

基于高光谱图谱融合技术的英德红茶等级快速无损判别

刘翠玲¹ 秦冬¹ 凌彩金^{2,3} 孙晓荣⁴ 郜礼阳^{2,3} 咎佳睿⁴

(1. 北京工商大学中国轻工业工业互联网与大数据重点实验室, 北京 100048;

2. 广东省农业科学院茶叶研究所, 广州 510640; 3. 广东省茶树资源创新利用重点实验室, 广州 510640;

4. 北京工商大学北京市食品安全大数据技术重点实验室, 北京 100048)

摘要: 茶叶等级评价是检测茶叶品质的一项重要技术指标。通过提取红茶高光谱成像技术下的图像特征和光谱特征,构建一种基于图谱融合方法、适用于英德红茶等级评价的快速无损判别模型。首先制备3种不同等级的红茶样本,采用 t 分布-随机近邻嵌入和主成分分析对光谱数据进行降维可视化分析,然后从影响内在品质角度用连续投影法提取每种化学值的特征波长,通过多模型共识策略和竞争性自适应重加权算法-连续投影法筛选得出表征其内在品质的最佳特征波长组合,并建立基于遗传算法优化支持向量机的等级判别模型;其模型的训练集准确率为88%,预测集准确率为78.33%。为了融合外形纹理差异,先提取最佳特征波长组合对应的高光谱图像;采用图像掩膜消除背景的干扰和采用图像主成分分析消除多波长图像间的冗余信息,然后采用灰度共生矩阵和局部二值化算法提取主成分前三维主成分图像与特征光谱融合,并建立基于特征融合的遗传算法优化支持向量机等级判别模型,且基于第三主成分图像特征融合模型判别效果最佳,训练集准确率提升至98%,预测集准确率提升至96.67%。

关键词: 红茶等级; 图谱特征融合; 主成分分析; 支持向量机; 高光谱成像

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-1298(2023)03-0402-09

OSID:



Fast Nondestructive Discrimination of Yingde Black Tea Grade Based on Fusion of Image Spectral Features of Hyperspectral Technique

LIU Cuiling¹ QIN Dong¹ LING Caijin^{2,3} SUN Xiaorong⁴ GAO Liyang^{2,3} ZAN Jiarui⁴

(1. Key Laboratory of Industry Industrial Internet and Big Data of China National Light Industry, Beijing Technology and Business University, Beijing 100048, China

2. Tea Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China

3. Guangdong Provincial Key Laboratory of Tea Innovation and Utilization, Guangzhou 510640, China

4. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China)

Abstract: Tea grade evaluation is an important technical index to detect the quality of tea leaves. By extracting image features and spectral features under hyperspectral imaging technique of black tea, a fast and nondestructive discriminative model based on the map fusion method was constructed to be applicable to the grade evaluation of Yingde black tea. Firstly, three different grades of black tea samples were prepared, and the spectral data were visualized by dimensionality reduction using t -distributed stochastic neighbor embedding and principal component analysis, and then the characteristic wavelengths of each chemical value were extracted from the perspective of influencing the intrinsic quality by successive projections algorithm, followed by the best combination of characteristic wavelengths characterizing its intrinsic quality by multi-model consensus strategy and competitive adaptive reweighted sampling - successive projections algorithm screening, followed by the establishment of a genetic algorithm optimization support vector machine based grade discrimination model, and the accuracy of its model was

收稿日期: 2022-06-24 修回日期: 2022-07-25

基金项目: 北京市自然科学基金项目(4222043)、广东省农业科学院院长基金项目(202032)、广州市科技计划项目(202002020079)、中国轻工业工业互联网与大数据重点实验室开放项目(IIBD-2021-KF09)和研究生科研能力提升计划项目

作者简介: 刘翠玲(1963—),女,教授,博士,主要从事智能建模控制和食品安全检测等研究,E-mail: liucl@btbu.edu.cn

通信作者: 凌彩金(1977—),女,研究员,主要从事茶叶品质生化研究,E-mail: lingcailin@gdaas.cn

88% for the training set and 78.33% for the prediction set. In order to fuse the shape and texture differences, the hyperspectral image corresponding to the best feature wavelength combination were firstly extracted; and then the image mask was used to eliminate the interference of the background and the principal component analysis was used to eliminate the redundant information between multi-wavelength images, and then the gray level covariance matrix and local binary pattern algorithms were used to extract the three-dimensional principal component images before principal component analysis and fuse them with feature spectra, moreover, the genetic algorithm optimized support vector machine grade discrimination model based on feature fusion was established, and the best discrimination effect was based on the third principal component image feature fusion model, which the accuracy of the training set was improved to be 98% and the accuracy of the prediction set was improved to be 96.67%.

Key words: black tea grade; fusion of image spectral features; principal component analysis; support vector machine; hyperspectral imaging

0 引言

英德红茶产于广东省英德市,属于全发酵茶。不同等级的红茶,不仅口感、外形色泽差别比较大,价格往往也相差甚远^[1],对其进行科学快速的等级判别有一定的价值和意义。

目前茶叶品质评价技术主要包括传统感官审评^[2]、化学成分分析、电子舌技术、电子鼻技术等,其中感官评审法是唯一评定茶叶等级的国家标准,根据口感滋味(35%)、香气(30%)、外形纹理(20%)、汤色(10%)、叶底(5%)共5种品质得分比重综合评审,其缺点是易受外部环境和主观因素的影响,缺乏足够的客观性,因此茶叶感官审评无法完全做到标准化和真正客观化。化学成分分析手段例如顶空固相微萃取^[3-4]、气相色谱-质谱联用法^[5-6]等,操作繁琐,前处理步骤复杂,测定时间长,效率低。电子舌、电子鼻^[7-8]等属于智能体感设备,是对香气、滋味检测的数据化设备,成本大,而且受传感器的灵敏度和标准样影响较大,不利于广泛推广。目前大多数传统检测方法有其局限性,检测效率较慢且多为有损检测。

