基于噪声鲁棒性特征提取的普洱茶品种鲁棒判别方法

赵秀芝[1,2]，谢德红[3]

（1. 浙江工贸职业技术学院 人工智能学院，浙江 温州 325002；2. 温州大学 计算机与人工智能学院，浙江 温州 325002；3. 南京林业大学 信息科学技术学院，江苏 南京 210037）

摘　要：利用近红外光谱与机器学习方法快速鉴别普洱茶叶品质时，中低端近红外光谱采集设备采集的光谱存在高维、重叠和噪声大的特性，严重影响了建模准确。本文提出了一个噪声鲁棒的特征提取方法，与支持向量机（SVM）分类器结合，建立普洱茶叶品质鉴别方法。首先，利用噪声鲁棒的特征提取方法、主成分分析（PCA）与连续投影算法（SPA）对获得的近红外光谱数据进行特征提取，获得特征空间；然后利用SVM对特征提取后的数据进行训练，获得鉴别模型。模型鉴别结果比较表明，对于噪声残留近红外光谱数据，本文提出的噪声鲁棒特征提取方法能够有效抵抗噪声的影响、从高维光谱中提出特征变量，以提高模型的鉴别精度。鉴别模型预测的正确率、召回率、特效度、准确率及平衡F分数分均明显高于其他两种方法所得模型。对于古树普洱茶叶与非古树普洱茶叶的鉴别，本文鉴别模型预测的正确率和召回率分别达到了92.06%和95.38%，表明本文方法训练所得模型具有较好的鉴别能力。研究结果为实现在实际应用中精准判别普洱茶品质提供理论参考和依据。

关键词：近红外光谱；噪声；快速鉴别；普洱茶；特征提取；机器学习

中图分类号：O657.3　　**文献标志码**：A　　**文章编号**：1671-5497(2025)05-1756-07

DOI：10.13229/j.cnki.jdxbgxb.20240566

Discrimination method for Pu-er tea varieties based on noise-robust feature extraction

ZHAO Xiu-zhi[1,2], XIE De-hong[3]

(1. *College of Artificial Intelligence, Zhejiang Industry& Trade Vocational College, Wenzhou*, 325002, *China*; 2. *School of Computer and Artificial Intelligence, Wenzhou University, Wenzhou* 325002, *China*; 3. *College of Information Science and Technology, Nanjing Forestry University, Nanjing* 210037, *China*)

Abstract: When using near-infrared spectroscopy and machine learning methods to quickly identify the quality of Pu-er tea, the spectra collected by medium and low-end near-infrared spectroscopy acquisition equipment have the characteristics of high dimension, overlap and large noise, which seriously affects the

收稿日期：2024-05-22.

基金项目：茶树生物学与资源利用国家重点实验室开发基金项目（SKLTOF090113）；食品安全大数据技术北京市重点实验室开放基金项目（BTBD-2019KF02）.

作者简介：赵秀芝（1978-），女，教授，硕士. 研究方向：图像处理和信号通信. E-mail：rgznzxz@zjitc.edu.cn

通信作者：谢德红（1979-），女，讲师. 研究方向：深度学习和计算机视觉. E-mail：dehong.xie@gmail.com

accuracy of modeling. This paper proposes a noise-robust feature extraction method, which is combined with support vector machine (SVM) classifier to establish the quality identification method of Pu-er tea. Firstly, the noise-robust feature extraction method, principal component analysis (PCA) and successive projections algorithm (SPA) are used to extract the features from the obtained near-infrared spectral data. Then, SVM is used to train the data after feature extraction to obtain the identification model. The comparison of the identification results of the model shows that for the noiseresidual near-infrared spectral data, the noise robust feature extraction method in this paper can effectively resist the influence of noise and propose feature variables from the high-dimensional spectrum to improve the accuracy of the identification model. The accuracy, recall, specificity, accuracy and F-score predicted by the identification model were significantly higher than those obtained by the other two methods. For the detection of ancient Pu-er tea and non-ancient Pu-er tea, the accuracy and recall predicted by the identification model in this paper have reached 92.06% and 95.38% respectively, indicating that the identification model has good identification ability. The research results provide theoretical reference and basis for accurately judging the quality of Pu-er tea in practical application.

