第42卷，第10期
2022年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 10, pp3249-3255
October, 2022

分数阶微分预处理及PCA-SRDA的多模型融合对红富士苹果产地溯源

黄 华[1]，南梦迪[1]，李政浩[1]，陈秋颖[1]，李廷杰[1]，郭俊先[2*]

1. 新疆农业大学数理学院，新疆 乌鲁木齐 830052
2. 新疆农业大学机电工程学院，新疆 乌鲁木齐 830052

摘 要 苹果产地溯源具有重要的应用价值和现实意义。为了探寻苹果产地溯源新方法，以红富士品种为研究对象，以新疆阿克苏、山东烟台、陕西洛川三个产地671个红富士苹果样本为试材，分别采集其590～1 250 nm的近红外透射光谱，然后基于分数阶微分(FD)及主成分分析(PCA)-谱回归判别分析(SRDA)进行多模型融合，构建红富士苹果产地溯源的集成学习模型。首先，将经过光谱校正后的光谱数据划分为训练集和测试集，并利用分数阶微分预处理训练集光谱，获取不同阶次(取0～2阶，步长为0.1)的分数阶微分光谱；结合不同阶次的分数阶微分光谱及PCA-SRDA算法构建基学习器，将基学习器预测结果构成一个新训练集，并通过决策树算法完成模型融合，得到最终分类预测模型；随后，采用对应阶次的分数阶微分预处理测试集光谱，并基于已建立的基学习器，获得测试集相应的预测结果；最后，将预测结果构成一个新测试集，并基于已建立的分类预测模型，输出最终的预测结果。按7∶3比例随机划分样本集，并进行200次重复实验。结果表明，结合不同阶次的分数阶微分预处理及线性判别分析(LDA)、SRDA、PCA-LDA、PCA-SRDA算法建立多模型融合集成学习模型，具有较好的鉴别效果和较强的鲁棒性，其中，FD-PCA-SRDA多模型融合集成学习模型为最优，其训练集的平均精度为97.33%，标准差为0.49%，测试集的平均精度为94.84%，标准差为1.48%。故，分数阶微分技术及PCA-SRDA算法结合近红外透射光谱可成功、有效地实现苹果产地溯源。

关键词 近红外透射光谱；分数阶微分；主成分分析-谱回归判别分析；苹果；产地溯源

中图分类号：S661.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)10-3249-07

引 言

我国是世界上最大的苹果栽培国家，也是产量最多的国家。据《2021年中国苹果行业分析报告》显示，2019年—2020年度，我国苹果产量约4 266万吨。在众多苹果品种中，红富士具有体积大，遍体通红，形状圆，果肉紧密，口感甜美、清脆等主要特点，备受消费者的广泛喜爱。红富士苹果多产于山东、甘肃、陕西、山西、河北、辽宁、河南、新疆等地，因受到环境、气候等众多外界因素影响，不同产地的红富士苹果主要成分含量存在差异，苹果的口感、水分、糖分等也存在明显的差别[1]，市场价格也随之不同。因此，从市场和消费者的需求出发，探索一种简单、快捷、无损的识别算法实现苹果产地溯源具有重要的应用价值和现实意义。

近红外光谱技术作为一种高效、快速的现代分析技术，随着计算机技术、光谱技术和化学计量学的不断发展，其以独特的优势在农业、化工、制药、石油等领域得到日益广泛的应用[2-5]。Zhang等运用近红外光谱及偏最小二乘回归对8个苹果品种的可溶性固形物和果实干物质含量进行无损预测[6]；马永杰等基于深度学习数据降维方法，结合近红外透射光谱研究了苹果产地溯源问题[7]。然而，在近红外光谱的实际应用中，采集的近红外光谱原始数据往往是高维的、复杂的，并且存在大量冗余信息和噪声，通常会面临光谱数据预处理、光谱数据的特征提取和特征选择、高精度的回归或判别模型建立等关键核心问题[8-9]。基于此，许多学者进行了广泛的研究。赵启东等将分数阶微分(fractional differential，FD)技术分别与极限学习机、随机森林、多元自适应回归样条函数、弹性网络回归和梯度提升回归树相结合，实现土

收稿日期：2021-08-08，**修订日期**：2021-11-16

基金项目：国家自然科学基金项目(61367001)，新疆维吾尔自治区教育厅面上重点项目(XJEDU2020I009)，新疆维吾尔自治区科技厅面上基金项目(2019D01A52)，2021年度新疆农业大学大学生创新项目资助

