	OS-SVM	1 363	92.47	0.94
2	Minmax-RF	1 363	77.42	0.79	22	Minmax-SVM	1 363	88.17	0.90
3	CWT-RF	1 363	88.17	0.90	23	CWT-SVM	1 363	90.32	0.91
4	SNV-RF	1 363	79.57	0.81	24	SNV-SVM	1 363	91.40	0.92
5	MSC-RF	1 363	81.72	0.82	25	MSC-SVM	1 363	90.32	0.92
6	OS-PCA-RF	6	82.80	0.86	26	OS-PCA-SVM	6	77.42	0.82
7	Minmax-PCA-RF	11	88.17	0.87	27	Minmax-PCA-SVM	11	90.32	0.92
8	CWT-PCA-RF	13	87.10	0.87	28	CWT-PCA-SVM	13	90.32	0.91
9	SNV-PCA-RF	12	88.17	0.88	29	SNV-PCA-SVM	12	89.25	0.90
10	MSC-PCA-RF	12	90.32	0.88	30	MSC-PCA-SVM	12	90.32	0.91
11	OS-LDA-RF	5	94.62	0.96	31	OS-LDA-SVM	5	100.00	1.00
12	Minmax-LDA-RF	5	88.17	0.92	32	Minmax-LDA-SVM	5	95.70	0.96
13	CWT-LDA-RF	5	91.40	0.94	33	CWT-LDA-SVM	5	95.70	0.96
14	SNV-LDA-RF	5	91.40	0.94	34	SNV-LDA-SVM	5	95.70	0.96
15	MSC-LDA-RF	5	91.40	0.94	35	MSC-LDA-SVM	5	95.70	0.96
16	OS-SPA-RF	15	59.14	0.70	36	OS-SPA-SVM	15	90.32	0.93
17	Minmax-SPA-RF	16	77.42	0.80	37	Minmax-SPA-SVM	16	82.80	0.86
18	CWT-SPA-RF	19	83.87	0.86	38	CWT-SPA-SVM	19	92.47	0.93
19	SNV-SPA-RF	15	78.49	0.79	39	SNV-SPA-SVM	15	90.32	0.91
20	MSC-SPA-RF	7	76.34	0.80	40	MSC-SPA-SVM	7	83.87	0.89

观察最佳模型混淆矩阵可知（见图5),OS-LDA-RF对乌龙茶、黑茶、白茶的识别正确率为100.00%,误判发生于绿茶、红茶及黄茶的识别中,三者识别正确率分别为89.28%、92.86%以及80.00%;OS-LDA-SVM对各茶类的识别正确率皆为100.00%,混淆矩阵中数值皆处于对角线,表明识别效果优异,模型质量好。

对性能最优模型OS-LDA-SVM进行三维空间可视化(见图6)。可发现,识别正确率达100.00%的茶类识别模型中,绿茶与黄茶的光谱特征于三维空间上的分布极为接近,可能与其加工工艺的高度相似有关,黄茶仅比绿茶多了闷黄工艺;还可能受黄茶样本数量较少导致特征信息不显的影响,黄茶在六大茶类中占比最小,仅在我国四川、湖南等少数省份小规模生产,后期将通过逐年增加黄茶样本数量的方式,强化黄茶光谱特征,提高模型性能。其他茶类的光谱特征在三维空间分布差异大,可能与加工工艺、茶树品种不同有关。



(a) OS-LDA-RF　　(b) OS-LDA-SVM

列为真实类别,行为预测类别。

图 5　最佳模型混淆矩阵

Fig. 5　Optimal model confusion matrix

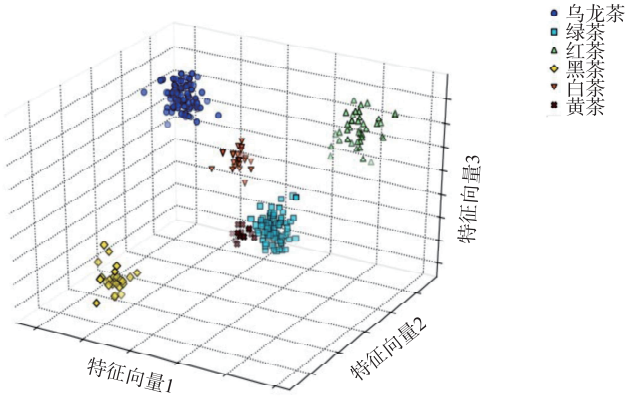


图 6　OS-LDA-SVM 模型三维空间效果

Fig. 6　Three-dimensional space effect of OS-LDA-SVM model

3 结　语

基于茶类的 370 个近红外光谱数据,剔除认为不含茶类相关信息的波段后,使用 Minmax、CWT、SNV 及 MSC 进行预处理,PCA、LDA 及 SPA 进行特征提取,最后基于 RF、SVM 构建茶类识别模型。主要结论如下:

1)茶类识别模型构建中,不同预处理对不同分类器的效果不尽相同。对 RF 模型而言,预处理对模型性能提升效果显著,而相同处理于 SVM 模型中效果欠佳,模型质量多不如原始光谱模型。

