基于近红外光谱技术的冀产黄芩鉴别方法研究

郭兆华[1]，文师召[2]，李亚薇[3]，李思凡[4]，王琪[4]，刘思琪[5]，王鑫国[5]，牛丽颖[5]，冯薇[5*]（1. 中国电子科技集团公司网络通信研究院，石家庄 050050；2. 南开大学统计与数据科学学院 天津 300192；3. 辽宁省重大技术装备战略基地建设工程中心，沈阳 110032；4. 东北大学理学院，沈阳 110167；5. 河北中医药大学药学院，中药材品质评价与标准化河北省工程研究中心，石家庄 050091）

摘要：目的 采用近红外光谱（near-infrared spectioscopy，NIRS）技术对黄芩的河北省道地性进行二分类实验，探究不同数据预处理算法及其组合、不同波段选择算法和不同分类方法对模型性能的影响。**方法** 研究采集 138 份黄芩样本，采用 12 500～4 000 cm^{-1}波段对不同黄芩样品进行近红外光谱采集。首先，对比不同光谱预处理方法的单一性能与组合性能；其次，对比竞争性自适应重加权采样方法（CARS）、无信息变量消除方法（UVE）、连续投影方法（SPA）和主成分分析（PCA）在红外光谱波段选择与特征提取方面的性能；最后，对比偏最小二乘判别分析（PLS-DA）、支持向量机（SVM）、人工神经网络（ANN）、随机森林（RF）、传统一维卷积神经网络（CNN）和堆叠自编码器（SAE）在建立中药属性分类模型中的性能差异。**结果** 最佳预处理算法是使用均值中心化（MC）和多元散射校正（MSC），总体正确率可以达到 92.9%；最佳波段选择算法是基于 PCA 选择出的 25 维变量，能将总正确率提升 10.7%；最佳分类算法是经过 MSC 处理和 PCA 降维后建立的一维 CNN 模型，可以实现冀产黄芩 100% 正确率的道地性二分类。**结论** 通过近红外光谱技术的道地性分类算法研究为黄芩道地性鉴别提供了快速、无损的检测手段及可靠的数据分析方法，为中药材产地溯源提供新的方法参考。

关键词：黄芩；近红外光谱；产地鉴别；机器学习

doi：10.11669/cpj.2024.20.008　　**中图分类号：**R282　　**文献标志码：**A　　**文章编号：**1001－2494（2024）20－1939－09

Identification Method of Hebei Produced *Scutellaria Baicalensis* Based on Near-Infrared Spectroscopy Technology

GUO Zhaohua[1], WEN Shizhao[2], LI Yawei[3], LI Sifan[4], WANG Qi[4], LIU Siqi[5], WANG Xinguo[5], NIU Liying[5], FENG Wei[5*] (1. *Network Communication Research Institute, China Electronics Technology Group Corporation, Shijiazhuang* 050050, *China*; 2. *School of Statistics and Data Science, Nankai University, Tianjin* 300192, *China*; 3. *Major Technical Equipment Construction and Engineering Center of Liaoning Province, Shenyang* 110032, *China*; 4. *College of Sciences, Northeastern University, Shenyang* 110167, *China*; 5. *Quality Evaluation & Standardization Hebei Province Engineering Research Center of Traditional Chinese Medicine, School of Pharmaceutical Sciences, Hebei University of Chinese Medicine, Shijiazhuang* 050091, *China*)

ABSTRACT: OBJECTIVE To explore the effects of different data preprocessing algorithms and their combinations, different band selection algorithms and different classification methods on the performance of the model by using near infrared spectroscopy to classify the genuineness of *Scutellaria baicalensis* Georgi in Hebei province. **METHODS** A total of 138 samples of *Scutellaria baicalensis* Georgi were collected, and the spectral acquisition of different *Scutellaria baicalensis* Georgi samples was carried out by using 12 500－4 000 cm^{-1} band. Firstly, the single performance and combined performance of different spectral preprocessing methods are compared. Secondly, the performance of competitive adaptive reweighted sampling (CARS), uninformative variable elimination (UVE), successive projections algorithm (SPA) and principal component analysis (PCA) in infrared spectral band selection and feature extraction are compared. Finally, the performance differences of partial least squares discriminant analysis (PLS-DA), support vector (SVM), artificial neural network (ANN), random forest (RF), traditional one-dimensional convolutional neural network (CNN) and stacked autoencoder (SAE) in establishing the attribute classification model of traditional Chinese medicine were compared. **RESULTS** The best preprocessing algorithm is to use mean centralization (MC) and multiple scattering correction (MSC), and the overall accuracy rate can reach 92.9 %. The optimal band selection algorithm is based on the 25-dimensional variables selected by PCA, which can increase the total accuracy by 10.7 %. The best classification algorithm is a one-dimensional CNN model established after MSC processing and PCA dimensionality reduction, which can achieve 100 % accuracy of geo-authentic

基金项目：河北省省级科技计划项目资助（21372503D）

作者简介：郭兆华，男，硕士，高级工程师　　研究方向：算法设计与模型分析　　*通讯作者：冯薇，女，博士，教授，博士生导师　　研究方向：中药材品质评价　　Tel：（0311）85216828

binary classification. **CONCLUSIONS** The study of genuineness classification algorithm by infrared spectroscopy provides a fast and non-destructive detection method and reliable data analysis method for the genuineness classification of *Scutellaria baicalensis*, and provides a new method reference for the traceability of Chinese medicinal materials.