作者简介：黄 华，1981年生，新疆农业大学数理学院副教授 e-mail：18899511151@163.com

*通讯作者 e-mail：junxianguo@163.com

壤有机碳含量的估算。结果表明，FD 的预处理效果优于整数阶微分[10]。通常情况下，FD 预处理除了能消除干扰、突出谱线的差别、增强信息量外，还可以挖掘光谱数据的 FD 层面信息，有助于提高模型精度。杨璐等利用红外光谱结合主成分分析(principal component analysis，PCA)-线性判别分析(linear discriminant analysis，LDA)，构建胶料种类判别模型，并对嘉峪关戏台文物胶料种类进行判别，其模型稳定有效[11]；Wu 等利用 PCA 和核 fisher 判别分析结合近红外光谱分析对苹果进行分级[12]；Lv 等提出一种基于近红外光谱和 PLS-DA 鉴别阿克苏红富士苹果品种[13]。PCA 作为一种无监督算法，在实现降维的过程中能尽可能多的保留方差信息，原始数据信息损失较少。LDA 作为一种有监督算法，也可用于数据降维，但是会面临复杂的广义特征分解问题，特别是当样本量和变量指标过大时，LDA 算法的计算量和内存占用较大。Gui 在 LDA 算法的基础上，基于谱图分析理论和回归模型提出了谱回归判别分析(spectral regression discriminant analysis，SRDA)算法，极大地简化了计算过程[14]。在光谱数据建模过程中，除了研究预处理技术、特征提取和特征选择算法外，基于模型融合、集成学习思路建立高精度的回归或判别模型也是建模研究的重点。已有研究预示利用近红外光谱，结合分数阶微分技术及 PCA，LDA 和 SRDA 算法可应用于苹果产地溯源，但是利用近红外透射光谱，基于分数阶微分技术及 PCA-SRDA 进行多模型融合构建集成学习模型，实现红富士苹果产地溯源的研究还未有相关报道。

因此，以红富士品种为研究对象，以新疆阿克苏、山东烟台、陕西洛川三个产地的红富士苹果为试验对象，利用近红外透射光谱，基于分数阶微分技术及 PCA-SRDA 进行多模型融合，构建红富士苹果产地溯源的集成学习模型，以期为苹果产地溯源的实际应用提供新思路。

1 实验部分

1.1 材料

本试验选取三个产地的红富士苹果，包括新疆阿克苏(产地经纬度：80°29′E，41°15′N)、山东烟台(产地经纬度：121°20′E，37°33′N)、陕西洛川(产地经纬度：109°42′E，35°76′N)。苹果试材于 2019 年 1 月 6 日和 10 日分两批购买于新疆乌鲁木齐市北园春水果批发市场。由经验丰富的果商对同批次同品牌苹果拆箱挑选大小适中、尺寸均匀、无明显损伤的苹果，套网套并打包装箱转运回无损检测实验室，然后开箱平铺、室温 20 ℃静置 24 h，擦净苹果表面浮土并逐个编号，共 671 个，其中，新疆阿克苏红富士苹果 241 个，山东烟台红富士苹果 215 个，陕西洛川红富士苹果 215 个。

1.2 设备

近红外透射光谱采集系统由苹果托架、配备小型风扇的光源套件(JCR12V 100 W 卤钨灯)、近红外光谱仪(美国海洋光学公司，USB 2000＋型)、大芯径双包层石英光纤(SMA905 接口)、铝合金机架、暗箱与计算机等组成。光纤探头一端连接光谱仪，另一端固定在苹果托架圆心正下方，实现对近红外透射光谱的高效采集。数据分析用 MATLAB 2019b 软件。

1.3 方法

1.3.1 光谱采集及校正

近红外透射光谱数据采集由配套软件美国海洋光学公司 USB 2000＋型近红外光谱仪实现，使用前开机预热 1 h，之后通过测试采样设置 SpectraSuite 软件界面参数，最后确定样本光谱采集参数为：平均次数 3；平滑度 5；积分时间 120 ms；波段数 512(波长范围：590～1 250 nm)。采集光谱时，将苹果置于光谱采集仪器的果托上，苹果与果托之间不留光缝，确保光纤接收光信号的点完全屏蔽光源，使其只能接收到透过苹果的光。待软件界面显示的光谱稳定后，保存光谱；然后将苹果分别顺时针旋转 120°两次，并分别采集光谱，最后将其平均光谱作为该样本的原始光谱。共采集 671 个苹果的近红外透射光谱信息。同时，为了消除因 USB2000＋光纤光谱仪预热不充分，导致暗光谱发生微小变化所产生的试验误差，每测量 10 个样本就需保存一次该时刻的暗光谱，用于后续光谱校正。