Key words: near-infrared spectroscopy; noise; rapid identification; Pu-er tea; feature extraction; machine learning

0　引　言

普洱茶茶香浓郁、回味无穷,深受我国广大消费者喜爱。其中,古树普洱茶叶滋味醇厚、浓烈、回甘快、韵味持久,质量远高于台地普洱(或非古树普洱)[1],且低污染、低农残,感官质量评分较高、市场价格也较高[2],深受茶商和消费者的喜爱,但是目前普洱茶市场比较混乱,以次充好、以非充古的现象比比皆是。普通消费者不具有普洱茶品质鉴别的专业知识,很难从混乱普洱茶市场上分辨普洱茶的品质(或辨别出价格昂贵的古树普洱茶叶)。这严重损害消费者权益,也给种植者和经营者带来负面影响。因此,方便、有效、准确的普洱茶品质鉴别是十分必要的。

近年,利用近红外光谱技术获得茶叶的内部特征信息,并结合机器学习技术,实现了茶叶品质鉴别[3-5]。此方法具有快速、无损、成本低等优点,在茶叶品质和品种的鉴别领域中得到广泛关注。然而,当此方法被广泛推广应用时,难以实现利用高端、大型近红外光谱采集设备采集近红外光谱数据,一般采用中端,甚至低端设备(如光纤光谱仪)。在利用这些中低端设备采集光谱时,由于设备的光学系统、光源、检测器、电子元件、电路设备,以及采集外部环境干扰等因素产生大量、复杂的噪声,严重影响光谱数据在机器学习中的准确性和精确性[6,7]。此噪声有高斯、脉冲等加性噪声,也有泊松等相关噪声,且有强有弱,因而降噪方法很难完全消除此噪声,取得理想效果。此外,近红外光谱维度高,且很多波段与噪声一样,为无效或干扰信息,易使后续的分类模型过拟合,从而严重影响鉴别结果。因此,在鉴别流程的特征提取或特征选取阶段,避免残留噪声影响、提取有利于鉴别的特征,对提升鉴别精度十分必要[8]。

针对高维近红外光谱,常见特征选取或提取方法有连续投影算法(Successive projections algorithm, SPA)[5,9]与主成分分析(Principal component analysis, PCA)[4,10]。SPA[11]利用向量的投影分析,将波段投影到其他波段上,比较投影向量大小,以投影向量最大的波段为待选波段。SPA选择的是含有最少冗余信息及最小共线性的变量组合,降低了波段之间相关性,有利于数据后续分类[12],但此方法假设光谱数据是干净的。相对SPA,PCA[13]可从残留噪声的光谱数据中去除冗余信息、提取有效特征,并在一些应用中取得了较好的效果。PCA特征提取的鲁棒性针对的是含高斯噪声的数据,受噪声破坏的、真实近红外光谱不符合PCA的假设。因此,在实际应用中,构建鲁棒的特征提取方法,对从已被噪声破坏的近红外光谱数据中学习有效特征,从而对茶叶品质进行鉴别非常关键。

针对中低端近红外光谱采集设备采集的近红外光谱,本文提出一种噪声鲁棒的特征提取方法,并结合支持向量机(Support vector machine,

SVM)[14]分类器建立茶叶品质分类模型,以改善模型鉴别普洱茶叶品质的精度,为在实际应用中精准判别普洱茶品质提供理论参考和依据。

1 材料与方法

1.1 样本制备

试验前于云南产地购买了两种普洱茶叶:古树普洱(勐宋古树)和非古树普洱(云饼茶)。试验从茶饼上取出茶叶、磨碎,装满直径4 cm、高2 cm的黑色广口试剂皿,然后利用压紧器压紧、排出空气,并使底部无缝隙,然后贴上标签。每种普洱茶叶取150份、每份构成1个样品,共计300个样品。将制备的样本用保鲜膜密封保存,等待近红外光谱采集设备采集。