KEY WORDS: Scutellariae Radix; near infrared spectroscopy; origin identification; machine learning

黄芩为唇形科植物黄芩(*Scutellaria baicalensis* Georgi)的干燥根,具有清热燥湿、泻火解毒、止血、安胎的功效。黄芩是我国常用的大宗药材,多地均大量栽培,河北、山西、内蒙古等地是黄芩的道地产区[1]。其中,冀产黄芩种植历史悠久,承德、保定、张家口等地区是黄芩的主产地,不仅是国内市场流通的重要来源,且作为河北的传统出口商品之一而享誉海外。

然而除黄芩的栽培管理技术参差、种质资源混杂等因素外,是否道地产区也对黄芩药材质量产生较大影响[2]。特定地域出产的道地药材是历史悠久、品质优良、疗效突出、优质高产等优点的代名词,其最基本的内涵为中药材质量优良[3]。无论道地药材还是植物类地理标志产品,都因为其生长环境和种植技术的差异而独具特色,因此对药材等天然产物进行快速产地溯源鉴别的需求日益迫切。

近红外光谱(near-infrared spectroscopy,NIRS)技术因其"快速、便捷、无损"等特点越来越多地应用于药材品质检测和产地溯源鉴别研究中。Li 等[4]使用 SG 平滑算法和标准正态变量变换(SNV)进行光谱数据预处理,结合线性判别分析等机器学习分类方法建立中药类别与产地鉴别模型。Li 等[5]对预处理后的山药光谱数据建立聚类判别和簇类独立软模式(SIMCA)判别模型,分类准确率均为 100%。Li 等[6]首先通过单一及组合方法对光谱数据进行预处理,再结合 Fisher 线性判别分析方法和主成分分析(PCA)分别构建不同产地茯苓块的鉴别模型,最佳鉴别率可达 91.7%。Zhou 等[7]采用 K 均值聚类法结合偏最小二乘法(PLS)对预处理光谱进行分析,能够有效地区分川贝母和其他贝母。Bai 等[8]采用不同的光谱预处理方法,结合 PCA 联用马氏距离法(PCA-MD)和偏最小二乘判别分析法(PLS-DA)对天麻进行产地鉴别,准确率分别达到 90.91% 和 95.45%。Wang 等[9]使用小波去噪处理光谱数据,采用 PLS 降维并将主因子得分作为广义回归神经网络(GRNN)的输入,从而建立豆粕的品质检测模型。Li 等[10]研究了偏最小二乘回归前波长变量选择在傅里叶变换近红外快速鉴别苹果汁掺假中的应用,采用连续投影算法(SPA)结合四种群体智能优化算法提取有效波长变量。

近红外光谱和化学计量学作为协同演进的两大技术,在提升数据解析效率与准确性方面展现出独特优势。在近红外光谱数据的处理中,化学计量学方法如光谱预处理算法、波段选择策略及定性模式识别技术等,扮演了至关重要的角色。不仅能够有效降低光谱数据中的噪声与无关信息干扰,还通过精细的数据优化,为构建更加精确、稳健的模型提供了坚实基础[11]。关于黄芩的近红外研究大多集中于黄芩或含黄芩复方中成分含量测定及栽培黄芩与对照药材的图谱比较,缺少对黄芩道地性鉴别的研究[12-14]。鉴于道地药材产地溯源对于确保中药品质与疗效的重要性,本研究聚焦于冀产黄芩,致力于开发高效的光谱预处理算法,通过化学计量学的数据处理手段实现对光谱数据的深度净化,不仅降低了模型构建的复杂性,同时显著提升了黄芩道地性产地溯源模型的预测精度与可靠性。为中药材的质量监控与道地性认证提供了强有力的技术支撑,对保障中药材"质优效佳"具有重要意义。

1 仪器与材料

1.1 实验仪器

MPA 型傅里叶变换近红外光谱仪(德国布鲁克光学仪器公司),配备 OPUS 光谱采集软件,测量方式选用固体积分球漫反射方式;YB-150 型多功能粉碎机(浙江永康市速锋工贸有限公司);DHG-9123A型电热恒温鼓风干燥箱(上海一恒科技有限公司)。

1.2 样品

所用药材从不同产地(河北省、内蒙古自治区、山西省、陕西省)采集,其中河北省采集样本来自承德、安国、张家口、衡水、行唐、博野等几个黄芩主要产区。经河北中医药大学侯芳洁副教授鉴定为唇形科植物黄芩(*Scutellaria baicalensis* Georgi)的干燥根,分别有采集自以上 4 个省(自治区)的 82、11、9 和 36 个共计 138 批样本,将除河北省以外的所有样品归为非冀产类别。将收集到的以上批次不同产地的黄芩样品,粉碎,过

80 目筛，混匀，样品详情见表 1。

表 1 138 批黄芩药材的详细信息

Tab. 1 Information of 138 batches of *Scutellaria Baicalensis*

Place of origin(in Chinese)	Sample No.	Collection time
Chengde, Hebei(河北承德)	S1-9; S87-92	2021-10
Zhangjiakou, Hebei(河北张家口)	S10-14; S81; S93-96	2021-10
Xingtang, Hebei(河北行唐)	S15-17; S97-99	2021-10
Hengshui, Hebei(河北衡水)	S18-22; S100-104	2021-10
Boye, Hebei(河北博野)	S23-27; S105-107	2021-10
Anguo, Hebei(河北安国)	S29-46; S85; S108-119;	2021-10
Quyang, Hebei(河北曲阳)	S83	2021-10
Wuji, Hebei(河北无极)	S28	2021-10
Chifeng, Inner Mongolia(内蒙古赤峰)	S47-51; S72; S120-124	2021-11
Yuncheng, Shanxi(山西运城)	S52-55; S73; S84; S125-127	2021-11
Weinan, Shaanxi(陕西渭南)	S56-64; S69; S74-76; S80; S86; S128-131; S134-135	2021-11
Yan'an, Shaanxi(陕西延安)	S65-68; S70-71; S77-79; S132-133; S136-138	2021-11
Yulin, Shaanxi(陕西榆林)	S82	2021-11