因苹果形状差异、摄像头中的暗电流噪声等会对苹果近红外透射光谱数据产生噪声影响，因此，采集原始光谱后，对获得的近红外透射光谱进行校正[7]。

1.3.2 分数阶微分技术

分数阶微分是由整数阶微分直接扩展而来的，它的定义包括 Cauchy 积分公式、Grünwald-Letnikov 分数阶微积分定义、Riemann-Liouville 分数阶微积分定义和 Capotu 定义等。本研究采用的是 Grünwald-Letnikov 分数阶微积分定义。根据该定义，可利用分数阶微分的数值近似求解公式进行计算，实现光谱数据的分数阶微分预处理[14]。

1.3.3 谱回归判别分析

SRDA 是在 LDA 算法的基础上，基于谱图分析理论和回归模型提出的，可用于有监督、半监督和无监督学习。其主要思想是通过类别标签或无类别标签的数据点来构建连接图，基于连接图可表征数据集内部的判别式结构，还可获得类别标签或无类别标签的数据点的学习响应，根据学习响应，利用回归模型可得嵌入函数，而通过向嵌入函数投影，则可以实现数据的降维。相比 LDA 算法，SRDA 算法在计算量和内存占用方面更具优势。

1.3.4 多模型融合

结合不同阶次的分数阶微分预处理及 PCA-SRDA 进行多模型融合，构建一种集成学习算法。多模型融合的具体流程如图 1 所示。基本思路为：(1)采用不同阶次(取 0～2 阶，步长为 0.1)的分数阶微分预处理训练集原始光谱；(2)基于不同阶次的分数阶微分预处理及 PCA-SRDA 算法构建基学习器，并输出相应的预测结果；(3)将基学习器的预测结果组成一个新训练集，并采用决策树算法完成模型融合，得到最终的分类预测模型；(4)采用对应阶次的分数阶微分技术预处理测试集原始光谱，然后基于已建立的基学习器，输出相应的预测结果；(5)将测试集的基学习器预测结果构成一个新测试集，并基于已建立的分类预测模型，输出最终的预测结果。

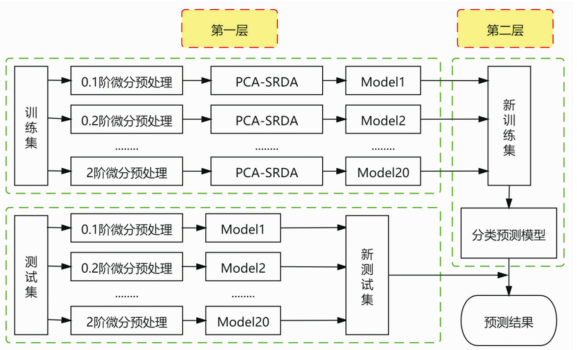


图 1　多模型融合流程图

Fig. 1　The flow chart of mult-model fusion

2　结果与讨论

2.1　不同产地红富士苹果重量、物理特性及可溶性固形物含量分析

对三个产地红富士苹果的重量、物理特性(横径、纵径)和可溶性固形物含量进行常规统计(见表 1)。

从表 1，可以得出，三个产地苹果重量的均值排序为新疆阿克苏＞陕西洛川＞山东烟台；三个产地苹果横径的均值排序为新疆阿克苏＞陕西洛川＞山东烟台；三个产地苹果纵径的均值排序为新疆阿克苏＞山东烟台＞陕西洛川；三个产地苹果可溶性固形物含量的均值排序为新疆阿克苏＞山东烟台＞陕西洛川，且三个产地的苹果可溶性固形物含量具有极其显著的差异。由此可知，新疆阿克苏红富士苹果的糖分明显高于其他两个产地的苹果，市场上也更受消费者喜爱，但仅从苹果重量、外形、物理特性等难以准确判断苹果产地。