1.2 光谱采集

本试验所采用的近红外光谱采集设备为NIR1.7光纤光谱仪(德国INSION),主要由光纤、光纤卤素灯、CCD构成。此光纤光谱仪的扫描步长为8 nm,扫描范围为844~1 894 nm。在室温为20 ℃的室内环境中,将光谱仪预热30 min。打开SPECview软件,将光源投射于标定白板上,手持光谱仪的探头对准标定白板上一点,进行校正;然后手持光谱仪探头对黑色广口试剂皿进行背景扫描,以消除仪器本身和环境中可能对测量结果产生的干扰;最后,手持光谱仪探头对准制备好的普洱茶叶样品,测量并保存茶叶的近红外漫反射光谱数据。为了避免采集光谱过度平滑、损失重要光谱信息,每个样品随机选取3个位置扫描、每个位置扫描5次,3个位置测量光谱的平均作为此样品的近红外漫反射光谱。测量多个不同位置可扫除样品不均匀引起的散射误差。两类普洱茶叶的近红外光谱如图1所示。

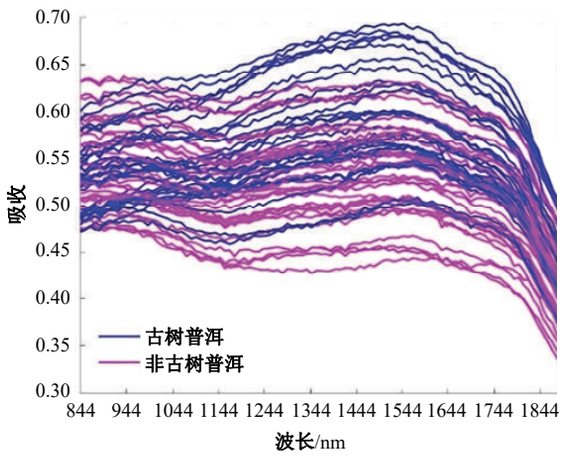

图1 普洱茶叶的近红外光谱

Fig. 1 NIR spectra of Pu-er tea samples

1.3 方法

本文利用机器学习方法创建普洱茶叶品质的鉴别模型,具体如图2所示。首先,将采集的数据分为训练集与测试集,利用特征提取或特征选取方法从训练集的光谱数据中提取特征,并获得相应的特征空间;其次,利用分类器训练样本在特征空间中的数据,获得分类模型;再次,将测试集中的近红外光谱映射到上述特征空间中;最后,利用分类模型鉴别映射后的数据,获得测试光谱所属普洱茶叶类别。

1.3.1 特征提取

如上所述,考虑实际应用中,近红外光谱数据中残留噪声,这不利于分类。本文提出一种噪声鲁棒的特征提取方法,方法的优化方程表示如下:

$$\min_{M,Z}\|Z\|_*+\alpha\|Z\|_1+\beta\|M^{\mathrm{T}}X-M^{\mathrm{T}}XZ\|_{2,1}$$
$$\text{s.t. } M^{\mathrm{T}}M=I,\ x_i=Xz_i,\ 1^{\mathrm{T}}z_i=1 \tag{1}$$

式中:$X=[x_1,x_2,\cdots,x_n]\in R^{m\times n}$为训练集中$n$个样本的近红外光谱数据矩阵,矩阵中列向量$x_i$为第$i$个样本的近红外光谱,$m$为样本近红外光谱的波段数量;$M\in R^{m\times d}$为特征选取的映射矩阵;$d$为提取特征的数量(即特征数);$I\in R^{d\times d}$为单位矩

