

陈书媛, 张友超, 杨杰, 等. 基于高光谱成像技术的白茶储藏年份判别 [J]. 食品工业科技, 2021, 42(18): 276–283. doi: 10.13386/j.issn1002-0306.2020110299

CHEN Shuyuan, ZHANG Youchao, YANG Jie, et al. Discrimination of Storage Time of White Tea Using Hyperspectral Imaging[J]. Science and Technology of Food Industry, 2021, 42(18): 276–283. (in Chinese with English abstract). doi: 10.13386/j.issn1002-0306.2020110299

· 分析检测 ·

基于高光谱成像技术的白茶储藏年份判别

陈书媛¹, 张友超², 杨杰², 蔡梅生³, 张青碧⁴, 何普明^{1,*}, 屠幼英¹

(1. 浙江大学茶学系, 浙江杭州 310058;

2. 浙江大学农业工程系, 浙江杭州 310058;

3. 福鼎市茶业发展领导小组, 福建福鼎 355200;

4. 福鼎市茶业管理局, 福建福鼎 355200)

摘要: 储藏年份是决定白茶经济价值的一大因素。为了实现快速便捷地判别白茶储藏年份, 本文提出了基于高光谱成像技术判别分析白茶储藏年份的无损检测方法。通过对 3、6、10 年寿眉高光谱图像感兴趣区域光谱数据的提取, 采用最小二乘平滑滤波、标准正态变换、归一化、多元散射校正预处理算法, 并用支持向量机、偏最小二乘联合线性判定法、逻辑回归建模对不同预处理后的光谱数据进行判别分析。最后, 通过分析混淆矩阵、精确率、召回率来评估模型性能。分析结果表明, 经过标准正态变换预处理结合支持向量机所建立的模型判别效果最佳, 训练集和测试集的精确率分别为 90.83% 和 86.02%。由此可见, 利用高光谱成像技术对白茶储藏年份进行快速无损的判别具有一定的可行性。

关键词: 白茶, 高光谱成像, 判别分析, 储藏年份

中图分类号: TS272.7

文献标识码: A

文章编号: 1002-0306(2021)18-0276-08

DOI: 10.13386/j.issn1002-0306.2020110299



本文网刊:

Discrimination of Storage Time of White Tea Using Hyperspectral Imaging

CHEN Shuyuan¹, ZHANG Youchao², YANG Jie², CAI Meisheng³, ZHANG Qingbi⁴,
HE Puming^{1,*}, TU Youying¹

(1. Department of Tea Science, Zhejiang University, Hangzhou 310058, China;

2. Department of Agricultural Engineering, Zhejiang University, Hangzhou 310058, China;

3. Organization Group of Fuding Tea Industry Development, Fuding 355200, China;

4. Fuding Tea Industry Bureau, Fuding 355200, China)

Abstract: The storage time is a main factor determining the value of white tea. To discriminate the storage time rapidly and nondestructively, a new method was applied in this paper. First, hyperspectral image data were captured from Shoumei of 3, 6 and 10 years storage time. Second, four kinds of algorithms were applied to preprocess the original data, savitzky-golay Filter, standard normal variate, minmaxscaler, and multiplicative scatter correction. Third, support vector machine, partial least squares with linear discriminant analysis and logistic regression were built based on the data of the full spectra. Finally, the best combination of preprocessing algorithm and model could be found by comparing the confusion matrix, precision and recall. The results showed that the best classification performances were obtained with the support vector machine after the pretreatment of standard normal variate. The precision of the calibration sets was 90.83%, and that of the prediction sets was 86.02%. Therefore, it is possible to use hyperspectral imaging to discriminate white tea of different storage time in tea industry.

Key words: white tea; hyperspectral imaging; discriminant analysis; storage time

收稿日期: 2020-12-02

作者简介: 陈书媛 (1995-), 女, 硕士研究生, 研究方向: 茶叶品质检测, E-mail: 584533826@qq.com。

* 通信作者: 何普明 (1965-), 男, 博士, 教授, 研究方向: 茶叶生物化学, E-mail: pmhe@zju.edu.cn。

白茶是我国六大茶类之一, 主要产区在福建福鼎、政和、松溪、建阳、云南景谷等地。根据原料的嫩度不同, 白茶可分为白毫银针、白牡丹、寿眉和贡眉。萎凋和干燥这两道工序形成了白茶独特的品质特征和生物活性。研究表明, 白茶具有许多强大的生物活性, 包括抗氧化、抗炎、抗癌、抗菌和神经保护活性^[1-5]。而在茶叶销售市场中, 长期储存的白茶更受到广大消费者的喜爱和追捧, 有着“一年茶, 三年药, 七年宝”的说法。有研究表明^[6], 白茶在储藏过程中, 其化学品质成分会发生改变, 与新白茶相比, 储藏白茶的醇含量更低, 碳氢化合物含量更高。Ning 等^[7]发现儿茶素和氨基酸的含量随着储藏年份的延长呈相似的下降趋势, 而没食子酸的含量则增加。相比之下, 由茶氨酸和儿茶素在储藏过程中形成的 7 种新化合物 8-C-N-乙基-2-吡咯烷酮 (8-C N-ethyl-2-pyrrolidinone-substituted flavan-3-ols, EPSFs) 取代黄烷-3-醇显著增加, 其含量与储藏年份呈正相关^[8]。随着消费者对储藏白茶的喜爱的增加, 市面上也出现了储藏年份造假的现象, 但当下关于白茶储藏年份预测的研究报道极少。Dongchao Xie 等^[9]对白茶中 24 种储藏相关化合物进行了绝对定量分析和线性回归, 分析表明基于 EPSFs 的 5 个指标组合对白茶储藏年份具有较好的判别能力, 其中模型和测试集的相关系数分别为 0.9294 和 0.8812。该模型对于储藏年份小于 10 年的白茶, 预测和实际储藏年份之间的误差在-1.75~1.84 年之间, 尚不能满足快速精确判别白茶年份的需求。