2.2　不同产地红富士苹果的平均透射光谱曲线

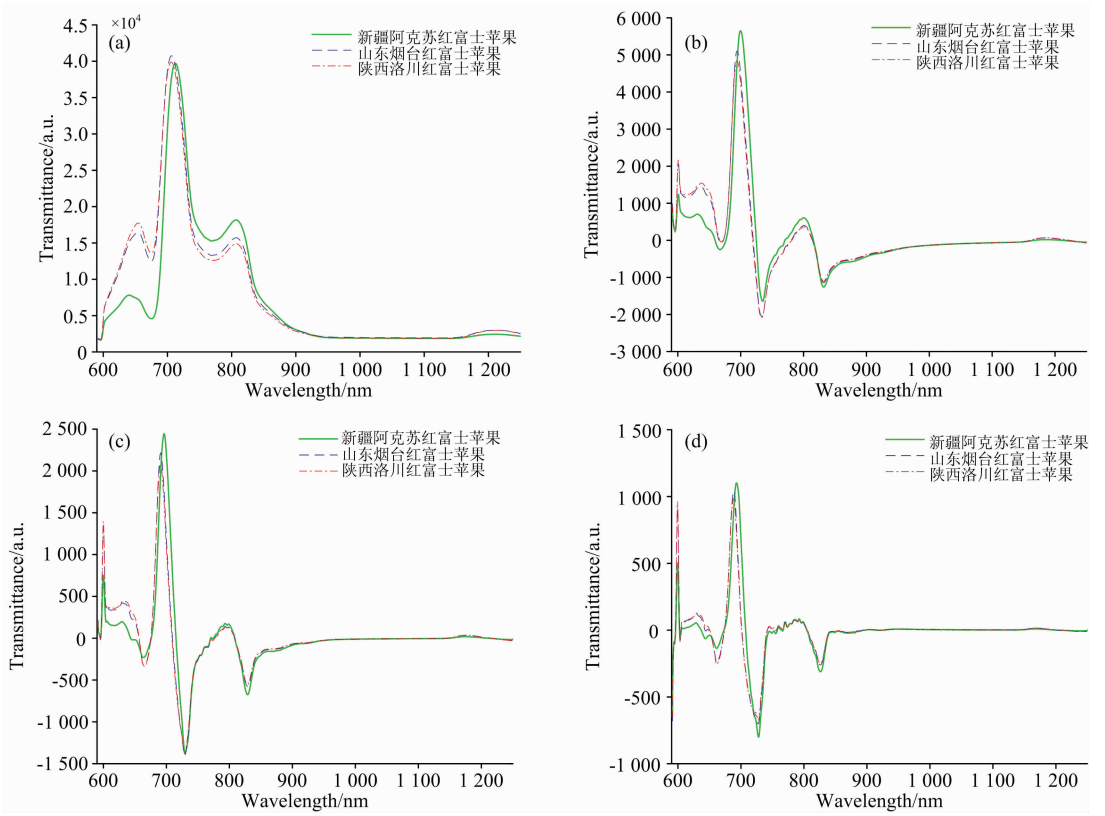
将采集的 671 个苹果的近红外透射光谱，经过光谱校正后，计算每个产地的苹果平均透射光谱曲线，如图 2(a)所示。可见，三条平均光谱曲线的形状、趋势非常一致，但是新疆阿克苏苹果在 600～900 nm 波长范围与其他两个产地的苹果光谱存在差异性分离，山东烟台与陕西洛川苹果之间的光谱吸光度差异较小而难以区分。

进一步，利用 Grünwald-Letnikov 分数阶微积分定义，计算三条平均光谱曲线的分数阶微分，如图 2(b)—(f)所示(仅列出 0.6 阶、0.9 阶、1.2 阶、1.5 阶、2 阶微分曲线)。可

表 1　三个产地红富士苹果的常规数据统计

Table 1　The routine data statistics of Red Fuji apples from three regions

产地	苹果重量/g	苹果横径/mm	苹果纵径/mm	可溶性固形物含量/Brix
新疆阿克苏	273.63±49.31	87.51±5.38	74.51±5.77	15.46±1.27
山东烟台	263.86±42.27	86.19±4.62	74.42±5.03	13.30±1.11
陕西洛川	270.23±37.58	87.40±4.44	74.31±4.91	12.73±1.15