近年来,高光谱成像技术在农产品检测方面得到广泛的应用,其具有获取信息量丰富、检测快速高效、能获取整个样本的空间特性和光谱特性等优点。其中光谱信息能够反映茶叶的内部属性,图像信息可以表征茶叶的大小、形状和纹理等外观特征。到目前为止,已有学者将高光谱成像技术应用在茶叶的等级识别领域。目前基于高光谱成像技术的茶叶等级判别模型大多数都是基于光谱特征的直接判别模型,鲜有从内在品质参数角度提取内在光谱特征,同时研究中少有融合图像信息表征其外在品质,总体上研究思路相对于感官评审法有一定的偏差。

综上所述,应用高光谱成像技术进行茶叶等级判别具有一定的理论基础和可行性,本文采用高光

谱成像技术,以英德红茶为研究对象,采集不同等级茶叶的高光谱信息和图像信息,从感官评审角度出发,以分析内在品质特征为主,结合化学品质参数筛选对应的最佳特征波长组合,并提取外在品质;通过多种图谱特征提取融合方法建模并对模型进行相关的比较和分析,以期红茶品质的等级判别提供新的分析思路与应用实例。

1 材料与方法

1.1 实验材料

英德红茶样本 210 个,其中英红九号 170 个,金萱 40 个。样本由广东省农业科学院茶叶研究所提供;3 种等级各 70 个,品质等级由专业品茶师按 GB/T 23776—2018《茶叶感官审评方法》划分,对应的各等级红茶样本及外形品质标准如图 1 所示,3 种等级内在品质判别标准如表 1 所示。



图 1 3 种等级红茶

Fig. 1 Three grades of black tea

每个样本有 10 种化学值,均由广东省农业科学院茶叶研究所提供,化学值分别为茶多酚、可溶性糖、EGC(表没食子儿茶素)、C(儿茶素单体)、EGCG(表没食子儿茶素没食子酸酯)、ECG(表儿茶素没食子酸酯)、GCG(没食子儿茶素没食子酸酯)、总儿茶素、酯型儿茶素、简单儿茶素含量;茶多酚、儿茶素类及其酯类化合物含量按 GB/T 8313—2018《茶叶中茶多酚和儿茶素类含量的检测方法》测得,可溶性糖含量采用高效液相色谱法按 GB 29704—

表 1 红茶各等级感官评价标准

Tab.1 Sensory evaluation criteria for each grade of black tea

茶叶等级	汤色	香气	滋味	叶底
一级	清澈明亮	高爽或高	醇厚鲜爽	嫩匀多芽
		鲜、纯正、有嫩茶香		尖,明亮,匀齐
二级	较明亮	较高爽、较高鲜	尚醇厚,尚鲜爽	嫩尚匀,尚明亮,尚匀齐
三级	欠明亮或有浑浊	尚纯,熟、老火或青气	尚醇,浓涩,青涩	尚嫩,尚亮,欠匀齐

2013《超高效液相色谱-串联质谱法》测得。

1.2 仪器与设备

芬兰 SPECIM 型高光谱成像光谱仪,其仪器参数设置:波长范围为 800 ~ 2 500 nm,波段数为 288 个,曝光时间为 3.38 μm,帧频为 40 Hz,空间采样分辨率为 640 像素。

1.3 高光谱信息采集

采用 SPECIM 型高光谱相机采集茶叶样本的高光谱图像。为了降噪及修正图像,需要对原始高光谱图像进行黑白板校正。校正公式^[9]为

$$R' = \frac{S - D}{W - D}$$

(1)

式中 R' ——校正后的光谱数据

S ——采集到的暗场原始光谱数据

W ——拍摄光源得到的白板光谱数据

D ——关闭镜头后的全黑光谱数据

通过 ENVI 3.5 软件划定同一位置大小的感兴趣区域 (Region of interest, ROI),提取 ROI 区域所有像素的光谱数据,并计算其平均值,作为该样本的光谱数据。

1.4 数据预处理

采用合适的光谱预处理方法可以提取有效的光谱信息,提升光谱质量^[10]。由于实验样本是干茶,在采集过程中,会有杂质和表面的漫反射影响,因此采取标准正态变换 (SNV),通过对光谱数据的处理,消除固体颗粒大小和光程变化对物体表面漫反射的影响。

1.5 特征波长提取

竞争性自适应重加权算法 (Competitive adaptive reweighted sampling, CARS)^[11]变量筛选的基本思想是根据自适应重加权对偏最小二乘回归系数的筛选寻优,从而权重分配较大的波长被选择出来,分配的权重较小的波长点被筛选除去,最终通过交互验证方法选出交叉验证均方根误差 (Root mean squared error, RMSECV) 最小的数据集。通过上述筛选原

理,可选出与标签值相关性最高的最优波长组合。

连续投影算法 (Successive projections algorithm, SPA)^[12]利用向量的投影分析,在光谱信息中充分寻找含有最低限度冗余信息的变量组,将变量间共线性的影响降到最低,从而减少信息的重叠,同时筛选出的几个变量就能代表原始数据的大部分信息,提高了建模的速度和效率。

1.6 数据可视化分析

1.6.1 t 分布-随机近邻嵌入

t 分布-随机近邻嵌入 (t -distributed stochastic neighbor embedding, T-SNE)^[13-14]算法对每个数据近邻分布进行建模,在原始高维空间中,将高维空间建模为正态分布,而在二维输出空间中,可以将其建模为 t 分布。目标是找到将高维空间映射到二维空间的变换,并且最小化所有点在这两个分布之间的差距。T-SNE 可以在保留主要信息情况下有效区分不同的数据类簇,获得更好的可视化效果。

1.6.2 主成分分析

主成分分析 (Principal component analysis, PCA)^[15-16]可以用于高维数据降维和图像矩阵降维。数据 PCA 降维则是为了简化计算量以及存储空间,需要对这些高维数据进行一定程度的降维,将 n 维高维数据映射到低维平面,并尽量保证数据不失真;图像 PCA 分析则是当提取的图像特征维度比较高时,以降维后的矩阵为基础提取图像特征。