光谱技术的发展为快速、准确地进行茶叶品控检测提供了基础, 近年来较多研究集中在近红外光谱技术在茶叶领域中的应用。刘鹏^[10]提出采用局部线性嵌入法和拉普拉斯特征映射法对近红外光谱数据进行非线性流形降维处理, 从而提高不同海拔茶叶品质的鉴别精度。Wang 等^[11]利用微型近红外光谱仪预测了红绿茶的判别模型, 基于标准正态变量的支持向量机模型对两类茶叶中的儿茶素和咖啡因都有较高的判别能力。刘洪林等^[12]对利用近红外光谱技术结合感官审评结果建立多个预测模型, 其预测性能优, 为客观评价功夫红品质提出新方法。李春霖^[13]采用化学计量方法和近红外光谱技术对龙井茶的感官和化学品质评价进行了系统性研究, 建立龙井茶鲜味和涩味的定性预测模型, 预测准确率达 75.18%。

相较于近红外光谱技术, 高光谱成像技术是一种更高效、无损的检测技术。近红外光谱技术对待检测的样本有一定的要求, 例如磨碎过筛, 而高光谱成像技术则不需要。并且高光谱成像技术是成像技术和光谱技术的结合体, 故可以同时获得待测样品的的光谱信息和空间信息, 以此检测样品的内外品质。目前高光谱成像技术在茶叶领域的研究主要集中在茶叶等级判别和茶类判别。于英杰等^[14]利用高光谱技术对不同等级的铁观音进行判别, 结合支持向量机的

模型对未知的铁观音茶样正确判别率可达 92.86%。李晓丽等^[15]结合高光谱成像仪和高效液相色谱法, 建立光谱与表没食子儿茶素没食子酸酯(epigallocatechin gallate, EGCG)浓度之间的回归模型, 其最优回归模型的确定系数达到 0.905。李瑶等^[16]以蒙顶黄芽、竹叶青、甘露茶叶为实验对象, 利用高光谱成像技术结合支持向量机建立茶叶品种判别模型和茶叶等级判别模型, 其精确率分别达到了 100% 和 96.67%。Guangxin Ren 等^[17]利用近红外高光谱成像结合多决策树的方法, 对祁门工夫红茶进行了品质和等级判别, 比较了三种不同类型的监督决策树算法, 其中基于数据融合的细树(fine tree, FT)模型预测效果最好, 红茶品质的评价正确率达到 93.13%。Zhiqi Hong 等^[18]采用 380~1030 和 874~1734 nm 两个光谱范围的高光谱成像系统对 6 个产地的龙井茶单叶进行了产地判别, 建立了支持向量机和偏最小二乘判别分析模型, 在两个光谱范围内均获得了良好的分类性能, 校正集和预测集的总体分类精确率均在 84% 以上。

鉴于白茶储藏年份判别的研究报道较少以及高光谱成像技术的优点, 本文探讨高光谱成像技术在白茶储藏年份判别中的应用, 在 450~998 nm 光谱范围采集白茶的高光谱图像并建模分析, 尝试为白茶储藏年份快速检测提供理论基础。

1 材料与方法

1.1 材料与仪器

实验样本 均来自福建省福鼎市(北纬 26°52'~27°26', 东经 119°55'~120°43')12 家茶叶企业。选取的白茶样本等级均为寿眉, 原料品种均为福鼎大毫。实验样本的储藏条件与白茶的一般储藏条件基本一致, 实验前样本均储藏在室温 25 °C, 通风干燥的茶叶储藏室中。在实验当月 2020 年 8 月, 将生产日期在 2017 年 5 月~10 月、2014 年 5 月~10 月、2010 年 5 月~10 月的样本储藏时间分别归为 3、6、10 年, 样本数量分别为 160、156、148, 每份样本约为 15 g。

Q285 成像光谱仪 来自德国品牌 Cubert; AZ-D100 卤素灯 来自德国品牌 Osram。为避免图像采集时环境光的干扰: 4 个 75 W 卤素灯对称安装在暗箱侧部, 保证整个视场光照均匀, 电控升降台安装在暗箱底部。成像光谱仪通过网线与计算机相连, 以控制数据的采集和传输^[19-20]。Q285 采用能在千分之一秒内获得样本高光谱图像立方体数据的画幅式成像技术, 相比于传统的推扫式成像技术, 不需要配备电动位移平台, 其数据采集过程更加快速、稳定。Q285 的主要技术参数如下: 光谱范围为 450~998 nm, 采样间隔 4 nm, 通道 125, 镜头焦距 23 mm, 像素分辨率 1000×1000, 数字分辨率 14 bit。