2 方 法

2.1 NIRS 信息采集

取适量的黄芩药材粉末，放入石英样品杯中至三分之二处，均匀铺平，45 ℃烘干至恒重。采用德国布鲁克 MPA 傅里叶变换近红外光谱仪，以空气为参比扣除背景采集光谱图，采用积分球漫反射采集光谱，扫描条件为：分辨率 8 cm^{-1}，扫描波段范围：12 500 ~ 4 000 cm^{-1}，样品背景和样品扫描时间：32 s，每批样品重复扫描 3 次，计算平均光谱。

黄芩近红外光谱数据集共有 138 个样本，所采取的黄芩样品中样本的波数变化范围为[3 999.879 22，12 493.354 66 cm^{-1}]，共 2 203 个波长点，见图 1。

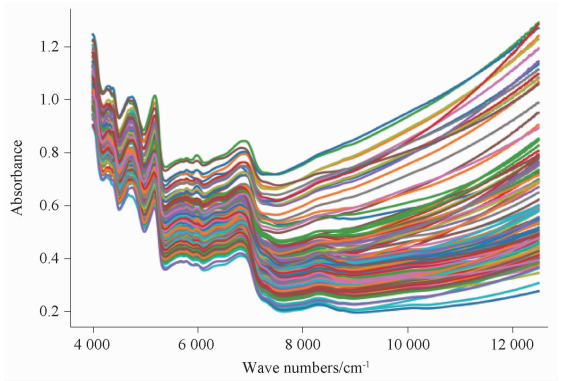


图 1 黄芩近红外光谱数据集图像

Fig. 1 Image of near-infrared spectral dataset of *Scutellaria baicalensis*

2.2 光谱信息的预处理

在光谱采集过程中，除含被测样本待测成分信息外，还包括测量环境及仪器产生的噪声，如高频随机噪声、基线漂移等无关信息。为减弱或消除各种非目标因素对光谱信息的影响，常对原始光谱数据进行预处理，以提高模型预测能力及稳定性。本研究对比原始光谱、去趋势校正[15]（detrending，DT）、SNV、多元散射校正[16]（MSC）、标准化[17]（SS）、均值中心化[17]（MC）、最小最大值归一化法[17]（MMS）、矢量归一化[17]（VS）、移动平均平滑[17]（MA）、Savitzky-Golay 卷积平滑法[18]、一阶导数变换[19]（1st Derivate）、二阶导数变换[19]（2nd Derivate）、连续小波变换[20]等光谱预处理方法的单一性能与组合性能。

2.3 光谱波段选择

同时，探究如何将传统模型识别方法和深度学习方法进行优势结合，建立简单有效稳定的中药属性分类模型。对比竞争性自适应重加权采样方法[21]（CARS）、无信息变量消除方法[22]（UVE）[23-24]、SPA 和 PCA 在红外光谱波段选择与特征提取方面的性能，具体波段算法选择参数见表 2。

表 2 黄芩产地鉴别模型波段选择算法参数设定

Tab. 2 Parameter setting of band selection algorithm in the model for identifying the model for identifying the origin of scufellaria baicalensis

	Parameter	Parameter value
CARS	Sampling frequency	$N = 50$
	Maximum number of principal components for cross-validation	$PC = 20$
	Number of cross-validation	$CV = 10$
UVE	PLS test set proportion	0.2
	Number of cross-validation	$CV = 10$
SPA	Total number of wavelength points	$h = 17$
	Number of cross-validation	$CV = 10$

2.4 分类算法选择

对比 PLS-DA[25]、支持向量机[26-27]（SVM）、随机森林[28-29]（RF）、传统一维卷积神经网络（CNN）、人工神经网络[30]（ANN）和堆叠自编码器[31]（SAE）在建立中药属性分类模型中的性能差异，具体算法参数见表 3 ~ 4。

2.5 数据处理

本实验中所使用的数据集为黄芩数据集，根据黄芩的近红外光谱数据进行黄芩的河北省道地性分类（即转化为二分类任务，将所有黄芩样本分为冀产与非冀产）。具体数据集信息见表 5。

表 3 黄芩产地鉴别模型中一维卷积神经网络(CNN)结构

Tab. 3 One-dimensional convolutional neural network (CNN) structure in scutelloria baicalensis origin identification model

Number of layers	Structure
1	1D Convolutional kernel(1,16,21)
2	Batch normalization
3	Relu activation function
4	1D Convolutional kernel(16,32,19)
5	Batch normalization
6	Relu activation function
7	1D Convolutional kernel(21,64,17)
8	Batch normalization
9	Relu activation function
10	Fully connected(137 536,512)
11	Fully connected(512,4)
12	Softmax

表 4 黄芩产地鉴别模型中堆叠自编码器(SAE)训练参数

Tab. 4 SAE training parameters in scutelloria baicalensis origin identification model

Parameters	Values
Number of training epochs for autoencoder	200
Loss function for autoencoder training	MSE loss function
Number of fine-tuning epochs for stacked autoencoder	200
Loss function for stacked autoencoder fine-tuning	Cross-entropy loss function/ focal loss function
Optimizer	Adam
Initial learning rate	0.001
Learning rate decay rate	0.000 1
Batch size	128

表 5 黄芩道地性分类数据集信息

Tab. 5 Geoherbalism classification data set information of Scutellariae Radix

Dataset metrics	Value	Dataset indicators	Ref.
Total number of samples	138	Category distribution	[82,56]
Proportion of test set	0.2	Training set sample distribution	[65,45]
Number of samples in training set	110	Test set sample distribution	[17,11]
Number of samples in test set	28	Sample dimension	2 203