1.7 图像特征提取

1.7.1 灰度共生矩阵

灰度共生矩阵 (Gray level concurrence matrix, GLCM)^[17-18]通常以像素灰度和位置信息为参考量,用于分析灰度图像纹理特性,分析图像空间中任意一个点与其所在的相邻区域内一定间隔的像元灰度组合情况。GLCM 的定义表达为从灰度级 i 的像元 (m, n) 离开某个固定点距离 d 到达灰度为 $j(m + a, n + b)$ 的像元的概率 $p(i, j, d, \theta)$,所有估计的值均可以表示成一个矩阵。其表达式为

$$p(i, j, d, \theta) =$$

$$\begin{bmatrix} (m, n) & (m + a, n + b) & f(m, n) & f(m + a, n + b) \end{bmatrix}$$

(2)

式中 θ ——灰度共生矩阵生成的方向,通常取 4 个方向为 0°、45°、90°和 135°

$f(\cdot)$ ——灰度计算函数

需要对共生矩阵提取二次统计量,选择了 4 个纹理特征的参数,分别为对比度、能量值、熵和相关性指数,其物理含义如下:

对比度 f_1 主要是用于反映灰度共生矩阵的各个像素点的概率分布情况,图像目标的局部发生变

化的程度以及图像的清晰程度和图像纹理的浅深。计算公式为

$$f_1 = \sum_i \sum_j (i-j)^2 p(i,j) \tag{3}$$

能量值 f_2 用来描述图像纹理分布的均匀程度。计算公式为

$$f_2 = \sum_i \sum_j p^2(i,j) \tag{4}$$

熵 f_3 用于描述图像大小。计算公式为

$$f_3 = - \sum_i \sum_j p(i,j) \lg p(i,j) \tag{5}$$

相关性指数 f_4 用来描述共生矩阵在某一行或者某一列方向上的相似程度,作为矩阵灰度线性关系的一项衡量值。计算公式为

$$f_4 = \frac{\sum_i \sum_j \left[(i-j)^2 p(i,j) - \mu_x \mu_y \right]}{\sigma_x \sigma_y} \tag{6}$$

式中 μ_x, σ_x ——相应行灰度均值及标准差
 μ_y, σ_y ——相应列灰度均值及标准差

1.7.2 局部二值模式

局部二值模式 (Local binary pattern, LBP)^[19-20] 是一种描述图像纹理特征的算子,被广泛用于模式识别和图像处理领域,其基本思想是利用像元与邻近像元的关系描述局部纹理特征,3×3 窗口原理如图 2 所示,比较窗口中心像元值与邻域像元值的大小关系,若邻域像元值大于等于中心像元值则该像素点被记为 1,否则被记为 0。将产生的 8 位二进制数转换为十进制数并记作中心单元的 LBP 值。

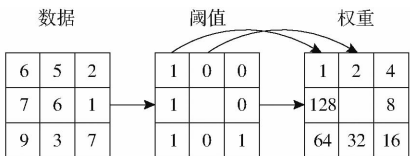


图 2 3×3 窗口原理
Fig. 2 Window principle

1.8 遗传算法 (GA) 寻优支持向量机 (SVM) 分类算法

1.8.1 支持向量机分类原理

支持向量机 (Support vector machine, SVM)^[21] 通过核函数定义映射关系,将待分类数据映射到更高维特征的空间中,来实现简单的线性可分性。然后在这个空间中构造最优的分类面,从而形成分类的决策规则。核函数、相应的核函数参数 g 和惩罚因子 c ,是支持向量机分类的 3 个可优化选项。高斯径向基核函数只需要设定一个参数,对模型的精度和稳定性更容易控制。因此,选择高斯径向基核函数;鉴于核函数参数 g 和惩罚因子 c 会对计算收敛速度和预测精度产生影响,故采用遗传算法对核函数参数 g 和惩罚因子 c 进行优化。

1.8.2 遗传算法优化支持向量机

遗传算法 (Genetic algorithm, GA)^[22] 是通过模拟达尔文生物进化论自然选择和遗传学机理来进行搜索最优解的一种方法。设置初代种群个数,按照“物竞天择,适者生存,优胜劣汰”的原理,来逐代演化出最优解,优化支持向量机的流程如图 3 所示。

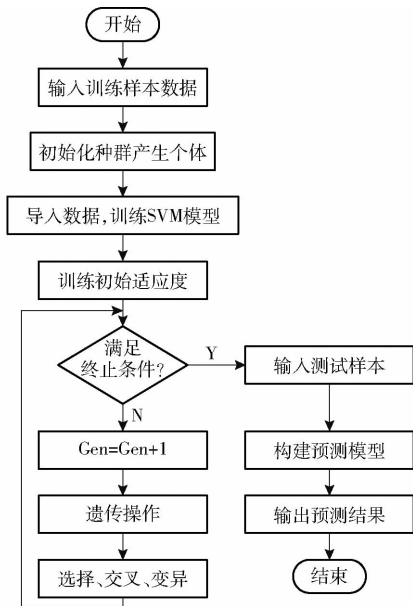


图 3 遗传算法优化支持向量机流程图
Fig. 3 Genetic algorithm optimization support vector machine flow chart

1.9 多模型共识策略建模分析方法

多模型共识判别策略^[23] 基本原则就是统计分析多个化学值对应的特征波长子模型的分析结果,综合判别茶叶的等级;假设子模型个数为 k ,采用 k 个子模型预测同一样本可以得到 k 个识别结果,当识别结果中识别为真的百分比大于 50%,则判定样本为真,反之则为假。设多模型共识策略的分析过程共用到 K 个化学值特征波长的光谱信息,则各化学值的特征波长分别建立 SVM 模型后就会得到 K 个品质等级判别结果,统计每个类别结果出现的概率,依照投票准则取概率最高的类别作为样本的分类结果。