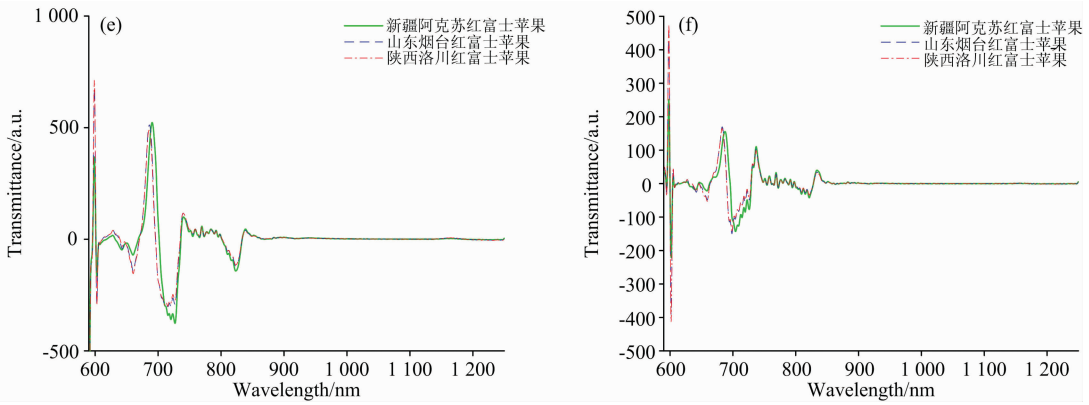


图 2 不同产地红富士苹果的不同分数阶次的近红外透射光谱

(a)：0 阶；(b)：0.6 阶；(c)：0.9 阶；(d)：1.2 阶；(e)：1.5 阶；(f)：2 阶

Fig. 2 The near-infrared transmission spectra of apples from different regions for different fractional orders

(a)：0 order；(b)：0.6 order；(c)：0.9 order；(d)：1.2 order；(e)：1.5 order；(f)：2 order

见，平均光谱曲线基于不同阶次的分数阶微分计算，不同产地的苹果光谱曲线呈现不同的变化。据此，可以从不同分数阶微分层面获取更多光谱数据的深层信息，比如光谱曲线的几何信息，所以利用分数阶微分技术，可挖掘光谱数据的分数阶微分层面的更多数据信息，这势必有助于提高苹果产地溯源模型的精确性和稳健性。

2.3 光谱数据的分数阶微分预处理与 PCA-SRDA 特征提取

光谱数据通过分数阶微分预处理，除了能挖掘分数阶微分层面的光谱曲线信息外，还可以消除干扰、突出谱线的差别、增强信息量。因此，按照多模型融合流程，对原始光谱的训练集和测试集分别进行分数阶微分预处理。如图 3 所示，为训练集和测试集的(0.6 阶)分数阶微分预处理结果。

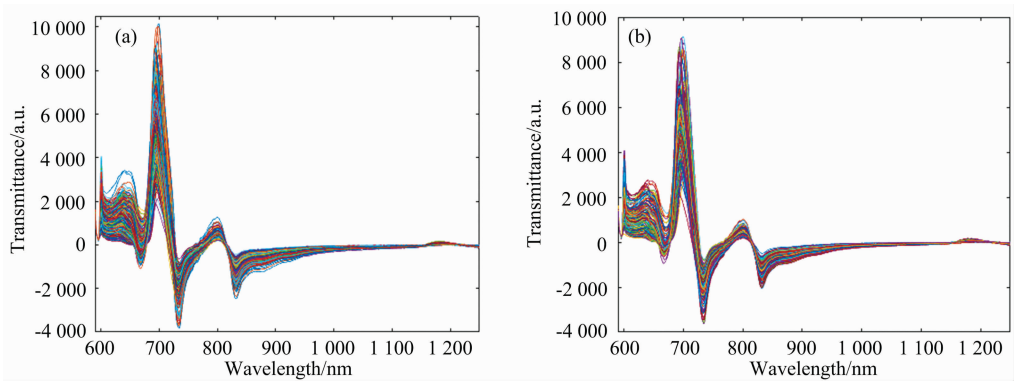


图 3 训练集(a)和测试集(b)的(0.6 阶)分数阶微分预处理结果

Fig. 3 The fractional differential (Step 0.6) preprocessing results of the training (a) and test (b) sets