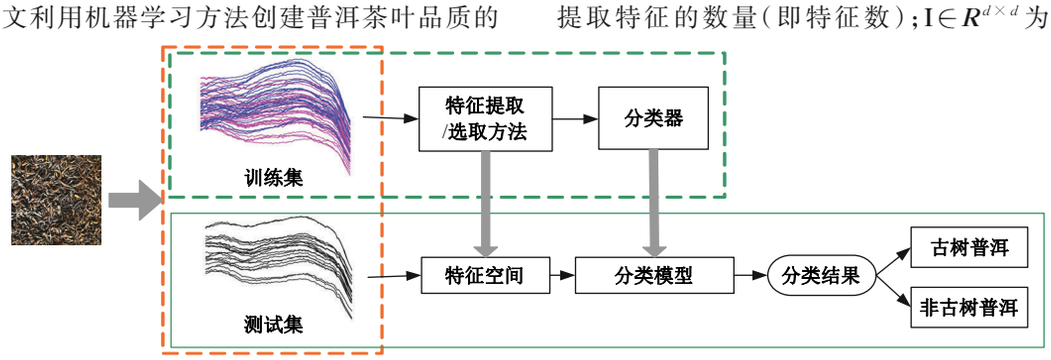

图2 茶叶品质的鉴别流程

Fig. 2 Flowchart of tea quality identification

阵；$Z=[z_1,z_2,\cdots,z_n]\in R^{n\times n}$为稀疏权重矩阵，$\mathbf{1}=[1\quad 1\quad \cdots\quad 1]^T\in R^n$；$\alpha>0$和$\beta>0$为正则化参数；$\|\cdot\|_1$为$l_1$范数；$\|\cdot\|_*$为核范数；$\|\cdot\|_{2,1}$为$l_{2,1}$范数；$E=M^TX-M^TXZ$为重建误差。由此正则化项$\|M^TX-M^TXZ\|_{2,1}$可控制提取特征受噪声的影响，最终得到特征提取后的数据：

$$\begin{cases} F=M^TX \\ F_t=M^TX_t \end{cases} \tag{2}$$

式中：$F\in R^{d\times n}$为训练光谱数据特征提取后的数据；$F_t\in R^{d\times t}$为测试光谱数据特征提取后的数据；$X\in R^{m\times t}$为测试光谱数据，t为测试集的样本数量。

1.3.2　分类器

在机器学习中，SVM是一个非常经典的分类器，在解决小样本分类问题上存在优势，因而在茶叶品质与品种鉴别中有较好的应用效果[14]。SVM分类器基本思想是寻找一个分类超平面使训练样本的各类样本分隔最大，即目标是寻找最优分类超平面。SVM利用核函数通过非线性映射，将输入向量映射到高维空间，在高维空间中构建分类超平面。在本文方法中，核函数采用径向基函数。

2　结果与分析

2.1　样本集的划分

训练集和测试集样本采用随机的方法进行划分，即从150个古树普洱样本和150个非古树普洱样本中分别随机选取70%作为训练集，剩余的30%作为测试集。为了增加所得分类模型稳定性，测试集将采取十折交叉的方式[15]，将原训练集随机分成10份，轮流将其中9份作为训练集，1份作为验证集，从而获得更好的分类模型。

2.2　特征提取结果与分析

本文的最终目的是为实现茶叶品质无损鉴别技术提供理论基础。但是，在实际应用中，中低端近红外采集设备采集的近红外光谱含复杂噪声，采集软件与后续降噪方法无法完全去除此噪声，而噪声在特征提取阶段会严重干扰有效特征的提取，最终影响鉴别结果。因此，本小节将通过样本在特征空间的分布以及分类模型的决策边界来验证和分析本文的特征提取方法在含噪近红外光谱特征提取中的有效性。