2 结果与分析

2.1 基于 T-SNE 和 PCA 聚类的可视化分析

采集 210 个红茶样本的高光谱信息,化学含量以及物理特征的不同,样本对特定的波长有着不同的吸光度,原始光谱如图 4a 所示;但是当茶叶样品为固体,且表面不均匀,光散射会导致加性干扰和乘性干扰,所以采用 SNV 进行预处理,SNV 预处理后的光谱如图 4b 所示,可知经过预处理后光谱信息更加集中;采集的高光谱图像如图 4c 所示。

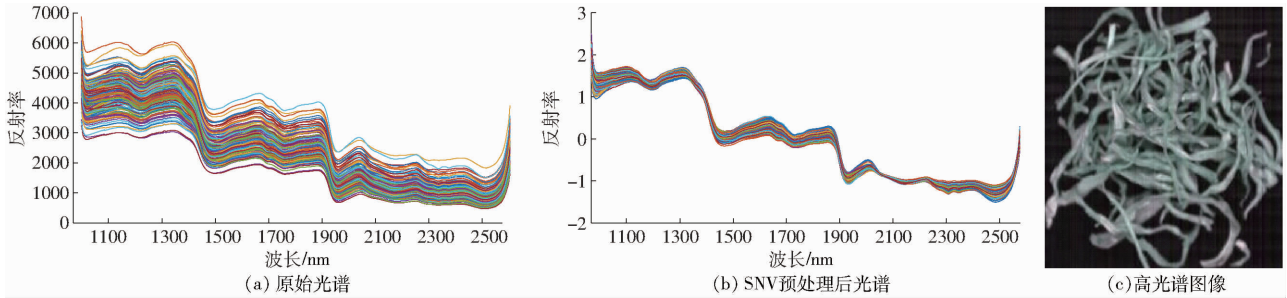


图 4 红茶高光谱信息

Fig. 4 Original spectrogram

采用 T-SNE 和 PCA 两种算法对 SNV 预处理后的 288 维高维光谱数据进行数据降维并进行可视化分析,采用 T-SNE 算法将数据降维到 2 维平面 (x,y),其聚类可视化如图 5 所示;为了方便对比分

析,采用 PCA 算法将数据映射到前 2 维特征上 (PC1、PC2),累积贡献率达 93.35%,其聚类可视化如图 6 所示。

对比图 5 和图 6 可知,T-SNE 较 PCA 的聚类

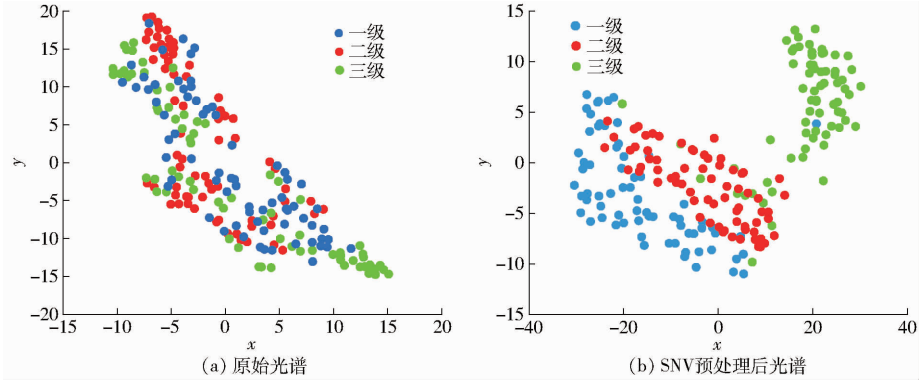


图 5 T-SNE 可视化结果

Fig. 5 T-SNE visualization

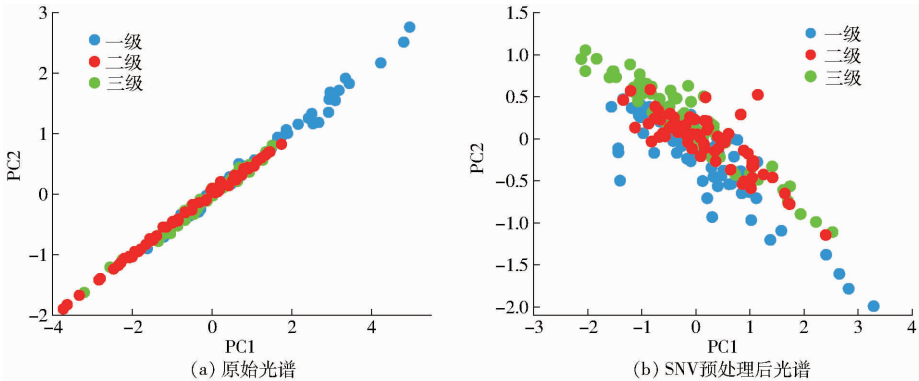


图 6 PCA 可视化结果

Fig. 6 PCA visualization

可视化效果较好,可以更明显地看出 3 种等级茶叶的光谱数据聚类分布情况;且从图 5a 和图 6a 对比可知,原始光谱存在小样本的数据簇聚集,但是 3 种等级茶叶存在较多的交叉混合,原始高光谱数据可分性不明显,所以需要对其进行预处理;从图 5b 和图 6b 对比可知,SNV 预处理后光谱数据有着各等级的数据簇,其中一级样本和三级样本数据簇的可分性比较明显,二级红茶样本与一级红茶样本和三级红茶样本存在部分交叉样本,这属于感官评审的主观性导致在相邻等级的交叉混合。因此经过

SNV 预处理后的光谱数据对于等级的分类有可分性,但对于等级交叉的样本纯光谱数据判别会存在一定的误判率。

建立基于 SNV 预处理后光谱数据和降维数据的 GA-SVM 模型,参数初始化,默认最大进化代数 100,种群最大数量为 20,交叉概率为 0.4,变异概率为 0.01,默认 SVM 模型参数 c 和 g 为 3,进行有限次迭代寻优,得出最佳的适应度。采用 Kennard-Stone(KS)法按 5:2 划分数据集,训练集 150 个,预测集 60 个,其模型分类结果如表 2 所示。

