基于近红外光谱结合化学计量学方法的山里红产地溯源分析

崔　萌[1]，段宝忠[1,3]，程　蕾[1]，张满常[2]，和福美[1,3*]，周　萍[1*]

1. 大理大学药学院，云南 大理　671000
2. 保山市食品药品检验检测中心，云南 保山　678000
3. 云南省中药资源开发利用国际联合实验室，云南 保山　678000

摘　要：目的　基于近红外光谱（near-infrared spectroscopy，NIRS）结合机器学习算法模型，建立山里红 *Crataegus pinnatifida* var. *major* 的产地溯源技术。**方法**　收集 6 个省份的 91 份山里红样本，采集其 NIRS，应用多种机器学习算法，包括主成分分析（principal component analysis，PCA）、正交偏最小二乘法判别分析（orthogonal pmjartial least squares-discriminant analysis，OPLS-DA）、K-最近邻（K-nearest neighbor，KNN）、决策回归树（classification and regression tree，CART）、随机森林（random forest，RF）、朴素贝叶斯（naive bayes，NB）、线性判别分析（linear discriminant analysis，LDA）和神经网络（artificial neural network，ANN）算法，探讨适合山里红产地溯源的模型。**结果**　ANN 模型的准确率和模型稳定性最优，可作为山里红产地识别的首选模型。**结论**　NIRS 结合 ANN 模型是山里红产地溯源的有效手段，为山里红的产地溯源提供了科学参考。

关键词：山里红；近红外光谱技术；化学计量学；神经网络；产地溯源

中图分类号：R286.2　　**文献标志码：**A　　**文章编号：**0253 - 2670(2024)14 - 4897 - 10

DOI: 10.7501/j.issn.0253-2670.2024.14.025

Geographical origin traceability of *Crataegus pinnatifida* var. *major* based on near-infrared spectroscopy combined with chemometrics

CUI Meng[1], DUAN Baozhong[1,3], CHENG Lei[1], ZHANG Manchang[2], HE Fumei[1,3], ZHOU Ping[1]

1.College of Pharmacy, Dali University, Dali 671000, China
2. Baoshan Food and Drug Inspection and Testing Center, Baoshan 678000, China
3. International Joint Laboratory for Development and Utilization of Traditional Chinese Medicine Resources in Yunnan Province, Baoshan 678000, China

Abstract: Objective　To develop a traceability technology system for determining the origin of *Crataegus pinnatifida* var. *major* through the integration of near-infrared spectroscopy (NIRS) technology and machine learning algorithms. **Methods**　A total of 91 samples of *C. pinnatifida* var. *major* were collected from six provinces in China, and their NIRS were acquired. Various machine learning algorithms, including principal component analysis (PCA), orthogonal partial least squares-discriminant analysis (OPLS-DA), k-nearest neighbor (KNN), classification and regression tree (CART), random forest (RF), naive bayes (NB), linear discriminant analysis (LDA) and artificial neural network (ANN), were employed to establish a model for the purpose of origin tracing. **Results**　Among the different algorithms tested, the ANN model demonstrated the highest accuracy and stability in identifying the origin of *C. pinnatifida* var. *major*, making it a reliable tool for traceability. **Conclusion**　The combination of NIRS technology and the ANN model can be used as an effective approach for tracing the geographical origin of *C. pinnatifida* var. *major*. This study contributes to the establishment of a scientifically rigorous foundation for the geographical origin tracing of *C. pinnatifida* var. *major*.

Key words: *Crataegus pinnatifida* Bge. var. *major* N. E. Br.; near-infrared spectroscopy; chemometrics; artificial neural network; origin traceability

收稿日期：2024-01-09

基金项目：云南省生物医药重大专项（202002AA100007）；云南省科技计划（202205AF150026）；云南省兴滇英才支持计划（YNWR-QNBJ-2020251）

作者简介：崔　萌（1999—），硕士研究生，研究方向为中药资源与鉴定。E-mail: 951591478@qq.com

***通信作者：**和福美，讲师，研究方向为中药药理药效，神经损伤性疾病的治疗。Tel: 15927504022　E-mail: hefumei2023@126.com

周　萍，教授，研究方向为中药品质评价。Tel: 13708668136　E-mail: zhouping725@126.com

山里红 *Crataegus pinnatifida* Bge. var. *major* N. E. Br.为蔷薇科山楂属植物，为山楂的主流栽培品种，其干燥果实具有消食健胃、行气散瘀、化浊降脂之功效[1-2]。山里红是重要的药食同源资源，广泛分布于河北、山东、辽宁等地，尤以山东为道地产区。大量研究表明，产地是影响药材次生代谢产物合成的重要因素，进而导致药效和临床应用存在差异[3-4]。因此，确定山里红的产地，是确保临床用药稳定和有效的基础。然而，目前相关研究不足，亟需建立山里红药材产地溯源方法。

近红外光谱（near-infrared spectroscopy，NIRS）是一种利用近红外区电磁波对物质中的化学键振动吸收信息，进行定性、定量检测的技术，其联合化学计量学方法可预测光谱与待测成分含量之间的关系。该技术具有无损、高效和操作简便等特点[5]，已广泛应用于多种原料，如枸杞[6]、天麻[7]、茯苓[8]等的产地识别。目前已有山里红品质评价的研究报道[9]，但尚未见山里红产地识别方面的研究。鉴于此，本研究采用 NIRS 技术，结合多种化学计量学方法，对 6 个省区的 91 批山里红样本进行检测，建立了山里红药材的 NIRS 图谱，并将多种化学计量学模型用于 NIRS 的分析，以期建立准确、可靠的山里红产地识别方法，为山里红药材的质量评价提供科学参考。