图3显示了PCA、SPA和本文方法从含噪近红外光谱中提取特征后在2D特征空间的样本分布和相应的决策边界。其中，决策边界是由训练集经特征提取后，利用SVM分类器训练所得分类模型决定，为在2D空间中样本分类边界。首先，从图1中看，古树普洱茶叶和非古树普洱茶叶两类样本的近红外光谱的波形是存在差异性的，因而理论上是可分的。然而，对分类而言，光谱数

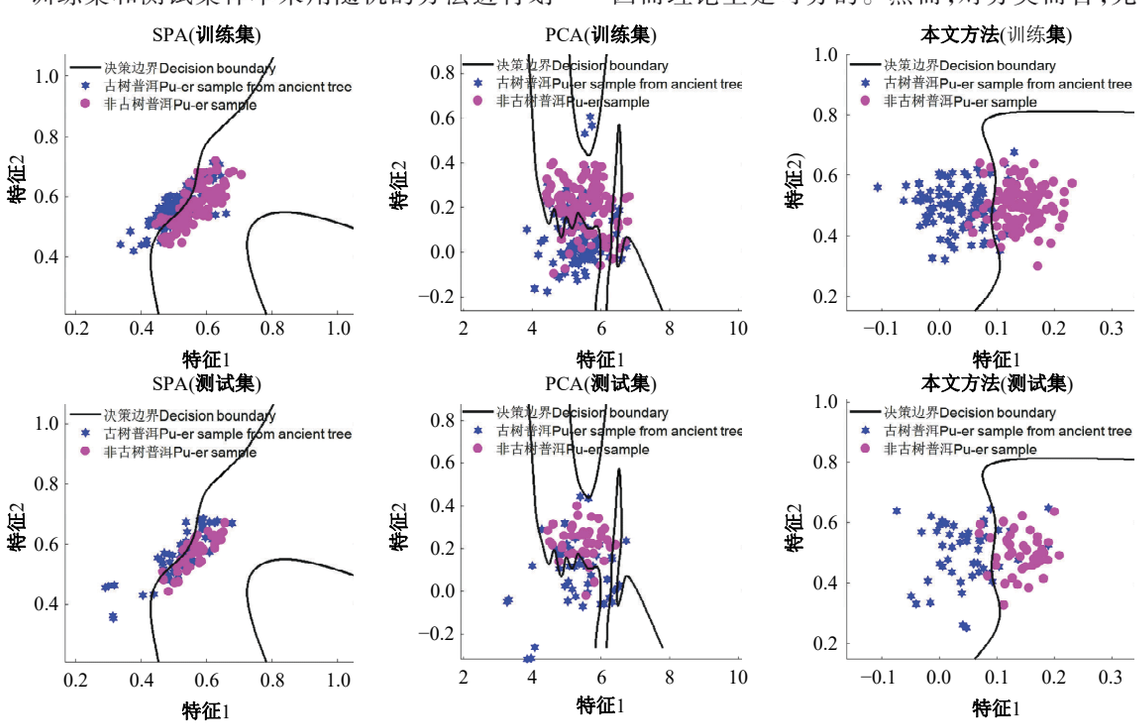

图3　样本点在2D特征空间的分布：PCA，SPA，本文方法

Fig. 3　2D projections of sample points by feature extraction method: PCA, SPA, our method

据维数越高、样本在原光谱空间中的差异性越不明显。特征提取的目的就是去除冗余信息,增加数据在其空间的差异性及样本分布的可分性。但是,从图3训练集的样本分布看,光谱数据经PCA和SPA方法提取特征后,古树普洱样本和非古树普洱样本在特征空间中分布的重叠性高,很难将两类样本分开。此现象说明,对PCA和SPA而言,提取特征时无法有效避开噪声的干扰,提取特征后数据的差异性不但未增加,反而有减小的迹象,从而使样本在特征空间中的分布高度重叠。图3右上图显示,本文方法特征提取后两类样本分布重叠程度低,表明本文方法在特征提取时受噪声影响小、特征提取后数据差异性变大,更有利于噪声干扰下的特征提取。其次,依据机器学习理论,训练数据特征提取后样本分布重叠度越高,所得决策边界集越复杂,则越易过度拟合训练数据,基于决策边界的测试数据分类准确率则越低。由此,如图3下排测试集样本分布图所示,相对PCA和SPA,本文方法分类的决策边界相对简单,且测试集特征提取后各类样本基本清晰地分布在决策边界划定的各类区域内,这表明利用本文方法所得分类模型的判别准确性高。