表 2 GA-SVM 建模结果

Tab.2 Modeling results of GA-SVM

数据集	最佳惩罚	最佳核函	训练集准	预测集准
	因子	数参数	确率/%	确率/%
原始光谱数据	5.49	9.89	65.33	58.33
SNV 预处理	6.17	6.39	75.33	71.66
T-SNE 降维	5.94	6.27	78.00	76.67
PCA 降维	6.84	5.73	75.33	73.33

从表 2 看出,基于全波长数据和 SNV 预处理后的光谱数据建模判别准确率较低,采用 T-SNE 和 PCA 降维后的数据建模的模型性能有略微提升,且 T-SNE 降维的数据建模效果较好;由于降维后数据维数过低会丢失部分信息导致建模效果不佳,但通过分析其二维平面聚类可视化图可得:基于光谱数据建模判别不同等级的红茶是可行的;由于全波长存在部分冗余波长之间的共性干扰,所以需要筛选特征波长来提高判别准确率。

2.2 化学成分与茶叶等级的相关性分析

计算 10 种化学值与红茶等级的相关性系数,由其相关性可知,总儿茶素、酯型儿茶素、简单儿茶素、茶多酚、EGCG、ECG 含量分别与红茶等级的相关系数在 0.7 以上,GCG、EGC、C、可溶性糖含量分别与

红茶等级的相关系数在 0.5~0.7 之间。因为茶多酚是决定茶叶的色、香、味及功效的主要成分,其中以儿茶素含量最高,质量分数占茶多酚的 60%~80%;EGCG、ECG、EGC、EC 是儿茶素的主要组成成分且是茶叶质量控制的关键成分,且 EGCG 含量最多,质量分数占儿茶素 50% 以上;其中酯型儿茶素主要为 EGCG、ECG、GCG 等的混合物,具有较强的苦涩味和收敛性,是茶汤的主体呈味物质;简单儿茶素主要为 C、EC、EGC 等的混合物,具有收敛性弱,使得茶汤鲜爽不苦涩的作用,总儿茶素为简单儿茶素和酯型儿茶素的总称,其包含的化学值基本涵盖了影响茶叶内在品质的主要化学成分。所以可以得出总儿茶素、酯型儿茶素、简单儿茶素、茶多酚、EGCG、ECG 含量这 6 种化学品质参数的化学特性与茶叶品质有着比较直接的关系,所以其组合可以表征红茶大部分的内在品质。

2.3 基于多模型共识策略的特征波长筛选

由于 6 种化学成分与红茶等级有较强的相关性,所以其对应的特征波长能表征其化学值特性,基于 6 种化学值的 SPA 特征波长提取结果如表 3 所示。

表 3 6 种化学值对应的 SPA 特征波长

Tab.3 Six chemical values corresponded to SPA characteristic wavelengths

化学值	特征波长数	最小均方根误差	对应的特征波长/nm
总儿茶素含量	7	7.47	990.72,1137.80,1368.75,1475.43,1665.88,1710.612,1855.87
酯型儿茶素含量	8	4.84	1137.80,1301.27,1660.18,1721.80,1800.03,1855.87,1872.61,1922.84
简单儿茶素含量	7	17.65	1132.16,1301.27,1509.08,1665.88,1716.21,1867.03,1939.58
茶多酚含量	7	4.01	1211.15,1278.76,1402.46,1593.12,1671.47,1939.58,2045.54
EGCG 含量	7	24.06	1092.60,1211.15,1357.51,1469.82,1621.11,1721.80,2039.96
ECG 含量	6	7.69	1089.26,1469.82,1649.09,1827.95,2028.81,2145.88

首先将 6 种化学值的特征波长分别作为输入单独建立 GA-SVM 红茶等级判别模型。采用 KS 法按 5:2 划分数数据集,训练集样本 150 个,3 种等级茶叶各 50 个,预测集样本 60 个,3 种等级茶叶各 20 个;6 种化学值对应特征波长的建模结果如表 4 所示。

表 4 单化学值特征波长建模结果

Tab.4 Single chemical characteristic wavelength modeling results

化学值	最佳惩罚	最佳核函	训练集准	预测集准
	因子	数参数	确率/%	确率/%
总儿茶素含量	6.17	6.38	73.33	66.67
酯型儿茶素含量	5.78	6.30	66.00	63.33
简单儿茶素含量	5.74	7.49	68.60	60.00
茶多酚含量	4.85	5.85	62.67	56.67
EGCG 含量	6.01	6.70	61.33	58.33
ECG 含量	5.26	6.60	59.33	53.33

为了保证对同一个样本综合判别结果出现真假的概率不同,确定联合子模型的个数为 4、5、6,首先根据表 4 中的单化学值特征波长子模型识别准确率从高到低排序,依次联合选取前三、前四、前五识别准确率的化学子模型作为固定子模型,依次选取第 4 维子模型、第 5 维子模型、第 6 维子模型的特征波长进行联合输入,建立模型集群进行判别;其联合模型的判别结果如表 5 所示。

从表 5 可知,当输入为 5 种化学值子模型,即:总儿茶素、酯型儿茶素、简单儿茶素、茶多酚、EGCG 子模型联合时,对红茶等级判别准确率最高,训练集的准确率为 86.00%,预测集的准确率为 78.33%,其组合波长共 35 个,占总波长数百分比为 12.15%。相对于单化学值模型组合波长的信息更丰富,但是同样存在相应的冗余波长,所以再采用