1 仪器与材料

1.1 仪器

Bruker MATRIX-F 型近红外光谱仪（德国布鲁克公司），800 A 型中草药粉碎机（浙江省永康市红太阳机电有限公司），DHG-9140A 型鼓风干燥箱（上海恒科学仪器有限公司），AX224ZH/E 型电子天平（奥豪斯仪器常州有限公司），称量瓶（北京万晶博美玻璃制品有限公司），60 目筛（浙江省上虞市五四纱厂）。

1.2 材料

91 份山里红新鲜果实，由经销商收集于新疆、山西、河北、河南、山东、辽宁等地，每份样品 1.0 kg，经自来水冲洗干净，切片后晒干，粉碎，过 60 目筛，存于干燥器中备用。经大理大学段宝忠教授鉴定为蔷薇科山楂属植物山里红 *C. pinnatifida* Bge. var. *major* N. E. Br 的果实，凭证标本保存于大理大学中药标本馆。样本地理信息详见表 1。

2 方法

2.1 NIRS 数据采集

所有样本的 NIRS 检测波数范围为 4 000～12 000 cm^{-1}，扫描次数 32 次，分辨率为 8 cm^{-1}；仪器控制、光谱参数和数据采集均使用 OPUS 7.8 软件进行。采集样本前，以仪器内置背景为参比，以消除仪器和环境因素干扰。每批样本重复测试 3 次，取平均值用于分析，以减小数据偏差。

2.2 NIRS 数据预处理

由于样本的原始光谱包含成分和仪器干扰信息，会影响模型的稳定性和可靠性。在进行模型建立之前，需要对采集的原始光谱进行预处理，以消除干扰信息的影响[10-11]。本研究样本的 NIRS 原始数据使用滤波平滑（savitsky-golay smoothing，S-G）、矢量归一化（standard normal variate，SNV）和一阶导数（first derivative，1st derive）进行预处理。

2.3 数据分析

NIRS 图谱采用 Origin 2018（Origin Lab，美国）绘制。主成分分析（principal component analysis，PCA）、正交偏最小二乘法判别分析（orthogonal partial least squares-discriminant analysis，OPLS-DA）采用 SIMCA 13.0 分析。线性判别分析（linear discriminant analysis，LDA）和神经网络（artificial neural network，ANN）采用 SPSS 26.0 分析，K-最近邻（K-nearest neighbor，KNN）、决策回归树（classification and regression tree，CART）、随机森林（random forest，RF）、朴素贝叶斯（naive bayes，NB）通过 SPSS 网络在线版 https://spssau.com/indexs.html 进行。

3 结果与分析

3.1 红外图谱共有特征峰解析及表征

样本原始叠加 NIRS 图（图 1-A）显示，在 4 000～12 000 cm^{-1} 内存在多个吸收带光谱，其中最显著的吸收带与 C-H、N-H 和 O-H 官能团的基频、倍频和合频振动信息相关[12]。值得注意是，不同产地的山里红原始 NIRS 吸收峰差异较小，根据原始光谱图难以区分产地，这可能由于仪器背景噪声和光散射效应等信息干扰所致[13-14]。

为提高光谱的专属性，本研究采用 SNV，S-G 和 1st derive 算法对光谱进行预处理，以去除干扰信息并提高重叠带的分辨率。样本预处理图如图 1-B 所示，其中在 7 200～4 000 cm^{-1} 内有 5 个明显的吸收峰。7 200～6 000 cm^{-1} 的吸收峰与 O-H 基团的二倍频振动有关，可能为糖、酸或黄酮类等成分的吸收峰[15]；6 000～5 500 cm^{-1} 的吸收峰与 C-H 键伸缩振动相关，可能为黄酮类化合物的吸收峰[16]；

表 1 样本信息

Table 1 Geographic information of samples

编号	地区	编号	地区	编号	地区
S1	河北石家庄	S32	辽宁沈阳	S63	山西运城
S2	河北石家庄	S33	辽宁沈阳	S64	山西运城
S3	河北石家庄	S34	辽宁沈阳	S65	山西运城
S4	河北石家庄	S35	辽宁丹东	S66	山西运城
S5	河北石家庄	S36	辽宁丹东	S67	山西运城
S6	河北石家庄	S37	辽宁丹东	S68	山西运城
S7	河北石家庄	S38	辽宁葫芦岛	S69	山西运城
S8	河北石家庄	S39	辽宁大连	S70	山西运城
S9	河北石家庄	S40	辽宁铁岭	S71	山西运城
S10	河北承德	S41	山东临沂	S72	山西运城
S11	河北承德	S42	山东临沂	S73	山西运城
S12	河北承德	S43	山东临沂	S74	山西运城
S13	河北承德	S44	山东临沂	S75	山西运城
S14	河北秦皇岛	S45	山东临沂	S76	山西运城
S15	河北秦皇岛	S46	山东临沂	S77	山西运城
S16	河北秦皇岛	S47	山东临沂	S78	山西吕梁
S17	河北唐山	S48	山东临沂	S79	山西吕梁
S18	河北唐山	S49	山东临沂	S80	山西吕梁
S19	河北保定	S50	山东临沂	S81	河南济源
S20	河北保定	S51	山东临沂	S82	河南济源
S21	辽宁鞍山	S52	山东临沂	S83	河南南阳
S22	辽宁鞍山	S53	山东临沂	S84	河南南阳
S23	辽宁鞍山	S54	山东潍坊	S85	河南南阳
S24	辽宁鞍山	S55	山东潍坊	S86	新疆维吾尔自治区阿勒泰地区
S25	辽宁鞍山	S56	山东潍坊	S87	新疆维吾尔自治区阿勒泰地区
S26	辽宁鞍山	S57	山东济宁	S88	新疆维吾尔自治区阿勒泰地区
S27	辽宁锦州	S58	山东日照	S89	新疆维吾尔自治区阿克苏地区
S28	辽宁锦州	S59	山东泰安	S90	新疆维吾尔自治区阿克苏地区
S29	辽宁锦州	S60	山东枣庄	S91	新疆维吾尔自治区阿克苏地区
S30	辽宁锦州	S61	山西运城		
S31	辽宁沈阳	S62	山西运城		