1.2 光谱采集

采集高光谱图像前的准备工作: 实验前 30 min 打开光源进行预热, 将寿眉样本均匀平铺在直径 150 mm、高度 20 mm 的培养皿中, 培养皿预先内置

反射率近似为零的黑色橡胶, 以免影响实验数据。

打开高光谱图像采集软件的操作界面, 采集寿眉样本的高光谱图像。使用 HSI Analyzer (ISUZU optics, 中国台湾) 软件对高光谱图像进行白板与暗电流校正。然后, 利用 ENVI 5.1 (Exelis VIS, 美国) 软件分析寿眉的光谱特征, 确定分割阈值, 将背景置 0, 并使用 Spectral Python 0.21, scikit-learn 0.23.1, numpy 1.18.1, pandas 1.0.3 做后续光谱数据分析。

1.3 数据处理

1.3.1 光谱提取和样本划分 由于寿眉样本等级基本一致, 且叶梗芽随机分布, 颜色花杂, 故高光谱采集到的图像数据对寿眉分类并无实际意义, 本文只分析光谱数据。高光谱图像在采集过程中会受到多种因素的干扰, 因此高光谱图像的去噪尤为重要^[21]。为了尽可能保留光谱的主要特征, 同时又能去除光谱中的噪音, 本文采用数学形态学中的二值形态学对采集到的寿眉样本图像进行降噪处理, 利用开运算和闭运算将图像上孤立的小点和毛刺去掉, 填平小孔, 并保持总体的位置和形状不变。

因为寿眉样本的不同部位, 如梗、叶、芽所含的内含物质成分差异明显, 故每个像素点之间的光谱数据差异较大。为解决这一问题, 如图 1 所示, 本文采用的方法是在样本区域内随机选取 100 个 20×20 像素的感兴趣区域 (Region of Interest, ROI), 其中一个像素点包含一条光谱信息, 再计算 ROI 内所有像素点的平均光谱作为建立模型的光谱数据进行下一步分析。

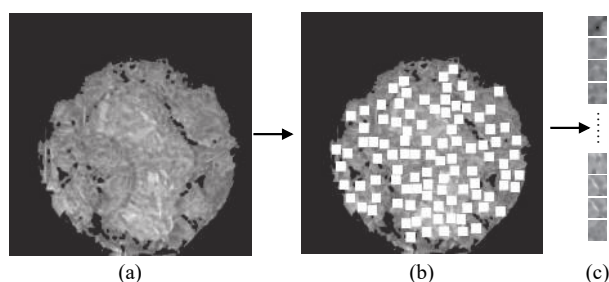


图 1 寿眉样本高光谱图像数据采集示意图

Fig.1 Sketch of data acquisition

注: (a) 高光谱图像去噪后的原图; (b) 在样本区域内随机选取 100 个 ROI 示意图; (c) 随机提取后得到的 100 个 ROI。

为了防止同一个样本的方块数据点同时出现在训练集和测试集, 本实验以每个寿眉样本的 100 个 ROI 作为一个单位进行判别分析。本次实验一共 464 个寿眉样本, 共采集到 46400 个 ROI, 464 个数据组。为了防止训练集过拟合, 同时防止因为训练集测试集划分引起的精确率的误差, 本文采用 5 折交叉验证法以增强模型的泛化能力, 即进行 5 次随机划分, 最后取各次结果平均值来评估模型效果。每一次都采用随机挑选法选取 371 个数据组作为训练集, 剩余 93 个数据组作为测试集。

1.3.2 四种预处理方法 因为电噪音、光散射、基线

漂移、光程变化等因素的干扰强度较大, 还需要对光谱数据做必要的预处理。本文采用 4 种光谱预处理算法: 最小二乘平滑滤波 (Savitzky-Golay Filter, SGF)、归一化 (MinmaxScaler, Minmax)、多元散射校正 (Multiplicative Scatter Correction, MSC)、标准正态变换 (Standard Normal Variate, SNV)。后续实验分别建立不同预处理后的光谱与储藏时间的不同判别模型, 通过对比不同预处理后的建模结果, 确定最佳处理组合。

1.3.2.1 最小二乘平滑滤波 (SGF) Savitzky-Golay 滤波器广泛地运用于数据平滑去噪, 是一种基于局部多项式最小二乘法拟合的滤波方法, 其最大特点在于去除噪声的同时可以保留信号的形状、宽度不变^[22]。该方法应用于光谱数据的主要作用是消除随机噪声, 提高信噪比。

1.3.2.2 标准正态变换 (SNV) 粒子大小、散射和多重共线性是漫反射光谱分析中长期存在的问题。这些效应的乘性组合是阻碍解释漫反射光谱的主要因素。样品粒子大小是影响方差的主要因素, 而由化学成分引起的方差很小。SNV 可有效地消除散射、粒径的乘性干扰和多重共线性^[23]。该算法处理光谱数据的基本思想是: 假设每一条光谱中各波长点的反射强度应符合正态分布, 再对每一条原始光谱数据进行标准正态化处理使得校正后的光谱数据均值为 0, 标准差为 1。具体如以下公式:

$$x' = \frac{x - \mu}{\sigma} \quad \text{式 (1)}$$

式中: x 是原始光谱数据; x' 是经过标准正态变换后的光谱数据; μ 为平均值; σ 为原始光谱数据的标准偏差。

1.3.2.3 归一化 (Minmax) Minmax 的原理与 SNV 相似, 目的都是为了方便处理数据, 使得整个数据分布更加均匀, 且两者都能使模型运行时的收敛速度加快^[24]。Minmax 与 SNV 的区别在于, Minmax 是将训练集中某一列数值特征 (假设是第 i 列) 的值缩放到 0 到 1 之间或者 -1 到 1, 从而降低特异样本数据导致的不良影响。计算公式如下:

$$x'' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad \text{式 (2)}$$

式中: x'' 是经过 Minmax 变换后的光谱数据; x_i 是第 i 列的任一原始光谱数据; $\max(x_i)$ 是该列中的最大值; $\min(x_i)$ 是最小值。