2.3 判别结果分析

为了进一步验证本文方法对含噪近红外光谱及其在茶叶品质鉴别中的有效性,本文利用正确率、召回率、特效度、准确率、平衡F分数($F1$_score)等指标[16]对分类模型的鉴别结果进行评价。正确率是指被分对的样本占所有样本的比例。通常,正确率越高,模型的鉴别效果越好。召回率是指被分对的正例样本占所有正例样本的比例,特效度是指被分对的负例占所有负例样本的比例。本文中,本古树普洱茶样本被标注为正例样本,非古树普洱茶样本则被标注为负例样本。因此,召回率衡量的是古树茶叶的鉴别能力,特效度衡量的则是非古树茶叶的鉴别能力。准确度是指被划分为正例的样本中实际为正例的比例,也称为查准率。平衡F分数为召回率和准确率加权平均,平衡F分数越高说明方法越有效。

图4为3种分类模型的训练集和测试集在不同特征数量下的鉴别正确率。其中,这3种分类模型分别由PAC、SPA和本文特征提取方法结合SVM分类器训练所得。图4中,一种颜色对应一种方法所得的分类模型;其中,同颜色的虚线表示训练集的鉴别正确率,实线则表示测试集的鉴别正确率。从图4发现,对于测试集,本文方法总体正确率最高,与PCA-SVM与SPA-SVM比存在明显优势;且当特征数大于10时,正确率基本稳定,而PCA-SVM和SPA-SVM随特征数有轻微的波动趋势。此外,图4还展示,本文方法的训练集与测试集总体正确率差异小,而PCA-SVM和