表 5 联合模型的判别结果

Tab.5 Discriminant results of combined model

固定子模型	联合子模型 维数	联合子模型	最佳惩罚 因子	最佳核函数 参数	训练集准确 率/%	预测集准确 率/%
总儿茶素、酯型儿茶素、简单儿茶素	4	茶多酚	4.58	6.86	78.00	70.00
		EGCG	4.09	6.47	72.00	66.67
		ECG	5.55	6.25	65.33	60.00
总儿茶素、酯型儿茶素、简单儿茶素、茶多酚	5	EGCG	5.13	6.28	86.00	78.33
		ECG	4.84	5.63	82.00	76.67
总儿茶素、酯型儿茶素、简单儿茶素、茶多酚、EGCG	6	ECG	5.13	6.29	81.33	73.33

CARS 算法进行二次特征波长筛选。CARS 波长筛选过程如图 7 所示,CARS 特征波长二次筛选建模的预测集判别结果如图 8 所示。

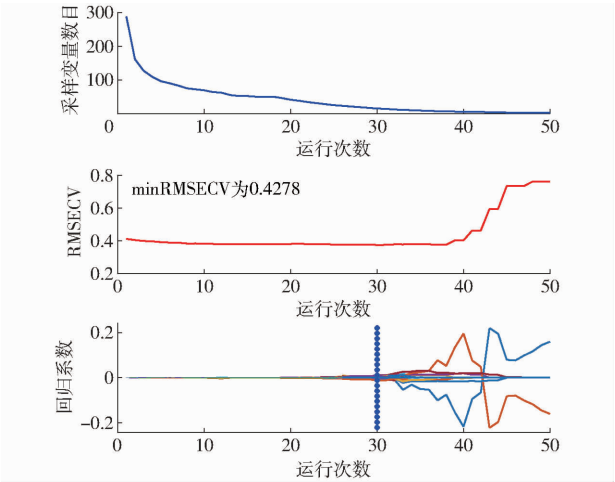


图 7 CARS 特征波长二次提取

Fig.7 CARS feature wavelength secondary extraction

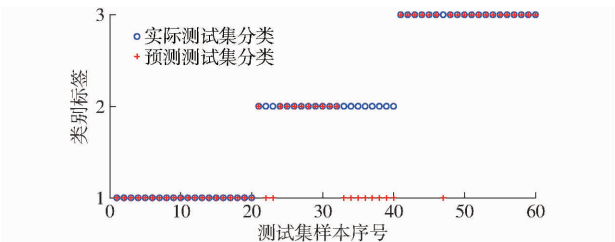


图 8 CARS 特征波长二次筛选模型的预测集判别结果

Fig.8 Prediction set discrimination results of CARS feature wavelength quadratic screening model

从图 7 可知,当采样次数在 30 次时,其最小均方根误差 (Minimum root mean squared error, minRMSE) 为 0.4278;通过 CARS 二次筛选的波长数为 8 个,特征波长比例下降近 10%。由图 8 可知,其特征波长建模将训练集的判别准确率提高至 88%,预测集的判别准确率提高至 81.67%,剔除了一部分冗余波长,使得特征波长信息更能准确判别各等级红茶。得出最佳的特征波长组合为 1 137.80、1 278.76、1 301.27、1 368.75、1 469.82、1 593.12、1 671.47、1 855.87 nm。其中 1 100 ~ 1 300 nm 附近的光谱反射率主要与茶多酚有关,

1 300 ~ 1 900 nm 的光谱反射率主要是由 EGCG、ECG、EC 等化学成分的 C=H、O=H、N=H 官能团所产生的倍频和合频吸收而形成的,所以二次筛选的特征波长更能够精准地表征内在品质特征。

2.4 基于图谱特征融合技术的建模分析

茶叶等级品质既包括内在品质,也包含外在特征。图像特征差异也是其品质表现之一,特征波长对应的高光谱图像如图 9 所示。

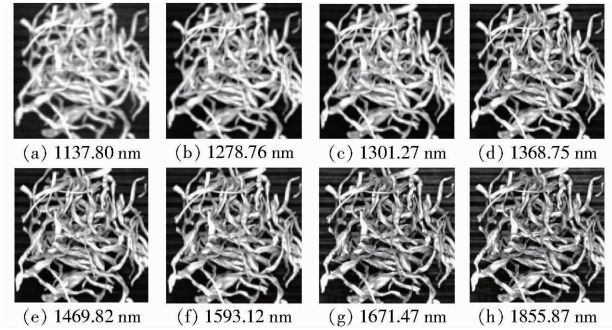


图 9 特征波长对应的高光谱图像

Fig.9 Hyperspectral image corresponded to characteristic wavelength

从图 9 可知,在不同波长下图像可以体现出样本的对比度、纹理等外在特征;从图 1 可知,不同等级茶叶的外形特征也是有区别的。图像的背景对茶叶的特征提取存在一定的影响且多波段的高光谱图像中也存在冗余信息。图像掩膜可以遮盖大部分背景的影响,PCA 变换可以将多波段图像中的有用信息集中到数量尽可能少的新的主成分图像中,并使这些主成分图像之间互不相关,从而大大减少总的数据量。建立圆形图像掩膜来遮盖大部分背景,突出茶叶为图像处理中心;采用 PCA 方法进行图像主成分分析,例:对波长 1 855.87 nm 处的图像进行 PCA 分解得到的第一、第二、第三主成分图像如图 10 所示,前 3 维累计贡献率达到 95% 以上。

从图 10 可以看出,第三主成分图像比第一主成分、第二主成分图像能够更好地区分背景与茶叶样本,不仅弱化了放置茶叶样本的数粒板背景,而且突

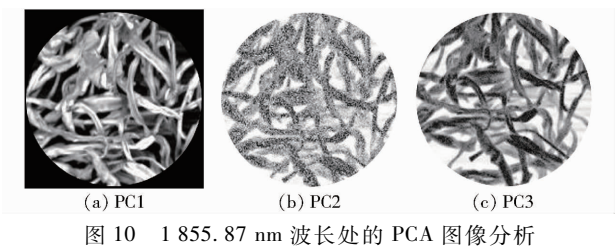


图 10 1 855.87 nm 波长处的 PCA 图像分析
Fig. 10 PCA image analysis at 1 855.87 nm

出了茶叶样本的图像特征,这是由于 PCA 变换对噪声比较敏感,本文中茶叶样本是信号,数粒板背景是噪声,第一、二主成分虽然信息含量比较高,但是此时噪声明显对信号有影响,因此第一、二主成分分量形成的图像质量不如第三主成分图像。