5 500～5 000 cm^{-1} 的吸收峰与 N-H 键、C-H 键、O-H 键和 C＝O 键的伸缩振动有关，可能为糖或蛋白质等化合物的吸收峰[17]；5 000～4 500 cm^{-1} 的 2 个明显吸收峰与 C-H 和 C=O 键的拉伸振动相关，可能为类黄酮的吸收峰[15]。4 500～4 000 cm^{-1} 的 2 个显著吸收峰与 C-H 键的伸缩和弯曲振动，以及-CH_2 的变形振动相关，可能为蛋白质、多糖和淀粉等物质的吸收峰[18]。总之，不同产地的山里红 NIRS 吸收峰高度相似，但吸收峰强度略有差异，这可能与不同产地山里红样本的化学成分含量差异有关[19]。然而，仅通过比较共有特征峰、特征波段以及全谱无法有效识别山里红的产地。

为了能够对山里红进行产地溯源，本研究基于近 NIRS 结合多种机器学习算法模型，通过比较不同算法的识别准确率，探讨适合山里红产地溯源的模型。特别地，在机器学习算法中，将数据集划分为测试集和训练集是评估模型性能和泛化能力的关键步骤，通常需要制定一个比例，但由于不同算法对训练集和测试集大小的敏感性不同，其中，7∶3 和 8∶2 是较为常见的划分方式。因此，选择不同的划分比例，可以更全面地评估模型性能。

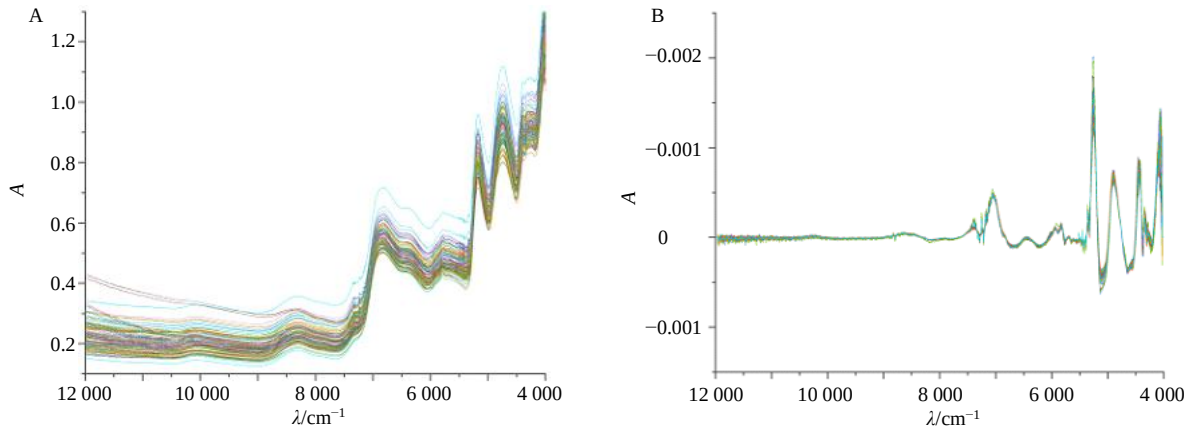

图 1　山里红原始 NIRS 图（A）和 S-G、SNV 和 1st derive 预处理光谱图（B）

Fig. 1　Original NIRS (A) and S-G, SNV, and 1st derive pre-treatment spectra (B) of *C. pinnatifida* var. *major*

3.2　基于 PCA 的产地溯源

PCA 是一种广泛应用的无监督学习算法，其可将原始变量转换为一组新的不相关变量，即主成分（principal components，PCs），从而降低数据的维度和复杂。该方法是通过线性变换将原始数据映射到新的坐标系中实现降维。然而，随着维度的减少，部分原始数据信息可能会丢失[20]。根据 PC1 和 PC2 生成的散点图（图 2-A），可见 91 批山里红样本的 PC1 和 PC2 方差贡献率为 43.6%，但判别率相对较低。具体而言，河南地区的样本和大部分山西地区的样本，在 PC1 和 PC2 上均显示为负值，而所有新疆的样本 PC1 值为负，部分 PC2 值为正。此外，大部分山东地区的样本和少数辽宁地区的样本 PC2 显示为负值。综合来看，山西地区与河南地区的样本