1.3.2.4 多元散射校正 (MSC) MSC 主要用来消除固体颗粒大小、表面散射对漫反射的影响。多元散射校正通过校正每个光谱的散射来消除颗粒分布不均匀及颗粒大小产生的影响, 增强了与成分含量相关的光谱吸收信息^[25]。广泛应用于固体漫反射光谱。使用该方法的前提是建立一个待测样品的“理想光谱”, 而实际上并不存在“理想光谱”。该光谱的特点

是样品中的成分含量与其满足线性关系,然后将该光谱作为标准对其他样品的光谱进行修正,其中包括基线平移和偏移校正。在实际应用中,经常将所有光谱的平均光谱视为一个理想的标准光谱。

1.3.3 三种判别模型 本文通过混淆矩阵、精确率、召回率对建立的模型进行评估。以最简单的二分类问题为例,预测结果可分为四类,分别是:原本为正例的样本被预测为正类(True Positive, TP)或负类(False Negative, FN),原本为负例样本被预测为负类(True Negative, TN)或正类(False Positive, FP)。精确率是针对总体预测结果而言的,它表示的是所有预测正确的样本占总观测值的比重,若用 P 表示精确率(Precision),则:

$$P = \frac{TP}{TP + FP} \quad \text{式 (3)}$$

而召回率是针对原来的样本而言的,它表示的是样本中的正例被预测正确的比例,若用 R 表示召回率(Recall),则:

$$R = \frac{TP}{TP + FN} \quad \text{式 (4)}$$

1.3.3.1 支持向量机 支持向量机(Support vector machine, SVM)是将低维空间线性不可分的数据映射到高维,利用支持向量用来寻找一个最优分类超平面,使得不同类别的边缘数据点之间的距离最大化,使得分类器的泛化能力最强,从而实现最优的二分类或多分类。SVM 的核心是将高维特征空间中的内积运算转化为低维空间中的内积核函数,以实现数据在经历非线性转化后的线性划分^[26]。核函数是支持向量机映射数据的重要手段^[27]。在光谱数据分析中,径向基函数(RBF)因其较强的非线性问题处理能力而成为应用最广泛的核函数。本文以径向基函数作为核函数,采用网格搜索法确定支持向量机的超参数(惩罚系数 c 和核宽度 g)。

1.3.3.2 偏最小二乘法联合线性判别 偏最小二乘法(Partial Least-square Method, PLS)是一种基于特征变量的回归方法,其实质是按照协方差极大化准则,在分解自变量变量数据矩阵 X 的同时,也在分解因变量数据矩阵 Y ,并且建立相互对应的解释隐变量与反应隐变量之间的回归关系方程^[28]。偏最小二乘回归的成分之间是相互正交的,这在一定程度上消除了多重线性相关性。线性判别(Linear discriminant analysis, LDA)分类器是一种监督学习的数据降维方法,该方法的核心是将输入变量空间投影到最佳鉴别矢量空间,达到样本在投影后的矢量空间中“类内方差最小,类间方差最大”的目的,即投影后每一种类别数据的投影点尽可能的接近,而不同类别的数据的类别中心之间的距离尽可能的大。之后通过计算待测样本与各中心的距离,待测样本的类别即为距离最短的^[29]。

1.3.3.3 逻辑回归 逻辑回归(Logistic Regression,

LR)是在线性回归的基础上构建一种分类模型,对于特征的线性组合来拟合真实标记为正例的概率的对数几率,其公式可表示为:

$$\left(\ln \frac{y}{1-y} = wx + b \right) \quad \text{式 (5)}$$

式中: w, b 均为参数。

LR 多应用于二分类问题,一个事件发生的机率定义为事件发生的概率与不发生的概率的比值,设 $p = P(y=1|x, w)$, 那么事件的几率是 $p/1-p$, 其对数几率是:

$$\ln \frac{p}{1-p} = \ln \frac{P(y=1|x, w)}{1-P(y=1|x, w)} = wx + b \quad \text{式 (6)}$$

可以看出,输出类别 1 的对数几率函数是输入 x 的线性函数,对数几率函数是 Sigmoid 函数的重要代表^[30]。对于多分类问题,LR 的方法是将多分类转化为所有的多个二分类问题,然后对样本,再依次进行二分类^[31]。

2 结果与分析

2.1 光谱分析

由于三个储藏年份的寿眉样本的光谱曲线在头尾处存在较大的随机误差,首先除去全波段中 450~461 和 963~998 nm 的首尾噪声,以获得波段范围为 462~962 nm 内的 126 个波段的光谱数据进行后续分析。统计每个样本的光谱数据并作图,结果如图 2 所示。根据图 2 可知,650~700 nm 处存在一个波谷,可能是因为叶绿素在 663 nm 处有较大的吸光系数,李晓丽等^[15]利用高光谱成像实现茶叶中 EGCG 可视化研究中有类似研究现象。700~900 nm 处寿眉样本对光谱的反射急剧上升,是因为寿眉对近红外波段吸收较少。每条光谱曲线的形状类似,说明寿眉样本内部的内含成分大致相同。随着储藏时间的变化,寿眉的内含成分的含量发生变化,导致在相同的波段有不同的反射率,这也为寿眉储藏年份测定奠定了客观基础。