2)通过特征提取优化模型,效果显著。PCA、LDA、SPA 方法皆大幅度降低了数据维数,提高了模型运算效率。其中 LDA 效果最佳,与不同预处理方法、分类器结合所得模型皆质量优异。

3)RF、SVM 皆适用于茶类识别模型构建,相较 RF 模型,SVM 模型总体效果略胜一筹。RF 模型中最佳模型为 OS-LDA-RF,RA 为 94.62%,对乌龙茶、黑茶及白茶的 RA 可达 100.00%,AUC 为 0.96,模型性能优异、稳定;OS-LDA-SVM 为 SVM 模型中的最优模型,不同茶类的 RA 皆达 100.00%,AUC 为 1.00,模型质量高、泛化能力好。

不同茶类的近红外光谱数据,经适当预处理、特征提取方法的选择后,结合 RF、SVM 分类器可挖掘出近红外光谱中茶类识别相关的关键信息,构建出高识别正确率的茶类识别模型。基于此,针对光谱特征接近的茶类,可逐年扩大样本量,优化与验证茶类识别模型性能,提高模型适用性。除此之外,本研究后续将利用互联网技术构建六大茶类的近红外光谱数据库,搭建在线茶类识别平台,开展远程茶类识别,以期早日在国际市场上实现产业化应用,促进我国茶产业的高质量发展。

参考文献:

[1] JIANG H,YU F,QIN L,et al. Dynamic change in amino acids,catechins,alkaloids,and gallic acid in six types of tea processed from the same batch of fresh tea(*Camellia sinensis* L.) leaves[J]. **Journal of Food Composition and Analysis**,2019,77:28-38.

[2] HUANG Y,WANG T J,FILLET M,et al. Simultaneous determination of amino acids in different teas using supercritical fluid chromatography coupled with single quadrupole mass spectrometry[J]. **Journal of Pharmaceutical Analysis**,2019,9(4):254-258.
[3] GAO Y,XU Y Q,RUAN J Y,et al. Selenium affects the activity of black tea in preventing metabolic syndrome in high-fat diet-fed Sprague-Dawley rats[J]. **Journal of the Science of Food and Agriculture**,2020,100(1):225-234.
[4] DUTTA R,HINES E L,GARDNER J W,et al. Tea quality prediction using a tin oxide-based electronic nose:an artificial intelligence approach[J]. **Sensors and Actuators B:Chemical**,2003,94(2):228-237.
[5] BANERJEE M B,ROY R B,TUDU B P,et al. Black tea classification employing feature fusion of E-nose and E-tongue responses[J]. **Journal of Food Engineering**,2019,244:55-63.
[6] ZHANG S F,ZHU D H,CHEN X J. Analysis of E-tongue data for tea classification based on semi-supervised learning of generative adversarial network[J]. **Chinese Journal of Analytical Chemistry**,2022,50(2):77-85.
[7] NING J M,LI D X,LUO X,et al. Stepwise identification of six tea(*Camellia sinensis* L.) categories based on catechins,caffeine,and theanine contents combined with fisher discriminant analysis[J]. **Food Analytical Methods**,2016,9(11):3242-3250.
[8] WU Q J,DONG Q H,SUN W J,et al. Discrimination of Chinese teas with different fermentation degrees by stepwise linear discriminant analysis(S-LDA) of the chemical compounds[J]. **Journal of Agricultural and Food Chemistry**,2014,62(38):9336-9344.
[9] JIANG H,ZHANG M T,WANG D X,et al. Analytical strategy coupled to chemometrics to differentiate *Camellia sinensis* tea types based on phenolic composition,alkaloids,and amino acids[J]. **Journal of Food Science**,2020,85(10):3253-3263.
[10] ZHANG Q F,WU S,LI Y,et al. Characterization of three different classes of non-fermented teas using untargeted metabolomics [J]. **Food Research International**,2019,121:697-704.
[11] WU X Y,YANG J Q,WANG S H. Tea category identification based on optimal wavelet entropy and weighted K-nearest neighbors algorithm[J]. **Multimedia Tools and Applications**,2018,77(3):3745-3759.
[12] NING J M,SUN J J,LI S H,et al. Classification of five Chinese tea categories with different fermentation degrees using visible and near-infrared hyperspectral imaging[J]. **International Journal of Food Properties**,2017,20:1515-1522.
[13] CHEN Q S,ZHANG D L,PAN W X,et al. Recent developments of green analytical techniques in analysis of tea's quality and nutrition[J]. **Trends in Food Science & Technology**,2015,43(1):63-82.
[14] POREP J U,KAMMERER D R,CARLE R. On-line application of near infrared (NIR) spectroscopy in food production[J]. **Trends in Food Science & Technology**,2015,46(2):211-230.
[15] WU X H,WU B,SUN J,et al. Discrimination of apples using near infrared spectroscopy and sorting discriminant analysis[J]. **International Journal of Food Properties**,2016,19(5):1016-1028.
[16] GANGIDI R R,PROCTOR A,POHLMAN F W,et al. Rapid determination of spinal cord content in ground beef by near- infrared spectroscopy[J]. **Journal of Food Science**,2005,70(6):397-400.
[17] WANG S T,LIU S Y,YUAN Y Y,et al. Simultaneous detection of different properties of diesel fuel by near infrared spectroscopy and chemometrics[J]. **Infrared Physics & Technology**,2020,104:103111.
[18] LONG J,WANG K, YANG M L,et al. Rapid crude oil analysis using near-infrared reflectance spectroscopy[J]. **Petroleum Science and Technology**,2019,37(3):354-360.
[19] WAHL P R,FRUHMANN G,SACHER S,et al. PAT for tableting:inline monitoring of API and excipients via NIR spectroscopy [J]. **European Journal of Pharmaceutics and Biopharmaceutics**,2014,87(2):271-278.
[20] MEES C,KAUFFMANN J M,PIERNA J A F,et al. Benchtop NIR data standardization on handheld spectrometers to identify paracetamol tablets[J]. **Journal of Chemometrics**,2022,36(3):1-17.
[21] HONG X Z,FU X S,WANG Z L,et al. Tracing geographical origins of teas based on FT-NIR spectroscopy:introduction of model updating and imbalanced data handling approaches[J]. **Journal of Analytical Methods in Chemistry**,2019,2019:1-8.
[22] 李尚科,杜国荣,丁胜华,等. 近红外光谱结合化学计量法快速无损鉴别燕麦[J]. 食品与机械,2019,35(2):72-76.
LI S K,DU G R,DING S H,et al. Rapid and non-destructive identification of different brands and inferior oats by near infrared spectroscopy combined with chemometrics methods[J]. **Food & Machinery**,2019,35(2):72-76. (in Chinese)
[23] MENG W J,XU X N,CHENG K K,et al. Geographical origin discrimination of oolong tea(Tieguanyin,*Camellia sinensis*(L.)