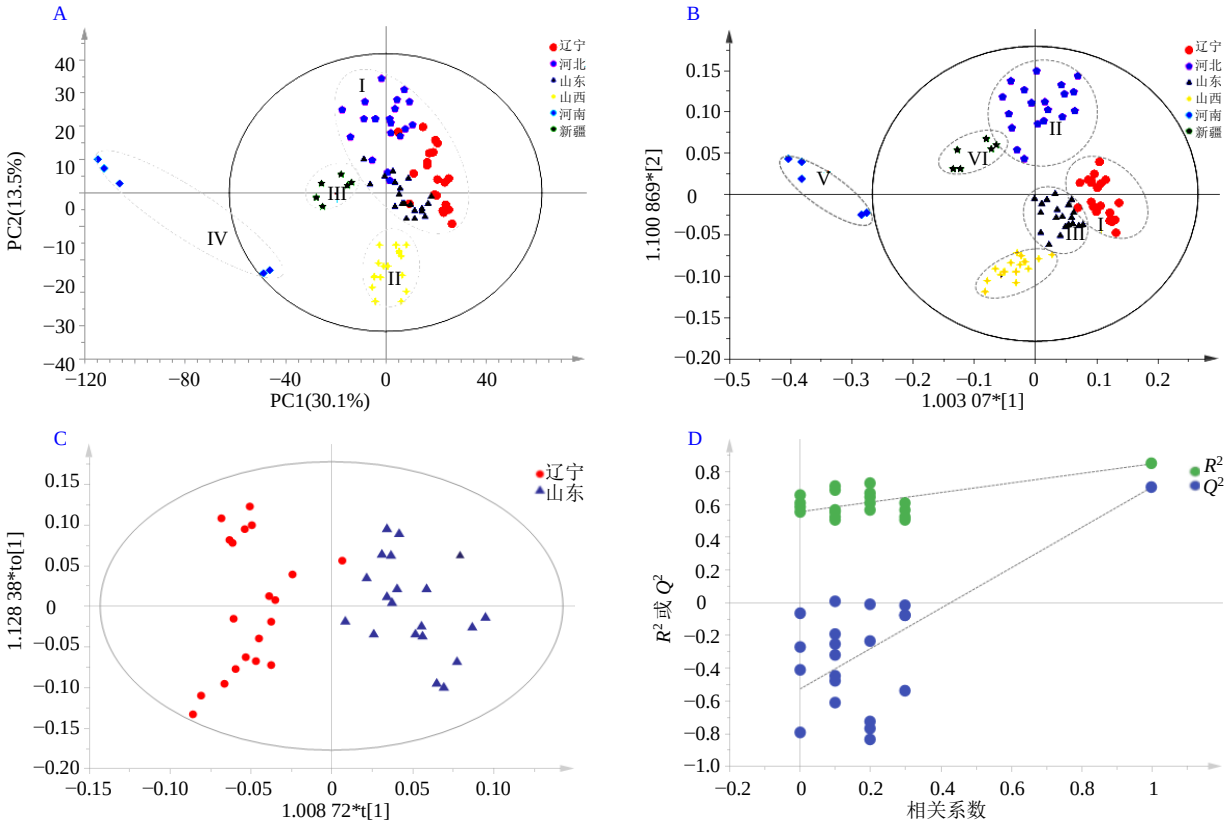

图 2　PCA 评分图（A）、91 批样品 OPLS-DA 评分图（B）、2 批样品 OPLS-DA 评分图（C）和 OPLS-DA 验证图（D）

Fig. 2　PCA score plot (A), OPLS-DA score plot of 91 batches of samples (B), OPLS-DA score plot of two batches of samples (C), OPLS-DA validation plot (D)

与其他产地的样本在主成分空间中有显著分离。相反，河北、辽宁和山东地区的样本（Group I）呈现出一定的聚集趋势。上述结果表明 PCA 能够有效区分新疆、山西和河南的样本明显区分，但难以将河北、辽宁和山东的样本区分开来。提示 PCA 不适合用于上述产地山里红样本的产地识别。

3.3 基于 OPLS-DA 的产地溯源

OPLS-DA 是一种有监督学习算法，通过结合代谢物表达量和样本类别来实现高效的数据分类[21]。相较于其他方法，OPLS-DA 可识别光谱中的非相关变异性，从而提高建模准确性。然而，在样本量较少的情况下，该模型可能出现对数据的过度拟合现象[22]。对 91 批山里红样本分析结果如图 2-B 所示，不同组别之间的分类较 PCA 更加明显，然而，辽宁地区（Group I）和山东地区（Group III）的样本在边缘处有轻微的重叠。评价参数 R^2_X=0.720、R^2_Y=0.664 和 Q^2=0.531 表明模型的拟合能力和预测能力略显不足。值得注意的是，当将 2 个地区的样本进行单独分析，这 2 个地区的样本分布较为分散（图 2-C），且左侧 R^2 和 Q^2 值均较右侧较低（图 2-D），Q^2 回归线的截距为负值，表明 OPLS-DA 不存在过拟合现象，并表现出一定的分类效果，其 Q^2 值超过了 0.5（0.706），提示该模型的相对稳定性[23-24]。综上，尽管 OPLS-DA 分析提供了聚类和距离趋势的信息，但仍无法对不同产地的山里红样品进行准确分类。