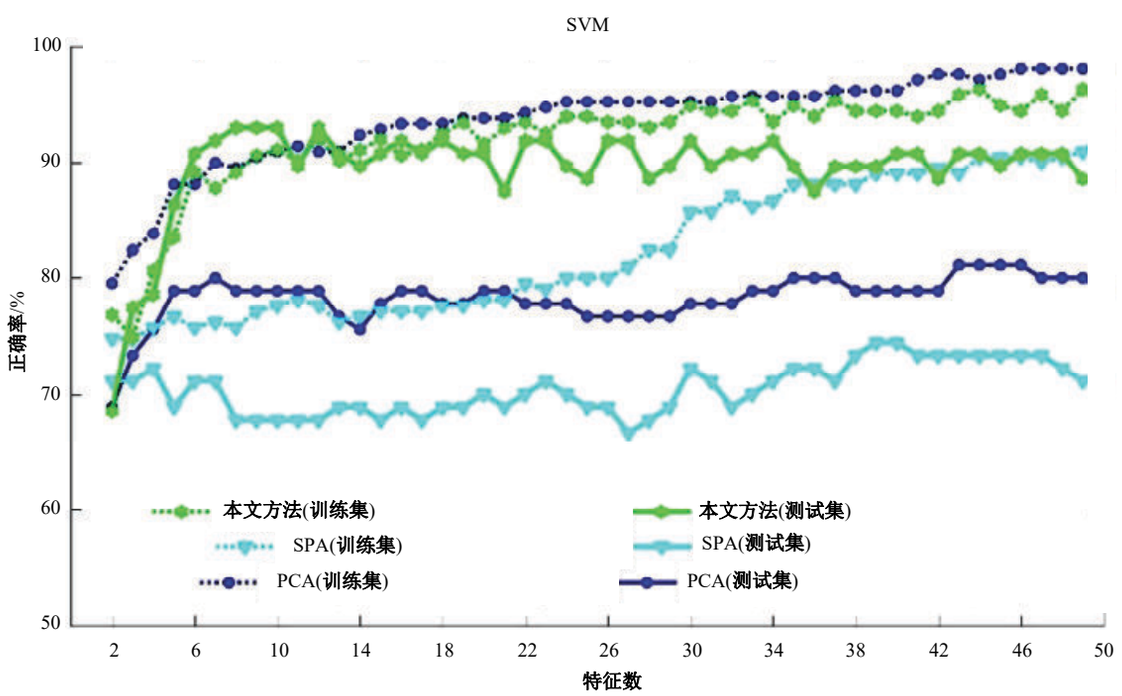

图4 鉴别的正确率与特征数的曲线图

Fig. 4 Accuracy of identification model varying with the extracted features count

SPA-SVM的训练集与测试集总体正确率差异稍大。此现象表明,相对PCA-SVM和SPA-SVM,本文方法获得的分类模型能更有效地避免过度拟合。

表1展示了特征数降到20时各方法的各评价指标值。由表1可看出,本文方法获得的普洱茶叶鉴别的正确率、召回率、特效度、准确率,以及平衡F分数分别达到了92.06%、95.38%、88.52%、89.86%、92.25%,均明显高于其他两种方法对应评价指标值,说明本文方法能较好地排除中低端光谱采集设备引入的噪声干扰,能有效地鉴别古树普洱茶叶与非古树普洱茶叶。此外,召回率和特效度可分别衡量方法对古树普洱茶叶和非古树普洱茶叶的鉴别能力。表1中,本文方法的召回率为95.38%,远高于其他方法,说明本文方法可很好地鉴别出价格昂贵的古树普洱茶叶,进一步说明本文方法对噪声残留的光谱具有较好的鲁棒性,适合中低端近红外光谱采集设备在茶叶鉴别中的应用。

表1　普洱茶叶品质鉴别结果

Table1　Quality identification results of Pu-er tea

方法	正确率	召回率	特效度	准确率	平衡F分数
PCA-SVM	0.842 9	0.853 2	0.831 7	0.845 5	0.849 3
SPA-SVM	0.700 0	0.629 6	0.805 6	0.829 3	0.715 8
本文方法	0.920 6	0.953 8	0.885 2	0.898 6	0.925 4

3　结束语

本文针对中低端近红外光谱采集设备的问题,提出的噪声鲁棒的特征提取方法较好地解决了采集的近红外光谱数据维度高、重叠、高噪声的问题。本文利用该方法与SVM分类器结合,对古树普洱茶叶与非古树普洱茶叶进行了鉴别研究,通过试验发现本文的鉴别方法总体的鉴别正确率可达92.06%,召回率达95.38%,远高于现有方法所得结果。研究结果表明:本文方法适合中低端近红外光谱采集设备在鉴别中的应用,能够实现普洱茶叶品质无损鉴别,可望解决实际应用环境中的一个鉴别难题。

参考文献:

[1] 赵阳, 龚加顺, 王秋萍. 古树普洱茶生茶贮藏过程中香气成分的变化[J]. 食品科学, 2022, 43(4): 241-248.
Zhao Yang, Gong Jia-shun, Wang Qiu-ping. Change in aroma components of raw pu-erh tea from ancient tea trees during storage[J]. Food Science, 2022, 43(4): 241-248.

[2] 曾敏, 龚正礼. 基于主成分分析法构建云南古树普洱生茶香气质量评价模型[J]. 食品工业科技, 2017, 38(15): 264-269.
Zeng Min, Gong Zheng-li. Modeling for aroma quality evaluation of Yunnan Pu-erh raw tea made from ancient trees based on principal component analysis[J]. Science and Technology of Food Industry, 2017, 38(15): 264-269.