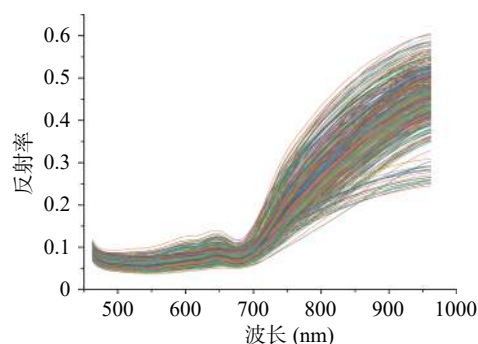


图 2 各储藏年份寿眉样本原始光谱数据图

Fig.2 Diagram of raw spectra of different Shoumei samples

2.2 预处理效果

由图 3 可知,在同一坐标系下,除了 SNV 预处理后的光谱,其余光谱的相对反射率均在 0 以上。

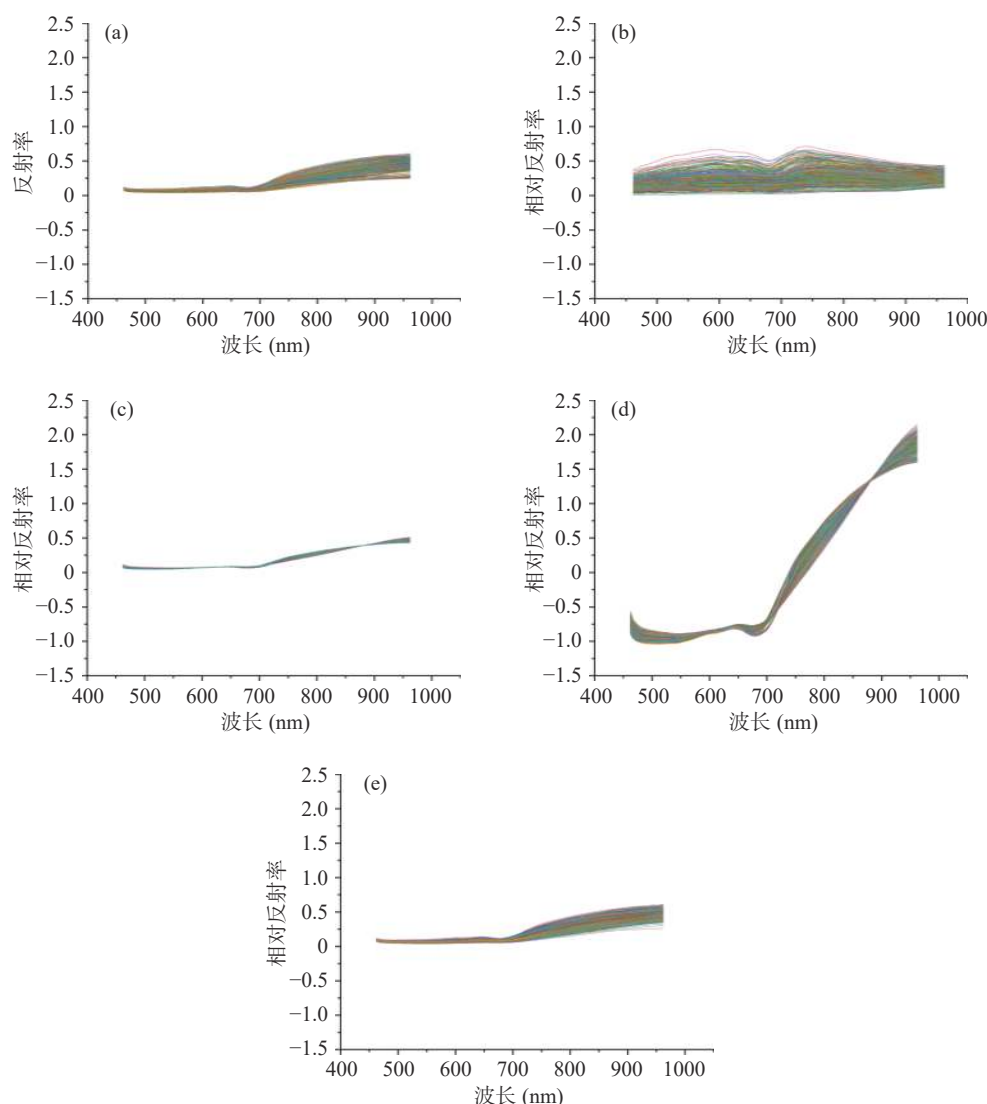


图3 预处理效果图

Fig.3 Spectra of all samples after being preprocessed

注: (a)原始光谱; 经 MinMax(b), MSC(c), SNV(d), SGF(e)预处理后的光谱。

由 1.3.2.3 中的公式可知, SNV 预处理后的相对反射率值为原始反射率减去该波段所有反射率的平均值再除以原始光谱的标准偏差后得到, 所以相对反射率会出现负值。从光谱变化情况可知, 相比于其余预处理, SNV 预处理后的光谱在不同波段的相对反射率变化较大。从光谱分布情况可知, MinMax 预处理后的光谱分布相比于原始光谱更为离散; MSC 预处理后光谱分布相比于原始光谱更为集中, 而 SGF 预处理前后差异不明显。光谱预处理后的具体建模效果还需结合建模结果进一步分析。