3.4 基于 KNN 分类算法的产地溯源

KNN 分类算法是一种惰性机器学习算法，具备增量学习特性，适用于非线性分类问题。该算法通过计算未知样本与训练集中样本的距离，并结合最近邻的类别信息，对未知样本进行分类。然而，在处理具有大量特征或属性的高维数据集时，KNN 性能可能下降[25]。如表 2 所示，72 个样本和 19 个样本分别被划分为训练集和测试集。在训练集数据中，除辽宁和山东地区的样本精确率、召回率、fl-score 未达到 100.0%，而其他地区的样本在这些指标上均达到了 100.0%，表明 KNN 在训练集上可以准确识别来自其他地区的样本。从测试数据集看，除了辽宁、山东和河南地区以外，其他地区的测试指标均达到 100%，表明 KNN 算法在山里红产地的识别方面具有一定识别能力，但对于辽宁和山东的样本分类效果较差，此外，河南地区样本所有性能指标均为 0，说明 KNN 算法无法将河南产地与其他地区的样品区分。

表 2 KNN 算法的分类结果

Table 2 Identification effects of KNN algorithm

样本	产地	样本数	精确率/%	召回率/%	fl-scroe/%
训练	辽宁	18	94.0	94.0	94.0
	河北	16	100.0	100.0	100.0
	山东	14	93.0	93.0	93.0
	山西	16	100.0	100.0	100.0
	河南	5	100.0	100.0	100.0
	新疆	3	100.0	100.0	100.0
	平均值		97.8	97.8	97.8
测试	辽宁	2	67.0	100.0	80.0
	河北	4	100.0	100.0	100.0
	山东	6	100.0	83.0	91.0
	山西	4	100.0	100.0	100.0
	河南	0	0.0	0.0	0.0
	新疆	3	100.0	100.0	100.0
	平均值		77.8	80.5	78.5

3.5 基于 CART 算法的产地溯源

CART 分析是一种基于决策树概念的机器学习算法。其通过选择最佳切分点来将数据分为更纯净的子集，然后在随机选择的子集上训练多个决策树，提高模型的稳健性和泛化能力[26]。该方法具备较强的数据解释性和非线性关系的建模能力，但对噪声敏感，容易出现过拟合现象。通过将参数设置为训练集比例 0.7，节点分裂标准为 Gini，节点划分方式为最佳切分点，不限制树深度。分析结果见表 3，63 个样本和 28 个样本分别被划分为训练集和测试集，在训练集方面，所有样本的精确率、召回率以及 fl-scroe 均达到 100.0%，表明该算法在训练集上相对

表 3 CART 算法的分类结果

Table 3 Identification effects of CART algorithm

样本	产地	样本数	精确率/%	召回率/%	fl-scroe/%
训练	辽宁	17	100.0	100.0	100.0
	河北	13	100.0	100.0	100.0
	山东	12	100.0	100.0	100.0
	山西	15	100.0	100.0	100.0
	河南	4	100.0	100.0	100.0
	新疆	2	100.0	100.0	100.0
	平均值		100.0	100.0	100.0
测试	辽宁	3	33.0	67.0	44.0
	河北	7	100.0	100.0	100.0
	山东	8	80.0	50.0	62.0
	山西	5	100.0	100.0	100.0
	河南	1	33.0	100.0	50.0
	新疆	4	100.0	50.0	67.0
	平均值		74.3	77.8	70.5

稳定。然而，在测试集方面，不同地区的样本分类结果存在差异。特别是，辽宁地区的样本的精确率仅达到 33.0%，召回率为 67.0%，fl-scroe 为 44.0%，提示 CART 算法在特定地区的应用需要进一步改进以提高精确率。总之，虽然训练集中的所有样本都被正确分类，但测试集的分类结果是评估模型在未知数据上可行性的重要参数[27]。因此，CART 算法在山里红产地的识别应用上存在一定局限。

3.6 基于 RF 算法的产地溯源分析

RF 是一种集成算法，通过构建多个决策树并采用投票或平均值的方式来提高分类的准确性。相较于单独的决策树，RF 在处理大规模高维数据集时表现出色，能够有效地避免数据过拟合问题[28]。然而，在处理少量数据集时，RF 的性能通常较差。通过将参数设置为训练集比例 0.7，节点分裂标准为 Gini，不限制树深度。结果见表 4，63 个样本和 28 个样本分别被划分为训练集和测试集。训练集中所有样本的各项指标均达到了 100.0%，这表明 RF 在训练数据上表现良好。然而，在测试集中，除辽宁地区的精确率为 75.0%，其他所有地区均为 100.0%。召回率方面，除山东地区的为 88.0%，其他地区均为 100.0%，相对于其他省份较低，提示需要进一步优化参数或者考虑使用更适合这些数据特征的替代模型。

表 4 RF 算法的分类结果

Table 4 Identification effects of RF algorithm

样本	产地	样本数	精确率/%	召回率/%	fl-scroe/%
训练	辽宁	17	100.0	100.0	100.0
	河北	13	100.0	100.0	100.0
	山东	12	100.0	100.0	100.0
	山西	15	100.0	100.0	100.0
	河南	4	100.0	100.0	100.0
	新疆	2	100.0	100.0	100.0
	平均值		100.0	100.0	100.0
测试	辽宁	3	75.0	100.0	86.0
	河北	7	100.0	100.0	100.0
	山东	8	100.0	88.0	93.0
	山西	5	100.0	100.0	100.0
	河南	1	100.0	100.0	100.0
	新疆	4	100.0	100.0	100.0
	平均值		95.8	98.0	96.5