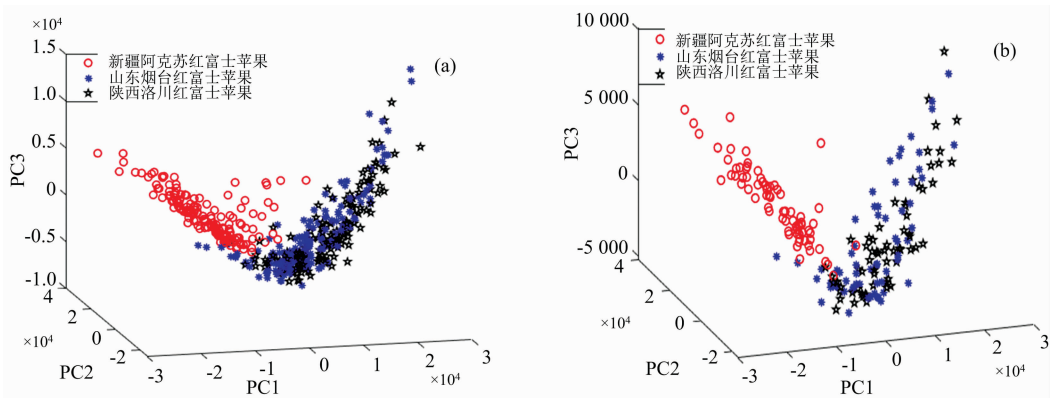


图 4 PCA 降维结果的可视化图

(a)：训练集；(b)：测试集

Fig. 4 The visualization of PCA dimension reduction results

(a)：Training set；(b)：Test set

在获取的分数阶微分光谱基础上，需要进一步降维和特征提取。利用 PCA-SRDA 进行特征提取。如图 4 所示，为训练集和测试集经过(0.6 阶)分数阶微分预处理的 PCA 降维结果(主成分个数取 16)。从图 4 可以看出，通过 PCA 降维处理后，新疆阿克苏苹果与山东烟台、陕西洛川红富士苹果具有较好的区分度，但是山东烟台与陕西洛川苹果之间仍有较多重叠，较为混淆。

如图 5 所示，为训练集和测试集经 PCA 降维后的 SRDA 特征提取结果。从图 5 可以看出，经过 PCA-SRDA 特征提取后，新疆阿克苏、山东烟台、陕西洛川苹果彼此之间的区分度显著提高。由此获悉，采用 PCA-SRDA 算法对本试验的苹果光谱数据进行特征提取是一种切实有效的技术。

2.4　多模型融合算法实现苹果产地溯源

将样本数据按 7∶3 比例随机划分训练集和测试集，然后根据多模型融合步骤予以实现。为比较多模型融合集成学习算法的优劣，同时给出基于 LDA，SRDA 和 PCA-LDA 的预测结果。如表 2 所示，为 200 次重复实验的结果，如图 6 为对应的箱图。

表 2　多模型融合集成学习模型的苹果产地溯源结果(200 次重复实验)

Table 2　The identification results of apple origin based on integrated learning model of multi-model fusion (200 experiments repeated)

模型	训练集精度/%	测试集精度/%
FD-LDA 多模型融合集成学习	86.27±2.09	81.86±3.60
FD-SRDA 多模型融合集成学习	89.37±4.47	82.16±5.48
FD-PCA-LDA 多模型融合集成学习	96.75±0.49	94.33±1.68
FD-PCA-SRDA 多模型融合集成学习	97.33±0.49	94.84±1.48