2.3 建模结果分析

2.3.1 SVM 建模结果 表 1 是 SVM 建模分析结果, 表 2 是该模型下的最优处理即 SNV-SVM 处理后得到的混淆矩阵。从表 1 可以看出, 经过 SNV 预处理后的建模结果最优, 训练集和测试集的精确率明显高于其他预处理, 分别为 90.83% 和 86.02%。其中没有经过预处理的建模效果在两个数据集上的表现与经过 Minmax 预处理后的差异较小, 说明原始光谱分布较为均匀, 导致该预处理在 SVM 模型中效果一般。从表 2 混淆矩阵中可以看出, SNV-SVM 处

表 1 SVM 建模分析结果

Table 1 Results of SVM modeling analysis

预处理算法	训练集精确率(%)	测试集精确率(%)
无	83.47	77.58
Minmax	81.67	77.34
SGF	87.31	83.29
SNV	90.83	86.02
MSC	84.43	80.67

理组合对于训练集不同年份寿眉样本的预测效果差异不大,但对于测试集中的预测效果差异明显,其中 3、10 年寿眉的召回率接近,分别为 90.32%、89.29%,而 6 年寿眉的召回率只有 79.41%。

2.3.2 PLS-LDA 建模结果 表 3 是 PLS-LDA 建模分析结果,表 4 是该模型下的最优处理即 SNV-PLS-LDA 处理后得到的混淆矩阵。由表 3 可见,经过 SNV 预处理之后的建模结果最佳,训练集和测试集中的精确率分别为 80.05% 和 74.19%;而经过部分预处理后的建模结果相较于无预处理并没有明显的提高。在 PLS-LDA 模型中,与 SVM 建模结果类似,不经过预处理与经过 Minmax 预处理后的建模结果几乎一致;不经过预处理的建模结果比经过 SGF 和 MSC 预处理的建模结果更好,可见原始光谱的噪声较小、基线漂移不明显,预处理可能会造成有效信息的损失。从表 4 混淆矩阵中可见,经过 SNV

预处理后,10 年的寿眉样本的精确率最高,但召回率最低,在测试集中分别为 86.67%、66.67%;3 年的寿眉样本情况则相反,在测试集中的精确率和召回率分别为 62.50%、80.00%。

2.3.3 LR 建模结果 表 5 是 LR 建模分析结果,表 6 是该模型下的最优处理即 Minmax-LR 处理后得到的混淆矩阵。从表 5 可知,与 PLS-LDA 建模结果类似,不经过预处理的建模结果和 SGF、MSC 预处理后的建模结果差异较小。但与其他两种模型不同的是 Minmax 相比于其他预处理在 LR 模型中的表现最佳,训练集和测试集的精确率分别为 84.64% 和 78.50%,这与 LR 模型本身对样本分布的均匀度更为敏感有关。从表 6 混淆矩阵中可以看出,Minmax 模型对三个储藏年份的寿眉样本判别存在差异,且在两个数据集中都对 3 年寿眉样本的判别最佳,6 年其次,10 年最差。

表 2 SNV-SVM 处理的混淆矩阵
Table 2 The confusion matrix in SNV-SVM model

训练集						测试集				
年份(年)	3	6	10	精确率(%)	召回率(%)	3	6	10	精确率(%)	召回率(%)
3	119	5	8	92.97	90.15	28	2	1	87.50	90.32
6	6	113	5	90.40	91.13	3	27	4	87.10	79.41
10	3	7	105	89.00	91.30	1	2	25	83.33	89.29
总值				90.83	90.83				86.02	86.02

注: 寿眉的真实储藏年份(行的数字),模型预测的储藏年份(列的数字),表4、表6同。

表 3 PLS-LDA 建模分析结果
Table 3 Results of PLS-LDA modeling analysis

预处理算法	训练集精确率(%)	测试集精确率(%)
无	77.12	70.73
Minmax	77.17	70.73
SGF	75.03	69.11
SNV	80.05	74.19
MSC	75.29	63.41

表 4 SNV-PLS-LDA 处理的混淆矩阵
Table 4 The confusion matrix in SNV-PLS-LDA model

训练集						测试集				
年份(年)	3	6	10	精确率(%)	召回率(%)	3	6	10	精确率(%)	召回率(%)
3	91	9	9	71.09	83.49	20	2	3	62.50	80.00
6	17	101	4	80.80	82.79	5	23	1	74.19	79.31
10	20	15	105	88.98	75.00	7	6	26	86.67	66.67
总值				80.05	80.05				74.19	74.19

表 5 LR 建模分析结果
Table 5 Results of LR modeling analysis

预处理算法	训练集精确率(%)	测试集精确率(%)
无	77.06	73.45
Minmax	84.64	78.50
SGF	77.06	71.22
SNV	83.69	71.39
MSC	77.06	74.09