3 结果与分析

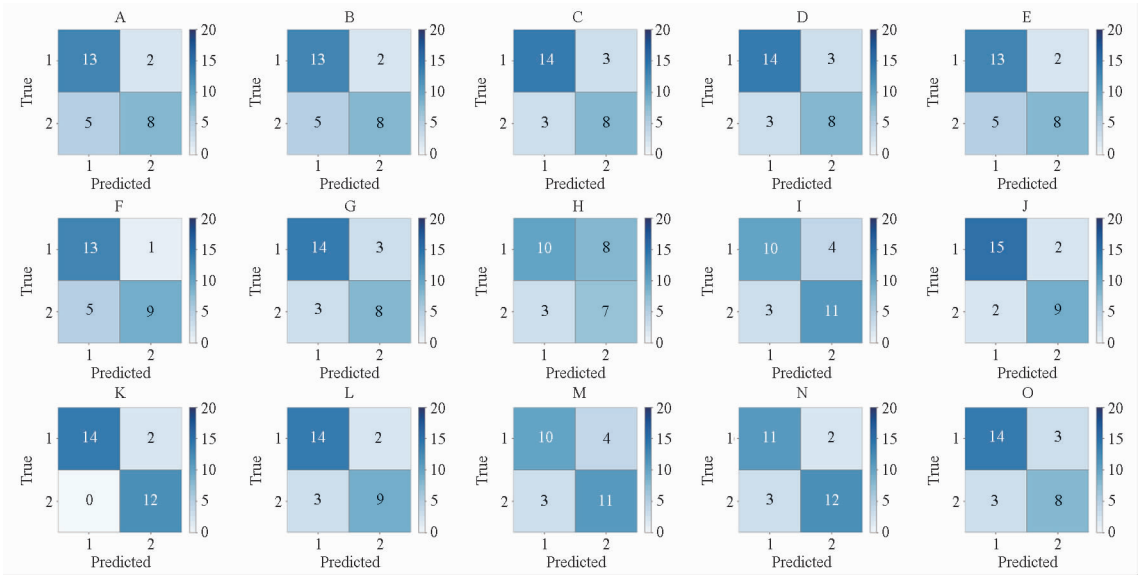
3.1 不同预处理算法对于模型性能的影响

实验数据编码方式使用顺序编码,使用全波长点进行建模,分类模型基于 SVM 算法,其中核函数使用高斯核,惩罚系数设为 30。图 2 展现了不同预处理算法下的混淆矩阵,具体实验结果见表 6。表 6 展示的是不同预处理算法下的实验结果的总体正确率(OA)、Kappa 系数、前两个类别的 F1 分数、准确率和召回率。观察表 6 可以发现,在不进行预处理情况下,对黄芩的道地性分类总体正确率为 75%,大部分预处理算法的使用可以改善模型的性能,其中使用 MC 和 MSC 后的模型总体正确率可以达到 92.9%,Kappa 系数也为 $\kappa = 0.858$ 说明分类器预测结果的一致性非常好。但是也存在一些预处理算法会损失有效信息而导致模型性能降低,如 DT,而经过 SNV 处理后再进行去趋势运算,便可将模型总体正确率提升至 82.1%。

图像展现出不同预处理方法后模型的正确率与 Kappa 系数对比(图 3A),能够发现先进行 MC,再进行 MSC 的模型正确率和 Kappa 系数最高,其他模型在一定程度上都能提升模型的性能,除单独进行 DT 会降低模型准确度与 Kappa 系数。展现出了不同预处理方法模型后的 F1 系数对比(图 3B),发现河北省安国类别的 F1 系数都在 0.8 左右。其中先进行 MC,再进行 MSC 处理的 F1 系数最高,单独进行 DT 的模型 F1 系数最低,其他非冀产类的 F1 系数与冀产类别类似。除此之外,可以发现结合多种算法进行预处理的模型性能明显好于单一预处理算法,这是因为在实际生产应用中,真实数据中的噪声和无用信息可能复杂多样。但是想要选择出更好的预处理算法组合,不仅需要一定的近红外光谱分析专业知识,同样也需要不断地重复组合实验。

3.2 不同波段选择算法对于模型性能的影响

图 4 展现出不同波段选择算法的模型的混淆矩阵,具体实验结果见表 7,展现出选择不同波段算法进行分类,最终模型的 OA、Kappa 系数、前两个类别的 F1 分数、准确率和召回率。通过表 6 ~ 7 可以发现,CARS 选择出的 71 维变量更具有代表意义,有助于模型性能的提高,而相对于传统的 SPA 和 UVE 可能由于在波段选择中损失一定的信息,导致模型性能下降。值得一提的是,基于 PCA 选择出的 25 维变量对于模型性能的提高最为显著,将总正确率提升 10.7%,Kappa 系数也超过 0.9,说明分类器预测结果的一致性非常好。



A－无；B－均值中心化；C－标准化；D－最小最大值归一化法；E－移动平均平滑；F－卷积平滑法；G－基于小波变换的聚类分析方法；H－去趋势算法；I－标准正态变量变换；J－多元散射校正；K－均值中心化-多元散射校正；L－标准化-多元散射校正；M－矢量归一化-标准正态变量变换；N－标准正态变量变换-去趋势算法；O－移动平均平滑-最小最大值归一化法。

A－None；B－MC；C－SS；D－MMS；E－MA；F－SG；G－WAVE；H－DT；I－SNV；J－MSC；K－MC-MSC；L－SS-MSC；M－VS-SNV；N－SNV-DT；O－MA-MMS.