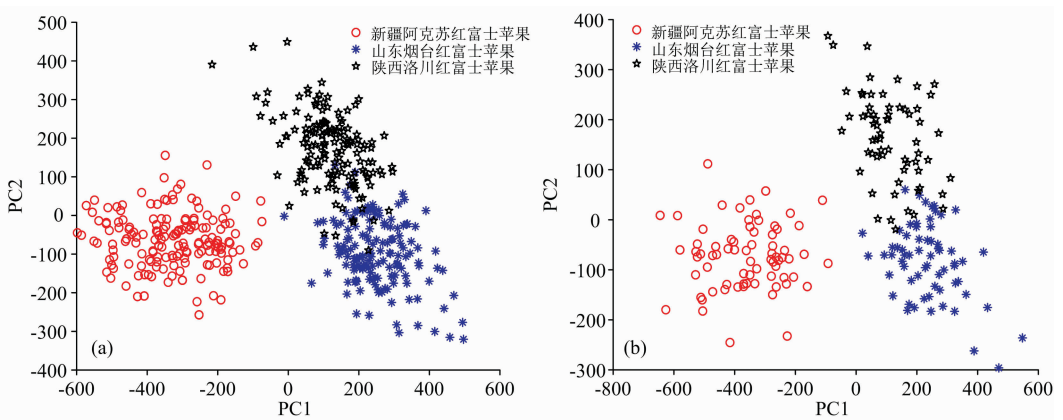


图 5　PCA-SRDA 特征提取结果的可视化图

(a)：训练集；(b)：测试集

Fig. 5　The visualization of PCA-SRDA feature extraction results

(a)：Training set；(b)：Test set

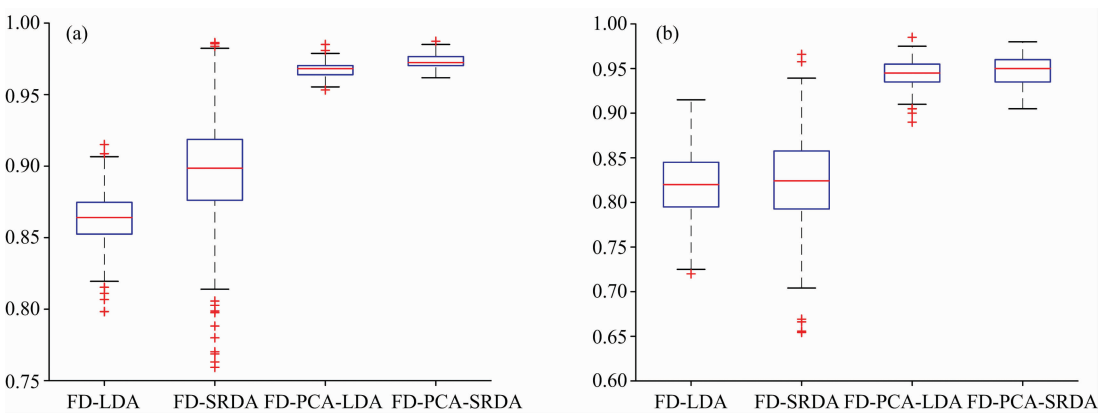


图 6　苹果产地溯源结果的箱图(200 次重复实验)

(a)：训练集；(b)：测试集

Fig. 6　The box diagram of Apple origin identification results (200 experiments repeated)

(a)：Training set；(b)：Test set

从表 3 和图 6 的结果可得，结合不同阶次的分数阶微分预处理及 LDA，SRDA，PCA-LDA 和 PCA-SRDA 算法建立多模型融合集成学习模型，具有较好的鉴别效果和较强的鲁棒性，其中，FD-PCA-SRDA 多模型融合集成学习模型为最

优，其训练集的平均精度为 97.33%，标准差为 0.49%，测试集的平均精度为 94.84%，标准差为 1.48%。此外，FD-PCA-SRDA 多模型融合集成学习模型与 FD-PCA-LDA 多模型融合集成学习模型在精度上没有显著差异，但在模型的运行时间上具有一定差异，单次实验的运行时间平均减少约 3.2 s。综上说明，分数阶微分技术及 PCA-SRDA 多模型融合结合近红外透射光谱技术对苹果产地溯源具有可行性。

3 结 论

结合近红外透射光谱，基于分数阶微分预处理技术及 PCA-SRDA 进行多模型融合构建集成学习模型，实现红富士苹果的产地溯源，得到如下主要结论：

(1)利用分数阶微分预处理光谱数据，除了能消除干扰、突出谱线的差别、增强信息量外，还可以通过计算光谱曲线不同阶次的分数阶微分，挖掘出分数阶微分层面的更多深层数据信息，比如光谱曲线的几何信息，这有助于提高模型的识别精度。

(2)利用 PCA-SRDA 算法对光谱数据进行特征提取，可以很好地将新疆阿克苏、山东烟台、陕西洛川苹果彼此分离开，区分度很好，说明 PCA-SRDA 算法是一种切实有效的特征提取技术。

(3)结合近红外透射光谱，基于分数阶微分技术及 PCA-SRDA 进行多模型融合，构建的苹果识别集成学习模型，取得了预期的识别效果，可成功、有效地实现苹果产地溯源。200 次重复实验结果表明，提出的多模型融合集成学习模型具有较好的鉴别效果和较强的鲁棒性，其中，FD-PCA-SRDA 多模型融合集成学习模型为最优，其训练集的平均精度为 97.33%，标准差为 0.49%，测试集的平均精度为 94.84%，标准差为 1.48%。

(4)本方法具有较强的适用性、较高的识别精度和泛化能力，可为红富士苹果产地溯源提供技术支持和科学支撑，还可以拓展到近红外光谱技术的其他应用领域。

References

[1] JIN Xin-xin, TIAN Ying-zi, YING Li, et al(靳欣欣, 田英姿, 英 犁, 等). Modern Food Science and Technology(现代食品科技), 2016, 32(7): 249.

[2] Jakubíková M, Sádecká J, Kleinová A, et al. Journal of Food Science and Technology, 2016, 53(6): 2797.