表 6 Minmax-LR 处理的混淆矩阵

Table 6 The confusion matrix in Minmax-LR model

年份(年)	训练集					测试集				
	3	6	10	精确率(%)	召回率(%)	3	6	10	精确率(%)	召回率(%)
3	113	8	4	88.28	90.40	27	4	2	84.38	81.82
6	5	103	16	82.40	83.06	0	24	6	77.42	80.00
10	10	14	98	83.05	80.33	5	3	22	73.33	73.33
总值				84.64	84.64				78.50	78.50

表 7 三种判别模型的最优处理汇总

Table 7 Results of three discriminant models with their best preprocessing algorithm

处理组合	训练集精确率(%)	测试集精确率(%)
SNV-SVM	90.83	86.02
SNV-PLS-LDA	80.05	74.19
Minmax-LR	84.64	78.50

2.3.4 建模结果小结 从表 7 可以得出, SVM 模型对于寿眉样本的判别效果要明显优于 PLS-LDA、LR 模型, 其中最佳处理组合 SNV-SVM 在训练集和测试集的精确率分别为 90.83%、86.02%。四种预处理方法中, SNV 在三种判别模型中的表现较好, Minmax 其次, SGF 和 MSC 表现一般。该结果与原始光谱数据特点相关, 原始光谱数据所受到随机噪声和基线漂移的影响都较小, SGF 和 MSC 预处理后可能降低原始光谱中的有效信息, 导致后续的建模效果相较于无预处理并不明显, 在 PLS-LDA 模型中甚至更差。

3 结论

本文通过高光谱成像技术获得 450~998 nm 波段范围内的高光谱数据对储藏年份为 3、6、10 年的寿眉样本进行无损检测研究。采用 Minmax、SGF、SNV、MSC 4 种算法来对寿眉原始光谱数据进行预处理, 并分别建立 SVM、PLS-LDA、LR 判别模型。建模结果表明, 相较于线性模型 PLS-LDA、LR, 非线性模型 SVM 对于寿眉样本的高光谱数据有更好的判别效果, 说明光谱数据和白茶储藏年份之间的关系更可能是一个非线性关系, 这与白茶中多种内含成分在储藏过程中不断发生复杂的化学变化有关。本文所建立的 SNV-SVM 处理组合针对 3、6、10 年这三个年份的寿眉有较高的判别精确率以及较强的泛化能力, 而对于判别其他品类和其他年份的白茶还需后续进一步研究。

参考文献

- [1] Liu L, Liu B, Li J, et al. Responses of different cancer cells to white tea aqueous extract: Response to white tea aqueous extract[J]. Journal of Food Science, 2018, 83(10-12): 2593-2601.
- [2] Orner G A. Suppression of tumorigenesis in the Apcmin mouse: Down-regulation of β -catenin signaling by a combination of tea plus sulindac[J]. Carcinogenesis, 2003, 24(2): 263-267.
- [3] Rasolifard M H, Nozohor Y. The effect of aqueous extract of white tea on the levels of antioxidant enzymes of rats' liver tissue exposed to arsenic[J]. Journal of Ilam University of Medical

Sciences, 2018, 25(5): 146-153.

- [4] Sangeeta Rangi, Sunil Kumar Dhatwalia, Priti Bhardwaj, et al. Evidence of similar protective effects afforded by white tea and its active component 'EGCG' on oxidative-stress mediated hepatic dysfunction during benzo(a)pyrene induced toxicity[J]. Food and Chemical Toxicology, 2018, 116: 281-291.
- [5] Camouse Melissa M, Domingo Diana Santo, Swain Freddie R, et al. Topical application of green and white tea extracts provides protection from solar-simulated ultraviolet light in human skin[J]. 2009, 18(6): 522-526.
- [6] 刘琳燕, 周子维, 邓慧莉, 等. 不同年份白茶的香气成分[J]. 福建农林大学学报(自然科学版), 2015, 44(1): 27-33. [Liu L Y, Zhou Z W, Deng H L, et al. Analysis of the aromatic components in white tea produced in different years[J]. Journal of Fujian Agriculture and Forestry University (Natural Science Edition), 2015, 44(1): 27-33.]
- [7] Ning J M, Ding D, Song Y S, et al. Chemical constituents analysis of white tea of different qualities and different storage times[J]. European Food Research and Technology, 2016, 242(6): 2093-2104.
- [8] Dai W, Xie D, Lu M, et al. Characterization of white tea metabolome: Comparison against green and black tea by a nontargeted metabolomics approach[J]. Food Research International, 2017, 96(6): 40-45.
- [9] Xie D, Dai W, Lu M, et al. Nontargeted metabolomics predicts the storage duration of white teas with 8-C N-ethyl-2-pyrrolidinone-substituted flavan-3-ols as marker compounds[J]. Food Research International, 2019, 125(11): 7209-7218.
- [10] 刘鹏, 艾施荣, 杨普香, 等. 非线性流形降维方法结合近红外光谱技术快速鉴别不同海拔的茶叶[J]. 茶叶科学, 2019, 39(6): 715-722. [Liu P, Ai S R, Yang P X, et al. Nonlinear manifold dimensionality reduction methods for quick discrimination of tea at different altitude by near infrared spectroscopy[J]. Tea Science, 2019, 39(6): 715-722.]
- [11] Wang Y J, Li T H, Li L Q, et al. Micro-NIR spectrometer for quality assessment of tea: Comparison of local and global models[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 237: 118403.