3.7 基于 NB 算法的产地溯源分析

NB 算法是一种基于概率统计的分类算法，用于数据分类和任务预测，并且生成的模型具有较好的解释性。该算法基于贝叶斯定理以及特征条件独立性假设，通过计算在给定类别情况下特征的条件概率来进行分类。然而，当数据量较少时，计算的概率可能不准确，影响其分类性能[29]。参数设置为高斯分布及训练集比例 0.8。根据表 5 的数据，72 个样本和 19 个样本分别被划分为训练集和测试集，在训练集中，辽宁和山东地区的 fl-scroe 为 97.0%，其他地区均达到了 100.0%，表明 NB 算法在大多数地区表现较优。从测试集数据看，河南和新疆的样本各指标均为 0，表明模型在这些地区的识别效果较差。

表 5 NB 算法的分类结果

Table 5 Identification effects of NB algorithm

样本	产地	样本数	精确率/%	召回率/%	fl-scroe/%
训练	辽宁	18	100.0	94.0	97.0
	河北	16	100.0	100.0	100.0
	山东	14	93.0	100.0	97.0
	山西	16	100.0	100.0	100.0
	河南	5	100.0	100.0	100.0
	新疆	3	100.0	100.0	100.0
	平均值		98.8	99.0	99.0
测试	辽宁	2	100.0	100.0	100.0
	河北	4	57.0	100.0	73.0
	山东	6	100.0	100.0	100.0
	山西	4	100.0	100.0	100.0
	河南	0	0.0	0.0	0.0
	新疆	3	0.0	0.0	0.0
	平均值		59.5	66.7	62.2

3.8 基于 LDA 算法的产地溯源分析

LDA 是一种用于降维和分类的机器学习算法，其通过线性组合特征以增大不同类别之间的距离，同时最小化类别之间的差异[30]。在处理具有明显类别分布且线性可分的问题，LDA 表现较为出色，具有显著的降维效果。然而，在解决非线性问题时，LDA 的性能通常较差。LDA 的分类结果如表 6 所示，32 个样本和 59 个样本分别被划分为训练集和测试集，在训练集中，不同产地的山里红样本的分类准确率为 100.0%。在测试集中，分类准确率降至 60.5%。20 个辽宁地区的样本中有 11 个被正确分类，河北地区的样本中有 1 个样本被误分类为新疆，山东地区的样本中仅有 9 个被正确分类，山西地区的样本分别有 2 个和 4 个样本被误分类为辽宁和山东，河南和新疆地区的样本均被正确分类。此外，LDA 图见图 3-A，可见 DF1 和 DF2 方差贡献率为 88.4%。来自河南地区的样本与其他 5 个产地的样本明显分离；辽宁、山东、山西聚集为一簇，新疆和河北的样本聚集为一簇。上述结果显示 LDA 算法不适用于山里红样本的产地识别。

3.9 基于 ANN 算法的产地溯源分析

ANN 是光谱数据分析的有效工具，由多个神经元层组成，用于处理复杂的非线性关系，具备计算、非线性拟合、自学习和容错特性。该方法可通过学习数据中的模式和特征来调整网络中的连接权重，以实现对复杂问题的建模和预测，因此对大规模数据集能进行有效的建模学习，但当数据量较少时会出现过度拟合的现象[31]。如表 7 所示，66 个样本和 25 个样本分别被划分为训练集和测试集，在训练集数据中，所有山里红样本均被成功分类，准确率为 100.0%。在检验集中，6 个省区的分类准确率也达 100.0%，表明 ANN 模型在山里红的产地识别方面较优，具有较高的准确性。NIRS 的变量重要性分析柱状图显示（图 3-B），排名前 7 位的正态化重

表 6　LDA 的分类结果

Table 6　Identification effects of LDA algorithm

样本	产地	辽宁	河北	山东	山西	河南	新疆	准确率/%
训练	辽宁	9	0	0	0	0	0	100.0
	河北	0	1	0	0	0	0	100.0
	山东	0	0	11	0	0	0	100.0
	山西	0	0	0	6	0	0	100.0
	河南	0	0	0	0	2	0	100.0
	新疆	0	0	0	0	0	3	100.0
	平均值							100.0
测试	辽宁	3	0	5	4	0	0	27.0
	河北	0	19	0	0	0	0	100.0
	山东	6	0	0	5	0	0	0.0
	山西	2	0	4	5	0	0	36.0
	河南	0	0	0	0	3	0	100.0
	新疆	0	0	0	0	0	3	100.0
	平均值							60.5