图 2 黄芩产地鉴别模型中不同预处理算法的混淆矩阵

Fig. 2 The confusion matrix of different preprocessing algorithms in scutelloria baicalensis origin identification model

表 6 黄芩产地鉴别模型中不同预处理算法对比实验结果

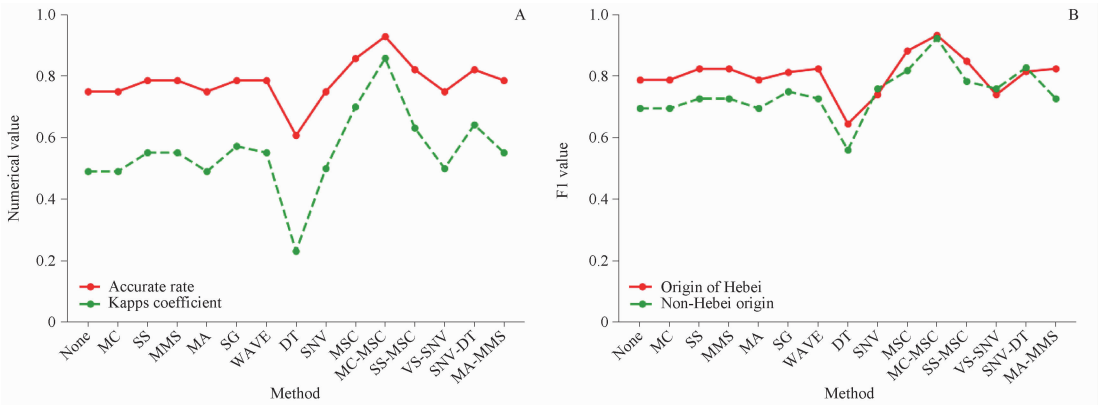
Tab. 6 Comparison of experimental results of different preprocessing algorithms in scutelloria baicalensis origin identification model

No.	Preprocessing algorithm	OA	κ	F_1,1	F_1,2	P_1	P_2	R_1	R_2
1	None	0.750	0.490	0.788	0.695	0.867	0.615	0.722	0.800
2	['MC']	0.750	0.490	0.788	0.695	0.867	0.615	0.722	0.800
3	['SS']	0.786	0.551	0.824	0.727	0.824	0.727	0.824	0.727
4	['MMS']	0.786	0.551	0.824	0.727	0.824	0.727	0.824	0.727
5	['MA']	0.750	0.490	0.788	0.695	0.867	0.615	0.722	0.800
6	['SG']	0.786	0.572	0.813	0.750	0.929	0.643	0.722	0.900
7	['WAVE']	0.786	0.551	0.824	0.727	0.824	0.727	0.824	0.727
8	['DT']	0.607	0.230	0.645	0.560	0.556	0.700	0.769	0.467
9	['SNV']	0.750	0.500	0.740	0.759	0.714	0.786	0.769	0.733
10	['MSC']	0.857	0.700	0.882	0.818	0.882	0.818	0.882	0.818
11	['MC','MSC']	0.929	0.858	0.933	0.923	0.875	1.000	1.000	0.857
12	['SS','MSC']	0.821	0.631	0.849	0.783	0.875	0.750	0.824	0.818
13	['VS','SNV']	0.750	0.500	0.740	0.759	0.714	0.786	0.769	0.733
14	['SNV','DT']	0.821	0.642	0.815	0.828	0.846	0.800	0.786	0.857
15	['MA','MMS']	0.786	0.551	0.824	0.727	0.824	0.727	0.824	0.727

注：['MA','MMS']表示先进行移动平均平滑处理(MA)，再进行最小最大归一化(MMS)，其他处理顺序亦同。

Note: "['MA','MMS']" indicates that a moving average smoothing (MA) process is conducted first, followed by a min-max normalization (MMS) process. The same applies to other processing sequences.

图 5A 展现出不同波段选择方法后模型的正确率与 Kappa 系数对比，能够发现进行 PCA 的正确率与 Kappa 系数在五种波段选择算法中最高，正确率能够达到 96.4%，Kappa 系数为 0.926，模型性能最好。图 5B 展现出不同波段选择的 F1 值，两种类别(冀产与非冀产)的 F1 值均在使用 PCA 下，达到最大。而在传统的 SPA 和 UVE 下，相较于未进行波段选择的模型来说，F1 值下降，说明使用这两种方法的模型性能下降。

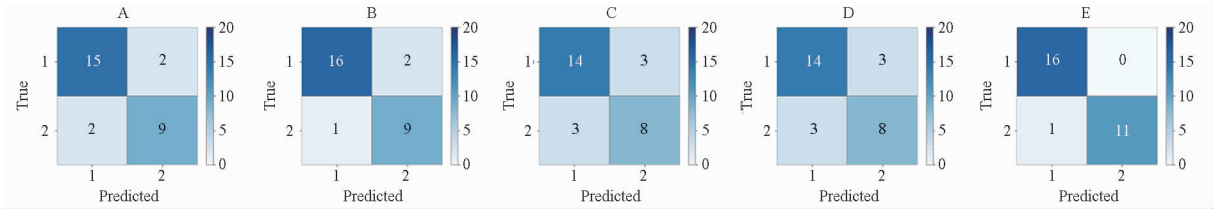


A – 不同预处理方法正确率和 Kappa 系数对比；B – 不同预处理方法 F1 系数对比

A – comparison of accuracy and Kappa coefficient for different preprocessing methods; B – comparison of F1 scores for different reprocessing methods

图 3 黄芩产地鉴别模型中不同预处理策略的分类结果对比

Fig. 3 Comparison of classification results of different preprocessing strategies in scutelloria baicalensis origin identification model



A – 无；B – 竞争性自适应重加权采样方法；C – 连续投影方法；D – 无信息变量消除方法；E – 主成分分析。

A – None; B – CARS; C – SPA; D – UVE; E – PCA.