分别采用 GLCM、LBP 两种算法对特征波长图像进行图像特征提取,采用 GLCM 提取 4 个方向为 0°、45°、90°和 135°对应的对比度、能量值、熵和相关性指数共 16 个特征;采用 LBP 提取到每幅图像共 256 个特征值,LBP 二值化的纹理图像如图 11 所示,PCA 图像分析后的 LBP 二值图如图 12 所示。将每个样本的 8 个特征图像特征的平均值作为该样本的图像特征,将图像特征与特征波长串行融合得到一组新的特征向量。采用 KS 法按 5:2 划分数据集,训练集 150 个,预测集 60 个,建立基于图谱特征融合的 GA-SVM 等级判别模型。其未进行 PCA 图像分析的判别结果如表 6 所示,PCA 图像分析的判别结果如表 7 所示,最佳预测集判别结果如图 13 所示。



图 11 未 PCA 图像分析 LBP 纹理图像
Fig. 11 LBP texture images without PCA image analysis

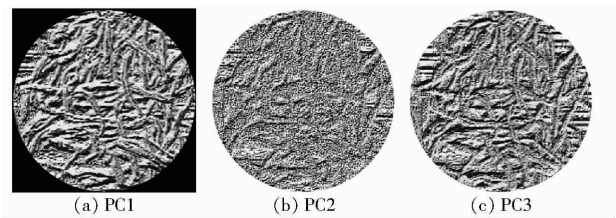


图 12 PCA 图像分析后的 LBP 纹理图像
Fig. 12 Image analysis of LBP texture image after PCA

对比表 6 和表 7 可知,PCA 图像分析的图谱融合模型的判别准确率相对于未 PCA 图像分析时整体明显提高,说明 PCA 分析后的图像更能凸显茶叶

表 6 未进行 PCA 图像分析的判别结果

Tab. 6 Discriminant results without PCA image analysis

图像特征 算法	最佳惩 罚因子	最佳核函 数参数	训练集准确率/ %	预测集准确 率/%
GLCM	5.68	4.03	91.33 (135/150)	88.33 (53/60)
LBP	6.17	4.55	84.67 (127/150)	80.00 (48/60)

表 7 PCA 特征提取的图谱融合模型判别结果

Tab. 7 Discriminant results of atlas fusion model
for PCA feature extraction

主成分	图像特征 算法	最佳惩 罚因子	最佳核函 数参数	训练集准 确率/%	预测集准 确率/%
PC1	GLCM	5.48	6.17	89.33	86.67
PC2		4.83	4.73	94.67	90.00
PC3		4.07	1.18	98.00	96.67
PC1	LBP	5.02	6.17	88.67	86.67
PC2		4.97	6.54	84.00	81.66
PC3		6.00	8.79	86.00	80.00

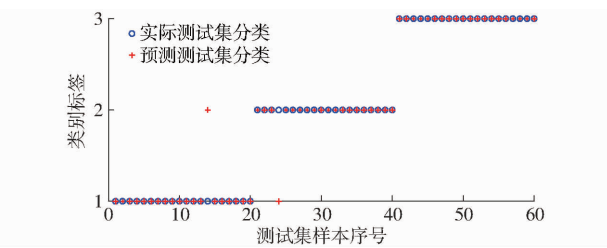


图 13 最佳预测集判别结果
Fig. 13 Best prediction set discriminant results

的图像特征,其进行图谱融合后的模型判别性能更好;另外由表 7 可以看出,基于 PC3-GLCM 图谱特征融合时,模型的判别性能最好,训练集的判别准确率达 98%,预测集的判别准确率达 96.67%,并且由图 13 可看出,对于三级红茶的判别准确率达 100%,一级和二级的判别准确率达 90%。因此,基于 PC3-GLCM 下的特征图谱融合模型可以更准确地判别不同等级的红茶。

3 结论

(1)从内在品质角度,采用多模型共识策略和 CARS-SPA 方法筛选最佳特征波长组合,建立基于最佳特征波长组合的 GA-SVM 等级判别模型,模型训练集判别准确率最高达 88%,预测集判别准确率最高达 81.67%。

(2)提取最佳波长下的高光谱图像,对其进行 PCA 图像分析,采用 GLCM 和 LBP 对 PCA 后的图像提取图像特征并与特征波长串行融合,建立基于图谱特征融合的 GA-SVM 等级判别模型,模型性能有了大幅度提高,PC3 成分下的特征融合模型判别效果最好,训练集准确率达 98%,预测集准确率达 96.67%。