[3] Abasi S, Minaei S, Jamshidi B, et al. Scientia Horticulturae, 2019, 252: 7.

[4] Maraa O M, Afseth N K, Knutset S H, et al. Postharvest Biology and Technology, 2021, 180: 111620.

[5] Dong J, Guo W. Food Analytical Methods, 2015, 8(10): 2635.

[6] Zhang Y, Nock J F, Shoffe Y A, et al. Postharvest Biology and Technology, 2019, 151: 111.

[7] MA Yong-jie, GUO Jun-xian, GUO Zhi-ming, et al(马永杰, 郭俊先, 郭志明, 等). Modern Food Science and Technology(现代食品科技), 2020, 36(6): 303.

[8] LIU Yan, CAI Wen-sheng, SHAO Xue-guang(刘 言, 蔡文生, 邵学广). Chinese Science Bulletin(科学通报), 2015, 60(8): 704.

[9] Dombi J, Dineva A. International Journal of Advanced Intelligence Paradigms, 2020, 16(2): 145.

[10] ZHAO Qi-dong, GE Xiang-yu, DING Jian-li, et al(赵启东, 葛翔宇, 丁建丽, 等). Laser & Optoelectronics Progress(激光与光电子学进展), 2020, 57(15): 9.

[11] YANG Lu, HUANG Jian-hua, CHEN Xin-nan, et al(杨 璐, 黄建华, 陈欣楠, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2021, 41(3): 796.

[12] Wu Xiaohong, Wan X, Wu Bin, et al. Advanced Materials Research, 2013, 710: 524.

[13] Lv C, Yang J, Liu Y, et al. IOP Conference Series: Earth and Environmental Science, 2019, 310: 042005.

[14] Gui J, Sun Z N, Cheng J, et al. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 211.

Multi-Model Fusion Based on Fractional Differential Preprocessing and PCA-SRDA for the Origin Traceability of Red Fuji Apples

HUANG Hua[1], NAN Meng-di[1], LI Zheng-hao[1], CHEN Qiu-ying[1], LI Ting-jie[1], GUO Jun-xian[2*]

1. College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi 830052, China

2. College of Mechanical and Electrical Engineering, Xinjiang Agricultural University, Urumqi 830052, China

Abstract Apple's origin traceability has important application value and practical significance. To explore new ways to trace apple's origin, taking 671 Samples of Red Fuji apples from Aksu of Xinjiang Province, Yantai of Shandong Province and Luochuan of Shanxi Province as the research objects. The near-infrared transmission spectra of the samples at 590～1 250 nm are collected respectively, and then the techniques of Fractional Differential (FD) and Principal Component Analysis (PCA)-Spectral Regression Discriminant Analysis (SRDA) are used to fuse multiple models. An integrated learning model of the red Fuji apple's origin traceability is constructed. Firstly, spectral data after spectral correction are divided into a training set and test set, and the fractional-order differential technique is used to preprocess the spectrum of the training set to obtain fractional-order differential spectra of different orders (order 0～2 and step size 0.1 in this paper). A new training set is constructed based on the prediction results of the base learner, built by combining different orders of fractional differential spectra and the PCA-SRDA algorithm, and the final classification prediction model is obtained by fusing the decision tree algorithm. Then, the corresponding order fractional differential is used to preprocess the spectrum of the test set, and the corresponding prediction results are obtained based on the established base learner. Finally, the results are formed into a new test set, and the final prediction results are output based on the established classification prediction model. The sample-set is randomly divided according to the ratio of 7∶3, and the experiment is repeated 200 times. The results show that the multi-model fusion and integration learning model combined with the fractional-order differential preprocessing, Linear Discriminant Analysis (LDA), SRDA, PCA-LDA and PCA-SRDA algorithms has a good Discriminant effect and strong robustness. Among them, The FD-PCA-SRDA multi-model fusion and integration learning model is the best, and the average accuracy and standard deviation of the training set are 97.33% and 0.49%, and the average accuracy and standard deviation of the test set are 94.84% and 1.48%, respectively. Therefore, the fractal-order differential technique and PCA-SRDA algorithm combined with the near-infrared transmission spectrum can successfully and effectively realize apple's origin traceability.

Keywords Near-infrared transmission spectrum; Fractional differential; Principal component analysis-spectral regression discriminant analysis; Apple; Origin traceability

(Received Aug. 8, 2021; accepted Nov. 16, 2021)

* Corresponding author