图 4 黄芩产地鉴别模型中不同波段选择算法的混淆矩阵

Fig. 4 The confusion matrix of different band selection algorithms in scutelloria baicalensis origin identification model

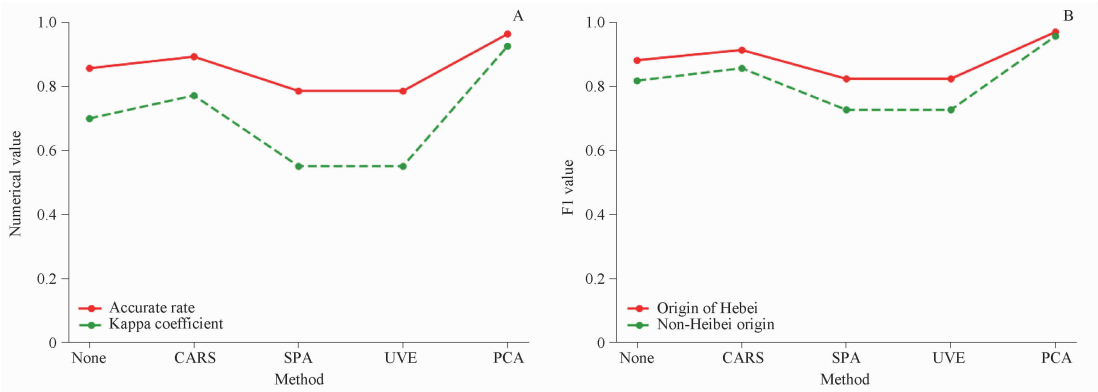
表 7 黄芩产地鉴别模型中不同波段选择算法对比实验结果

Tab. 7 Comparison of experimental results of different band selection algorithms in scutelloria baicalensis origin identification model

No.	Band selection algorithm	Dim	OA	κ	F_1 ,1	F_1 ,2	P_1	P_2	R_1	R_2
1	None	2203	0. 857	0. 700	0. 882	0. 818	0. 882	0. 818	0. 882	0. 818
2	CARS	71	0. 893	0. 772	0. 914	0. 857	0. 889	0. 9	0. 941	0. 818
3	SPA	17	0. 786	0. 551	0. 824	0. 727	0. 824	0. 727	0. 824	0. 727
4	UVE	13	0. 786	0. 551	0. 824	0. 727	0. 824	0. 727	0. 824	0. 727
5	PCA	25	0. 964	0. 926	0. 970	0. 957	1. 000	0. 917	0. 941	1. 000

注：Dim – 经过波段选择后的样本波段数。

Note: Dim – the number of sample bands after band selection.



A – 不同波段选择方法正确率和 Kappa 系数对比；B – 不同波段选择方法 F1 系数对比。

A – comparison of accuracy and Kappa coefficient for the different band selection method; B – comparison of F1 coefficients for the different band selection method.

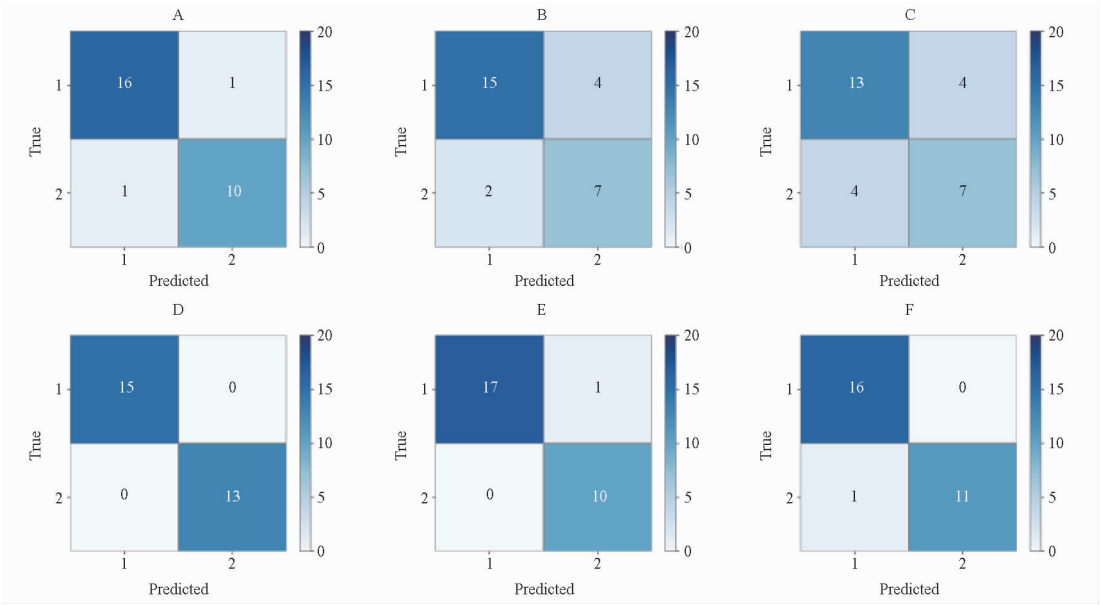
图 5 黄芩产地鉴别模型中不同波段选择方法的分类结果对比

Fig. 5 Comparison of the classification results for the different band selection method in scutelloria baicalensis origin identification model

3.3 不同分类算法对于模型性能的影响

在本阶段实验中，主要探究不同分类算法对于模型性能的影响。实验所使用的数据集仍为表5所描述的黄芩地道性分类数据集，其中一维卷积神经网络和其他传统分类器的重要参数见表3～4。实验数据编码方式使用顺序编码，采用MSC进行光谱预处理，利用PCA算法进行数据降维，实验结果见图6、表8。