- [12] 刘洪林. 基于近红外光谱技术(NIRS)对工夫红茶审评品质客观评价研究[J]. 食品工业科技, 2016, 37(5): 311–315. [Liu H L. Research of evaluation the quality of Congou black tea by near infrared spectroscopy[J]. Science and Technology of Food Industry, 2016, 37(5): 311–315.]
- [13] 李春霖. 基于化学计量学和近红外光谱技术的龙井茶感官及化学品质评价研究[D]. 杭州: 浙江大学, 2019. [Li C L. Sensory and chemical quality evaluation of Longjing tea using chemometrics and near-infrared spectroscopy technique[D]. Hangzhou: Zhejiang University, 2019.]
- [14] 于英杰, 王琼琼, 王冰玉, 等. 基于高光谱技术的铁观音茶叶等级判别[J]. 食品科学, 2014, 35(22): 159–163. [Yu Y J, Wang Q Q, Wang B Y, et al. Identification of Tieguanyin tea grades based on hyperspectral technology[J]. Food Science, 2014, 35(22): 159–163.]
- [15] 李晓丽, 魏玉震, 徐劼, 等. 基于高光谱成像的茶叶中EGCG分布可视化[J]. 农业工程学报, 2018, 34(7): 180–186. [Li X L, Wei Y Z, Xu J, et al. EGCG distribution visualization in tea leaves based on hyperspectral imaging technology[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(7): 180–186.]
- [16] 李瑶. 基于高光谱成像的茶叶品质判别方法研究[D]. 成都: 四川农业大学, 2018. [Li Y. A method of distinguishing tea quality based on hyperspectral imaging[D]. Chengdu: Sichuan Agricultural University, 2018.]
- [17] Ren G, Wang Y, Ning J, et al. Using near-infrared hyperspectral imaging with multiple decision tree methods to delineate black tea quality[J]. Spectrochimica Acta Part A Molecular and Biomolecular Spectroscopy, 2020, 237: 118407.
- [18] Hong Z, He Y. Rapid and nondestructive discrimination of geographical origins of Longjing tea using hyperspectral imaging at two spectral ranges coupled with machine learning methods[J]. Applied Science, 2020, 10(3): 1173–1184.
- [19] 魏玉震. 基于光谱和光谱成像技术的茶叶含水率检测机理和方法研究[D]. 杭州: 浙江大学, 2019. [Wei Y Z. Moisture content detection of tea leaves based on spectral and spectral imaging technologies[D]. Hangzhou: Zhejiang University, 2019.]
- [20] 芦兵, 孙俊, 杨宁, 等. 基于荧光透射谱和高光谱图像纹理的茶叶病害预测研究[J]. 光谱学与光谱分析, 2019, 39(8): 2515–2521. [Lu B, Sun J, Yang N, et al. Prediction of tea disease based on fluorescence transmission spectrum and texture of hyperspectral image[J]. Spectroscopy and Spectral Analysis, 2019, 39(8): 2515–2521.]
- [21] 田欣. 高光谱图像的去噪模型与算法[D]. 长沙: 湖南大学, 2019. [Tian X. Models and algorithms for hyperspectral image restoration[D]. Changsha: Hunan University, 2019.]
- [22] H Martens, S Å Jensen, P Geladi. Multivariate linearity transformation for near-infrared reflectance spectrometry[C]// Proc Nordic Symp on Applied Statistics(O. H. J. Christie, ed.), Stokkand Forlag Publ, Stavanger, Norway, 1983: 208–234.
- [23] 郭辉, 彭彦昆, 江发潮, 等. 手持式牛肉大理石花纹检测系统[J]. 农业机械学报, 2012, 43: 207–210. [Guo H, Peng Y K, Jiang F C, et al. Development of conveyable beef-marbling detection system[J]. Transactions of the Chinese Society of Agricultural Machinery, 2012, 43: 207–210.]
- [24] Shaheen H, Agarwal S, Ranjan P. MinMaxScaler binary PSO for feature selection[M]. First International Conference on Sustainable Technologies for Computational Intelligence, 2020: 705–716.
- [25] M Dhanoa, S Lister, R Sanderson, et al. The link between multiplicative scatter correction(MSC) and standard normal variate(SNV) transformations of NIR spectra[J]. Journal of Near Infrared Spectroscopy, 1995, 2(1): 43–47.
- [26] Saunders C, Stitson M O, Weston J, et al. Support vector machine[J]. Computer Ence, 2002, 1(4): 1–28.
- [27] MATLAB 中文论坛. MATLAB 神经网络 30 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010. [Matlab Chinese Forum. Matlab neural network analysis of 30 cases[M]. Beijing: Beihang University Press, 2010.]
- [28] Brereton R G, Lloyd G R. Partial least squares discriminant analysis: Taking the magic away[J]. Journal of Chemometrics, 2014, 28: 214–225.
- [29] Miguel Pérez-Enciso, Tenenhaus M. Prediction of clinical outcome with microarray data: A partial least squares discriminant analysis(PLS-DA) approach[J]. Human Genetics, 2003, 112(5-6): 581–592.
- [30] Sabouri S, Hajrasouliha A, Song Y, et al. Logistic regression[M]. Basic Quantitative Research Methods for Urban Planners, 2020: 15–24.
- [31] Allison P D. Logistic regression using the SAS system: Theory and application[M]. SAS Publishing, 1999: 81–90.