O. Kuntze）using proton nuclear magnetic resonance spectroscopy and near-infrared spectroscopy[J]. **Food Analytical Methods**，2017，10(11)：3508-3522.

[24] CHEN Q S，CHEN M，LIU Y，et al. Application of FT-NIR spectroscopy for simultaneous estimation of taste quality and taste-related compounds content of black tea[J]. **Journal of Food Science and Technology**，2018，55(10)：4363-4368.

[25] SUN Y M，WANG Y J，HUANG J，et al. Quality assessment of instant green tea using portable NIR spectrometer [J]. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**，2020，240：118576.

[26] 李水芳，单杨，朱向荣，等. 近红外光谱结合化学计量学方法检测蜂蜜产地[J]. 农业工程学报，2011，27(8)：350-354.
LI S F，SHAN Y，ZHU X R，et al. Detection of geographical origin of honey using near-infrared spectroscopy and chemometrics [J]. **Transactions of the Chinese Society of Agricultural Engineering**，2011，27(8)：350-354.（in Chinese）

[27] HUANG L，ZHAO J W，CHEN Q S，et al. Rapid detection of total viable count（TVC）in pork meat by hyperspectral imaging[J]. **Food Research International**，2013，54(1)：821-828.

[28] ZHENG J H，JIAO S K，CUI D. Application of principal component analysis-assisted neural networks for the rotor blade load prediction[J]. **International Journal of Aerospace Engineering**，2021，2021：1-13.

[29] 罗玉琴，韦燕菊，林琳，等. 基于 GC-IMS 技术的福建白茶产地判别[J]. 农业工程学报，2021，37(6)：264-273.
LUO Y Q，WEI Y J，LIN L，et al. Origin discrimination of Fujian white tea using gas chromatography-ion mobility spectrometry [J]. **Transactions of the Chinese Society of Agricultural Engineering**，2021，37(6)：264-273.（in Chinese）

[30] 张楠楠，张晓，王城坤，等. 基于高光谱和连续投影算法的棉花叶面积指数估测[J]. 农业机械学报，2022，53(增刊 1)：257-262.
ZHANG N N，ZHANG X，WANG C K，et al. Cotton LAI estimation based on hyperspectral and successive projection algorithm [J]. **Transactions of the Chinese Society for Agricultural Machinery**，2022，53(Suppl. 1)：257-262.（in Chinese）

[31] YAO Z F，LEI Y，HE D J. Early visual detection of wheat stripe rust using visible/near-infrared hyperspectral imaging [J]. **Sensors**，2019，19(4)：952.

[32] GALVÃO R K H，ARAUJO M C U，FRAGOSO W D，et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm[J]. **Chemometrics and Intelligent Laboratory Systems**，2008，92(1)：83-91.

[33] HONG X Z，WANG J. Use of electronic nose and tongue to track freshness of cherry tomatoes squeezed for juice consumption：comparison of different sensor fusion approaches[J]. **Food and Bioprocess Technology**，2015，8(1)：158-170.

[34] DELGADO M F，CERNADAS E，BARRO S，et al. Do we need hundreds of classifiers to solve real world classification problems?[J]. **Journal of Machine Learning Research**，2014，15(1)：3133-3181.