图6展现了在相同预处理方式以及波段选择算法条件下，不同分类算法下的混淆矩阵。表8更为详细地展现出不同分类算法对于模型OA、Kappa系数、前两个类别的F1分数、准确率和召回率的影响。通过表8，可以发现这些方法中，一维卷积神经网络的方法效果最优，在现有数据下分类正确率达到100%，能够正确地分类出黄芩的产地，Kappa系数也达到1。相比较而言，采用高斯核的SVM和SAE效果仅次于CNN，采用非线性核函数高斯核的SVM可以找到良好的区分平面对样本进行划分，而SAE的结构也有助于模型学习到更好的样本特征，从而进行分类。在6种分类算法中，RF和PLS-DA的模型准确率不达80%，并且Kappa系数不足0.6，说明分类器预测结果的一致性一般，分类器性能不高。



A－人工神经网络；B－随机森林；C－偏最小二乘判别分析；D－一维卷积神经网络；E－堆叠自编码器；F－支持向量机。

A－ANN；B－RF；C－PLS－DA；D－CVV；E－SAE；F－SVM.

图6 黄芩产地鉴别模型中不同分类方法的混淆矩阵

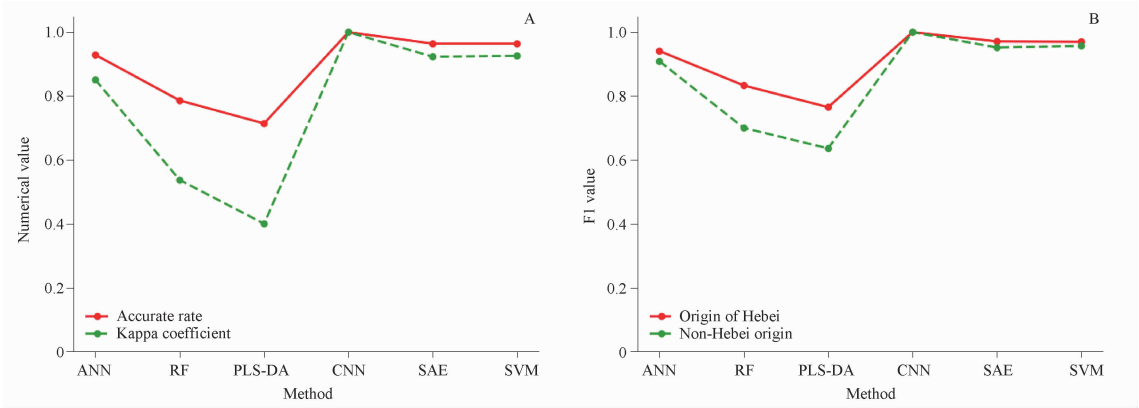
Fig. 6 Confusion matrix of different classification methods in scutelloria baicalensis origin identification model

表8 黄芩产地鉴别模型中不同分类算法对比实验结果

Tab. 8 Comparison of experimental results of different classification algorithms in scutelloria baicalensis origin identification model

No.	Classification algorithm	OA	κ	F_1,1	F_1,2	P_1	P_2	R_1	R_2
1	ANN	0.929	0.851	0.941	0.909	0.941	0.909	0.941	0.909
2	RF	0.786	0.537	0.833	0.700	0.789	0.778	0.882	0.636
3	PLS-DA	0.714	0.400	0.765	0.636	0.765	0.636	0.765	0.636
4	CNN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	SAE	0.964	0.923	0.971	0.952	0.944	1.000	1.000	0.909
6	SVM	0.964	0.926	0.970	0.957	1.000	0.917	0.941	1.000

图7A展现了不同分类方法的正确率与Kappa系数对比，通过图7A发现PLS-DA正确率与Kappa系数都偏低。而ANN、CNN、SAE与采用高斯核的SVM这几种分类算法效果都相对较好，其中尤其CNN模型能够在现有数据下达到100%正确率。图7B展现出两个类别（冀产与非冀产）的F1值，也是采用CNN模型时模型效果最好。



A – 不同分类方法正确率和 Kappa 系数对比;B – 不同分类方法 F1 系数对比。

A – comparison of accuracy and Kappa coefficient for different classification methods; B – comparison of F1 coefficients for different classification methods.

图 7 黄芩产地鉴别模型中不同分类方法的分类结果对比

Fig. 7 Comparison of classification results for different classification methods in scutelloria baicalensis origin identification model

4 讨 论

在使用传统方法建立中药属性分类模型时,通常采取“光谱预处理-波段选择-分类模型构建”的策略。但是由于光谱数据中包含干扰因素较多,通常需要采取不同类型预处理策略的组合。对于波段选择策略,若选取不当,也容易损失部分重要光谱信息导致模型性能下降。在不同的分类方法中,SAE 和一维 CNN 的效果相对较好,但是参数量大,且也需要一定的预处理策略。其他分类方法如 SVM 则会受到重要参数的影响,往往需要参数调优才能达到更理想的效果。本研究存在样本量较小的局限性,希望未来能够扩大样本量、进行更多的算法比较以及采用数据增强等方法,以提高我们研究的准确性和普遍性。

REFERENCES

[1] WANG H Y, YANG L, WANG D Y, *et al*. Herbal textual research and analysis on the traceability and origin change of Scutellaria baicalensis varieties [J]. *J Shaanxi Univ Chin Med*(陕西中医药大学学报), 2021, 44(3): 20-25.

[2] CUI L, LU J X, LIN H S, *et al*. Investigation on the resources and production status of Scutellaria baicalensis in China [J]. *LiShizhen Med Mater Med Res*(时珍国医国药), 2009, 20(9): 2279-2280.

[3] MENG X C, DENG D Q, DU H W, *et al*. Scientific connotation of high-quality genuine medicinal materials [J]. *Chin Tradit Herb Drugs*(中草药), 2023, 54(3): 939-947.