参 考 文 献

- [1] REN G, LIU Y, NING J, et al. Assessing black tea quality based on visible-near infrared spectra and kernel-based methods[J]. *Journal of Food Composition and Analysis*, 2021, 98: 103810.
- [2] 中国国家标准化管理委员会, 中华人民共和国国家质量监督检验检疫总局. 茶叶感官审评方法: GB/T 23776—2018[S]. 北京: 中国标准出版社, 2018.
- [3] BHUYAN L P, BORAH P, SABHAPONDIT S, et al. Spatial variability of theaflavins and thearubigins fractions and their impact on black tea quality[J]. *Journal of Food Science and Technology*, 2015, 52(12): 7984–7993.
- [4] SOLTANI S, SERESHTI H. A green alternative QuEChERS developed based on green deep eutectic solvents coupled with gas chromatography-mass spectrometry for the analysis of pesticides in tea samples[J]. *Food Chemistry*, 2022, 380: 132181.
- [5] WANG B, CHEN H, QU F, et al. Identification of aroma-active components in black teas produced by six Chinese tea cultivars in high-latitude region by GC-MS and GC-O analysis[J]. *European Food Research and Technology*, 2022, 248(3): 647–657.
- [6] REN G, LI T, WEI Y, et al. Estimation of Congou black tea quality by an electronic tongue technology combined with multivariate analysis[J]. *Microchemical Journal*, 2020, 163: 105899.
- [7] 张晓云, 赵艳, 钱晔, 等. 高压脉冲电场对普洱生茶香气和陈化时间的影响[J]. *食品科学*, 2020, 41(9): 43–49.
- ZHANG Xiaoyun, ZHAO Yan, QIAN Ye, et al. Effect of high voltage pulsed electric field on aroma and aging time of unfermented Pu'er tea[J]. *Food Science*, 2020, 41(9): 43–49. (in Chinese)
- [8] 徐敏. 基于电子鼻、电子舌和电子眼的多源信息融合技术对龙井茶品质的检测[D]. 杭州: 浙江大学, 2020.
- XU Min. Quality detection of Longjing tea based on the multi-source information fusion of E-nose, E-tongue and E-eye[D]. Hangzhou: Zhejiang University, 2020. (in Chinese)
- [9] 陈全胜, 赵杰文, 蔡健荣, 等. 利用高光谱图像技术评判茶叶的质量等级[J]. *光学学报*, 2008, 28(4): 669–674.
- CHEN Quansheng, ZHAO Jiwen, CAI Jianrong, et al. Estimation of tea quality level using hyperspectral imaging technology[J]. *Acta Optica Sinica*, 2008, 28(4): 669–674. (in Chinese)
- [10] 黄燕, 王璐, 关海鸥, 等. 基于优选 NIR 光谱波数的绿豆产地无损检测方法[J]. *光谱学与光谱分析*, 2021, 41(4): 1188–1193.
- HUANG Yan, WANG Lu, GUAN Haiou, et al. Nondestructive detection method of mung bean origin based on optimized NIR spectral wavenumber[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(4): 1188–1193. (in Chinese)
- [11] 袁莹, 王伟, 褚璇, 等. 光谱特征波长的 SPA 选取和基于 SVM 的玉米颗粒霉变程度定性判别[J]. *光谱学与光谱分析*, 2016, 36(1): 226–230.
- YUAN Ying, WANG Wei, CHU Xuan, et al. Selection of characteristic wavelengths using SPA and qualitative discrimination of mildew degree of corn kernels based on SVM[J]. *Spectroscopy and Spectral Analysis*, 2016, 36(1): 226–230. (in Chinese)
- [12] 郭阳, 郭俊先, 史勇, 等. CARS-SVM 预测哈密瓜可溶性固形物含量[J]. *食品与机械*, 2021, 37(6): 81–85.
- GUO Yang, GUO Junxian, SHI Yong, et al. Prediction of soluble solids in Hami melon by CARS-SVM[J]. *Food & Machinery*, 2021, 37(6): 81–85. (in Chinese)
- [13] YI C, TUO S, TU S, et al. Improved fuzzy C-means clustering algorithm based on t-SNE for terahertz spectral recognition[J]. *Infrared Physics & Technology*, 2021, 117: 103856.
- [14] 陈俊佟. t-SNE 结合支持向量机的降维分类研究[D]. 大连: 大连理工大学, 2021.
- CHEN Juntong. Research on dimensionality reduction classification of t-SNE combined with support vector machine[D]. Dalian: Dalian University of Technology, 2021. (in Chinese)
- [15] AGRAHARI A, GHOSH R. Multi-oriented text detection in natural scene images based on the intersection of MSER with the locally binarized image[J]. *Procedia Computer Science*, 2020, 171: 322–330.
- [16] 张颖, 马承泽, 杨平, 等. 基于小波变换和改进 PCA 的人脸特征提取算法[J]. *吉林大学学报(理学版)*, 2021, 59(6): 1499–1503.
- ZHANG Ying, MA Chengze, YANG Ping, et al. Face feature extraction algorithm based on wavelet transform and improved principal component analysis[J]. *Journal of Jilin University(Science Edition)*, 2021, 59(6): 1499–1503. (in Chinese)
- [17] UTAMININGRUM F, SAROSA S J A, KARIM C, et al. The combination of gray level co-occurrence matrix and back propagation neural network for classifying stairs descent and floor[J]. *ICT Express*, 2022, 8(1): 151–160.
- [18] LALIT R, PURWAR R K. Crowd abnormality detection using optical flow and GLCM-based texture features[J]. *Journal of Information Technology Research (JITR)*, 2022, 15(1): 1–15.
- [19] LIU X, LI L, WANG S, et al. Local-binarized very deep residual network for visual categorization[J]. *Neurocomputing*, 2021, 430: 82–93.
- [20] TAN Q, LI W, CHEN X. Identification the source of fecal contamination for geographically unassociated samples with a statistical classification model based on support vector machine[J]. *Journal of Hazardous Materials*, 2021, 407(14): 124821.
- [21] ZHOU J, HUANG S, WANG M, et al. Performance evaluation of hybrid GA-SVM and GWO-SVM models to predict earthquake-induced liquefaction potential of soil: a multi-dataset investigation[J]. *Engineering with Computers*, 2021(2): 1–19.
- [22] GUAN S, WANG X, HUA L, et al. Quantitative ultrasonic testing for near-surface defects of large ring forgings using feature extraction and GA-SVM[J]. *Applied Acoustics*, 2021, 173: 107714.
- [23] 张征, 马晓军, 张运银, 等. 一种基于多模型融合的电驱车辆车速估计方法[J]. *火力与指挥控制*, 2020, 45(12): 160–164.
- ZHANG Zheng, MA Xiaojun, ZHANG Yunyin, et al. A speed estimation method for electric drive vehicles based on multi-model fusion[J]. *Fire Control & Command Control*, 2020, 45(12): 160–164. (in Chinese)