[3] 吴全金, 周喆, 孙威江. 近红外光谱技术在茶叶品质调控中的应用[J]. 广东农业科学, 2019, 46(1): 91-100.
Wu Quan-jin, Zhou Zhe, Sun Wei-jiang. Review on the application of near-infrared spectroscopy technology in tea quality management[J]. Guangdong Agricultural Sciences, 2019, 46(1): 91-100.

[4] 王胜鹏, 龚自明, 高士伟, 等. 基于近红外光谱技术的恩施玉露茶保存年份的快速无损鉴别[J]. 华中农业大学学报, 2015, 34(5): 111-114.
Wang Sheng-peng, Gong Zi-ming, Gao Shi-wei, et al. Identification of Enshi yulu tea conserved years based on near infrared spectroscopy[J]. Journal of Huazhong Agricultural University, 2015, 34(5): 111-114.

[5] Ren G, Wang Y, Ning J, et al. Highly identification of keemun black tea rank based on cognitive spectroscopy: near infrared spectroscopy combined with feature variable selection[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 230: 118079.

[6] 韩广, 王小燕, 陈思琪, 等. 提高近红外光谱法检测人体血液等复杂溶液成分准确度的研究进展[J]. 光谱学与光谱分析, 2021, 41(7): 1993-1997.
Han Guang, Wang Xiao-yan, Chen Si-qi, et al. Research progress on improving the accuracy of near inf rared spectroscopy detection of human blood and other complex solution components[J]. Spectroscopy and Spectral Analysis, 2021, 41(7): 1993-1997.

[7] 谢德红, 李俊锋, 刘葫, 等. 基于改进Hodrick-Prescott分解模型的近红外自适应降噪方法[J]. 光谱学与光谱分析, 2020, 40(5): 1650-1655.
Xie De-hong, Li Jun-feng, Liu Di, et al. An improved hodrick-prescott decomposition based near-infrared adaptive denoising method[J]. Spectroscopy and Spectral Analysis, 2020, 40(5): 1650-1655.

[8] Yang H, Li L L, Li G H, et al. A novel feature ex-

traction method for ship-radiated noise[J]. Defence Technology, 2022, 18(4): 604-617.

[9] 董春旺, 梁高震, 安霆, 等. 红茶感官品质及成分近红外光谱快速检测模型建立[J]. 农业工程学报, 2018, 34(24): 306-313.
Dong Chun-wang, Liang Gao-zhen, An Ting, et al. Near-infrared spectroscopy detection model for sensory quality and chemical constituents of black tea[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(24): 306-313.

[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 Peng, Ai Shi-rong, Yang Pu-xiang, et al. Nonlinear manifold dimensionality reduction methods for quick discrimination of tea at different altitude by near infrared spectroscopy[J]. Journal of Tea Science, 2019, 39(6): 715-722.

[11] Canova L D S, Vallese F D, Pistonesi M F, et al. An improved successive projections algorithm version to variable selection in multiple linear regression[J]. Analytica Chimica Acta, 2023, 1274: 341560.

[12] Pang L, Wang L, Yuan P, et al. Rapid seed viability prediction of Sophora japonica by improved successive projection algorithm and hyperspectral imaging[J]. Infrared Physics & Technology, 2022, 123: 104143.

[13] Ghosh T, Kirby M. Linear centroid encoder for supervised principal component analysis[J]. Pattern Recognition, 2024, 155: 110634.

[14] Cardoso V G K, Poppi R J. Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine[J]. Microchemical Journal, 2021, 164: 106052.

[15] Pang Y, Wang Y, Lai X, et al. Enhanced kriging leave-one-out cross-validation in improving model estimation and optimization[J]. Computer Methods in Applied Mechanics and Engineering, 2023, 414: 116194.

[16] Luque A, Carrasco A, Martín A, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix[J]. Pattern Recognition, 2019, 91: 216-231.