[4] LI X J, WANG T. Classification and origin identification of Chinese medicinal materials based on infrared spectroscopy analysis [J]. *Math Mod Appl*(数学建模及其应用), 2022, 11(3): 50-59.

[5] LI C B, NIU C W, SU L, *et al*. Identification and variance analysis of Chinese yam from different origins by near infrared spectroscopy [J]. *Food Res Dev*(食品研究与开发), 2022, 43(15): 175-181.

[6] LI J Y, YU M, ZHENG Y, *et al*. Nondestructive identification of Poria cocos blocks from different origins based on near infrared spectroscopy [J]. *Chin J Anal Lab*(分析试验室), 2021, 40(12): 1381-1386.

[7] ZHOU T, FU S B, XIE H M, *et al*. Identification and validation of bulbs of Frillariae species using near-infrared spectroscopy data [J]. *West China J Pharm Sci*(华西药学杂志), 2021, 36(2): 193-197.

[8] BAI Q X, HOU Y, YANG P P, *et al*. Identification method of the producetion site of gastrodia elata blume based on near infrared spectroscopy [J]. *J West China For Sci*(西部林业科学), 2021, 50(3): 124-130.

[9] WANG L Q, YAO J, WANG R Y, *et al*. Research on Detection of Soybean Meal Quality by NIR Based on PLS-GRNN [J]. *Spectrosc Spectr Anal*(光谱学与光谱分析), 2022, 42(5): 1433-1438.

[10] LI Y, GUO Y, LIU C, *et al*. SPA combined with swarm intelligence optimization algorithms for wavelength variable selection to rapidly discriminate the adulteration of apple juice[J]. *Food Anal Method*, 2017, 10: 1965-1971.

[11] CHU X L, CHEN P, LI J Y, *et al*. Progresses and perspectives of near infrared spectroscopy analytical technology [J]. *J Instrum Anal*(分析测试学报), 2020, 39(10): 1181-1188.

[12] WANG F. Rapid and simultaneous determination of active components in raw and processed root samples of scutellariae radix by near-infrared spectroscopy [D]. Xinxiang: Xinxiang Medical University, 2022.

[13] MA J F, WANG X L, XIAO X, *et al*. Near infrared spectroscopy non-destructive assay of key quality-indicative ingredients of Antai Pills [J]. *Mod Tradit Chin Med Mater Med-World Sci Technol*(世界科学技术-中医药现代化), 2018, 20(5): 651-659.

[14] ZHAO J J, GAO X J, WANG Y H, *et al*. Comparative studies on HPLC fingerprint and near-infrared spectra of cultivated and reference crude Scutellaria baicalensis [J]. *China J Chin Mater Med*(中国中药杂志), 2016, 41(22): 4204-4209.

[15] BARNES R J, DHANOA M S, LISTER S J. Standard normal

variate transformation and de-trending of near-infrared diffuse reflectance spectra [J]. *Appl Spectrosc*, 1989, 43(5): 772-777.

[16] ISAKSSON T, NÆS T. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy [J]. *Appl Spectrosc*, 1988, 42(7): 1273-1284.

[17] CHU X L. *Chemometrics Methods in Modern Spectral Analysis* [M]. Beijing: Chemical Industry Press, 2022: 10-14, 79-82.

[18] SAVITZKY A, GOLAY M J E. Smoothing and differentiation of data by simplified least squares procedures [J]. *Anal Chem*, 1964, 36(8): 1627-1639.

[19] HOPKINS D W. What is a Norris derivative? [J]. *NIR News*, 2001, 12(3): 3-5.

[20] LEUNG A K, CHAU F, GAO J. Wavelet transform: a method for derivative calculation in analytical chemistry [J]. *Anal Chem*, 1998, 70(24): 5222-5229.

[21] LI H, LIANG Y, XU Q, *et al*. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. *Anal Chim Acta*, 2009, 648(1): 77-84.

[22] CENTNER V, MASSART D L, DE NOORD O E, *et al*. Elimination of uninformative variables for multivariate calibration [J]. *Anal Chem*, 1996, 68(21): 3851-3858.

[23] AŔAUJO M C U, SALDANHA T C B, GALVAO R K H, *et al*. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis [J]. *Chemometr Intell Lab*, 2001, 57(2): 65-73.

[24] SOARES S F C, GOMES A A, ARAUJO M C U, *et al*. The successive projections algorithm [J]. *Trac-trend Anal Chem*, 2013, 42: 84-98.

[25] GELADI P, KOWALSKI B R. Partial least-squares regression: a tutorial [J]. *Anal Chim Acta*, 1986, 185: 1-17.

[26] ZHANG X G, BIAN Z Q. *Pattern Recognition* (2*nd Edition*) [M]. Beijing: Tsinghua University Press, 2004.

[27] LI H, LIANG Y, XU Q. Support vector machines and its applications in chemistry [J]. *Chemometr Intell Lab*, 2009, 95(2): 188-198.

[28] BREIMAN L. Random forests [J]. *Mach Learn*, 2001, 45: 5-32.

[29] CHEN G, ZHANG X, WU Z, *et al*. An efficient tea quality classification algorithm based on near infrared spectroscopy and random Forest [J]. *J Food Process Eng*, 2021, 44(1): e13604.

[30] CHEN Y, THOSAR S S, FORBESS R A, *et al*. Prediction of drug content and hardness of intact tablets using artificial neural network and near-infrared spectroscopy [J]. *Drug Dev Ind Pharm*, 2001, 27(7): 623-631.

[31] SUN Z X, ZHAO Z G, LIU F. Near-infrared spectral modeling based on stacked supervised auto-encoder [J]. *Spectrosc Spectr Anal*(光谱学与光谱分析), 2022, 42(3): 749.

(收稿日期:2024-01-20)