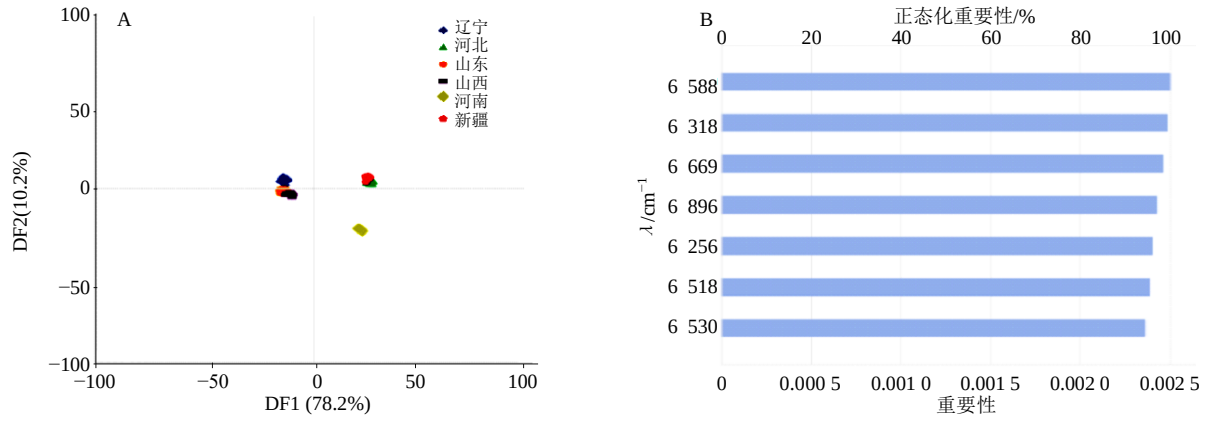

图 3　NIRS 图谱的 LDA 分析 (A) 和变量重要性分析柱状图 (B)

Fig. 3　LDA of NIRS (A) and bar chart of variable importance analysis (B)

表 7　ANN 的分类结果

Table 7　Identification effects of ANN algorithm

样本	实测	辽宁	河北	山东	山西	河南	新疆	准确率/%
训练	辽宁	13	0	0	0	0	0	100.0
	河北	0	12	0	0	0	0	100.0
	山东	0	0	18	0	0	0	100.0
	山西	0	0	0	15	0	0	100.0
	河南	0	0	0	0	4	0	100.0
	新疆	0	0	0	0	0	4	100.0
	平均值							100.0
测试	辽宁	7	0	0	0	0	0	100.0
	河北	0	8	0	0	0	0	100.0
	山东	0	0	2	0	0	0	100.0
	山西	0	0	0	5	0	0	100.0
	河南	0	0	0	0	1	0	100.0
	新疆	0	0	0	0	0	2	100.0
	平均值							100.0

要性波长分别为 6 588、6 318、6 669、6 896、6 256、6 518、6 530 cm^{-1}，这些波长为 6 256～6 896 cm^{-1}，之前的研究提示这些波段范围为糖、酸或黄酮类化合物[16]；大量的研究已经证实山里红的主要含有多糖、有机酸及黄酮类成分，是其发挥药效的物质基础[32-33]。本研究中，所筛选出的特征波段与这些成分相符，提示多糖、有机酸和黄酮类化合物在山里红产地识别中具有重要作用。

综上，6 种算法的识别准确率依次为 ANN（100.0%）＞RF（96.5%）＞KNN（78.5%）＞CART（70.5%）＞NB（62.2%）＞LDA（60.5%），计算运行时间在 15.6～70.5 s，其中运算最快的为 ANN（15.6 s），最慢的为 KNN（70.5 s）。

4 讨论

本研究对全国 6 个省份的 91 批山里红材料，采用 NIRS 结合 PCA、OPLS-DA、KNN、CART、RF、NB、LDA 和 ANN 等多种算法对山里红的产地溯源进行了探讨。研究表明，基于运行时间、准确率及模型稳定性综合考虑，ANN 算法是山里红产地识别的最优方法。在训练集和预测集上的识别率均达到了 100.0%，为山里红产地识别提供了科学参考。此外，ANN 算法所筛选出的特征波段与山里红所含的主要成分多糖、有机酸和黄酮类化合物一致，提示这些成分在山里红产地识别中具有重要作用，在将来的山里红质量评价中应重点关注这些成分。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 中国药典 [S]. 一部. 2020: 253.

[2] Cui M, Cheng L, Zhou Z Y, *et al*. Traditional uses, phytochemistry, pharmacology, and safety concerns of hawthorn (*Crataegus* genus): A comprehensive review [J]. *J Ethnopharmacol*, 2024, 319(Pt 2): 117229.

[3] 杨钰颖, 苗水, 李雯婷, 等. 植物代谢组学在根及根茎类中药材中的研究进展 [J]. 中草药, 2023, 54(20): 6856-6865.

[4] 王梦迪, 雍旭红, 印敏, 等. 代谢组学技术在植物次生代谢调控研究中的应用 [J]. 植物科学学报, 2023, 41(2): 269-278.

[5] 黄志伟, 郭拓, 黄文静, 等. 近红外光谱技术在名贵中药材质量评价中的研究进展 [J]. 中草药, 2022, 53(20): 6328-6336.

[6] 王磊, 覃鸿, 李静, 等. 近红外高光谱图像的宁夏枸杞产地鉴别 [J]. 光谱学与光谱分析, 2020, 40(4): 1270-1275.

[7] 白庆旭, 候英, 杨盼盼, 等. 基于近红外光谱技术的天麻产地鉴别方法 [J]. 西部林业科学, 2021, 50(3): 124-130.

[8] 李嘉仪, 余梅, 郑郁, 等. 基于近红外光谱技术的不同产地茯苓块无损鉴别 [J]. 分析试验室, 2021, 40(12): 1381-1386.

[9] 崔洁, 刘心悦, 常冠华, 等. 基于 HPLC 和层次分析法评价不同品种山楂综合品质 [J]. 辽宁中医药大学学报, 2020, 22(9): 41-44.

[10] 李文龙, 徐金钟, 刘绍勇, 等. 近红外漫反射光谱法测定 5 种中药提取物中水分含量的研究 [J]. 药物分析杂志, 2009, 29(10): 1602-1606.

[11] 蒲婷婷, 刘杰, 周忠瑜, 等. 基于化学成分表征的附子饮片“白缓黑急”合理性研究 [J]. 中草药, 2022, 53(22): 1-8.

[12] Kumar Y, Chandrakant Karne S. Spectral analysis: A rapid tool for species detection in meat products [J]. *Trends Food Sci Technol*, 2017, 62: 59-67.

[13] Chen J, Liu H G, Li J Q, *et al*. A rapid and effective method for species identification of edible boletes: FT-NIR spectroscopy combined with ResNet [J]. *J Food Compos Anal*, 2022, 112: 104698.

[14] Peng Q, Chen J L, Meng K, *et al*. Rapid detection of adulteration of glutinous rice as raw material of Shaoxing Huangjiu (Chinese Rice Wine) by near infrared spectroscopy combined with chemometrics [J]. *J Food Compos Anal*, 2022, 111: 104563.

[15] Zhao F Y, Du G R, Huang Y. Exploring the use of Near-infrared spectroscopy as a tool to predict quality attributes in prickly pear (*Rosa roxburghii* Tratt) with chemometrics variable strategy [J]. *J Food Compos Anal*, 2022, 105: 104225.

[16] Tahir H E, Arslan M, Mahunu G K, *et al*. Authentication of the geographical origin of Roselle (*Hibiscus sabdariffa* L) using various spectroscopies: NIR, low-field NMR and fluorescence [J]. *Food Contr*, 2020, 114: 107231.

[17] Qian L L, Li D W, Song X J, *et al*. Identification of Baha'sib mung beans based on Fourier transform near infrared spectroscopy and partial least squares [J]. *J Food Compos Anal*, 2022, 105: 104203.

[18] Chen X, Liu H G, Li J Q, *et al*. A geographical traceability method for *Lanmaoa asiatica* mushrooms from 20 township-level geographical origins by near infrared spectroscopy and ResNet image analysis techniques [J]. *Ecol Inform*, 2022, 71: 101808.

[19] 李长滨, 牛畅炜, 苏丽, 等. 不同产地山药的近红外鉴别和差异分析 [J]. 食品研究与开发, 2022, 43(15): 175-181.

[20] Yang Y, Yang L C, He S Y, *et al*. Use of near-infrared

spectroscopy and chemometrics for fast discrimination of *Sargassum fusiforme* [J]. *J Food Compos Anal*, 2022, 110: 104537.

[21] 林欣, 黄世安, 张琴, 等. 采用UPLC-MS/MS分析低温贮藏期间‘空心李’果实初生代谢物 [J]. 植物生理学报, 2022, 58(10): 1982-1994.

[22] Varrà M O, Fasolato L, Serva L, *et al*. Use of near infrared spectroscopy coupled with chemometrics for fast detection of irradiated dry fermented sausages [J]. *Food Contr*, 2020, 110: 107009.

[23] 王震. 不同生长期穿心莲药材 HPLC 指纹图谱及化学模式识别 [J]. 药物分析杂志, 2021, 41(3): 410-420.

[24] Guan Q, Pu T T, Zhou Z Y, *et al*. Multi-element and metabolite characterization of commercial *Phyllanthi Fructus* with geographical authentication and quality evaluation purposes [J]. *Food Contr*, 2023, 151: 109787.

[25] 丁茹茗, 徐晓光, 刘瑞, 等. 基于RGB空间的非经典K最近邻算法应用研究 [J]. 井冈山大学学报: 自然科学版, 2023, 44(3): 70-75.

[26] 符保龙, 陈如云. 分类回归树在高校计算机联考数据分析中的应用 [J]. 计算机时代, 2010(1): 33-34.

[27] 赵邑新, 王建国, 吴建平. 测试集自动生成方法中的可执行化研究 [J]. 计算机研究与发展, 2001, 38(1): 74-80.

[28] 邓红卫, 罗亮. 基于SMA算法优化随机森林的PPV预测模型 [J]. 黄金科学技术, 2023, 31(4): 624-634.

[29] 李兴鑫, 朱友文, 王箭. 安全高效的加密数据朴素贝叶斯训练和分类 [J]. 密码学报, 2022, 9(3): 448-467.

[30] Ribeiro J P O, Medeiros A D, Caliari I P, *et al*. FT-NIR and linear discriminant analysis to classify chickpea seeds produced with harvest aid chemicals [J]. *Food Chem*, 2021, 342: 128324.

[31] Liang N, Sun S S, Zhang C, *et al*. Advances in infrared spectroscopy combined with artificial neural network for the authentication and traceability of food [J]. *Crit Rev Food Sci Nutr*, 2022, 62(11): 2963-2984.

[32] 陈新谦. 山楂药用简史 [J]. 中国科技史料, 1985, 6(4): 18-22.

[33] 蒋昊. 北山楂、南山楂和广山楂性状鉴别和有机酸成分研究进展 [J]. 辽宁中医药大学学报, 2023, 25(1): 132-137.

[责任编辑 时圣明]