基于近红外光谱和梯度提升决策树建立当归药材及伪品的定性判别模型*

拱健婷[1,2]，李　莉[1,2]，邹慧琴[3]，徐　东[3]，王大仟[1,2]，丛　悦[1,2]，刘长利[4]

（1. 北京市卫生局临床药学研究所　北京　100035；2. 首都医科大学附属北京中医医院　北京　100010；3. 北京中医药大学中药学院　北京　102488；4. 首都医科大学中医药学院　北京　100069）

摘　要：目的 建立NIRS技术快速无损鉴别当归药材及其伪品的方法。方法 采集当归及伪品断面的近红外光谱，结合模式识别法分析药材，用主成分分析（Principal component analysis，PCA）进行定性分析；对比梯度提升决策树（Gradient Boosting Decision Tree，GBDT）、支持向量机（Support Vector Machine，SVM）、人工神经网络（Artificial Neural Network，ANN）3种当归真伪判别模型的分类效果；利用RF筛选特征波长优化所建模型。结果 PCA无法有效区别当归及其伪品；与ANN、SVM相比，GBDT具有更高的准确性，训练集与预测集的总体准确率分别为94.39%和90.38%；而后以RF选取出20个特征波长，建立的近红外特征光谱判别模型训练集和预测集的总体准确率也达到了91.59%和86.54%。结论 近红外光谱技术结合GBDT鉴别当归药材真伪鉴别是可行的，为当归药材真伪快速无损鉴别提供了一种新方法。

关键词：当归　对比梯度提升决策树　近红外　模式识别　判别模型　真伪鉴别

doi: 10.11842/wst.20181023001　　中图分类号：R282.5　　文献标识码：A

当归为伞形科植物当归*Angelica sinensis*（Oliv.）Diels.的干燥根[1]，始载于《神农本草经》[2]，为常用的名贵大宗药材，素有“十药九归”之称，临床上多用于治疗血瘀证、血虚证，为“血中之圣药”[3]。当归是卫计委公示的“药食两用”品种之一[4]，也是保健品、化妆品、饮品、香料的原料，市场需求量大，是产销量位居第二的大宗中药商品[5]。随着其需求的增长和价格的提高，当归混伪品也日渐增多，近年来笔者发现市场上部分药材因名称与当归相似而混为当归药用，如云南野当归（*Angelica sp.*）、欧当归（*Levi sticum officinale* Koch.）和华中前胡（*Peucedanum medicum*）混作当归。这些植物在根茎形态上与正品当归十分类似，但在药效方面相差甚远[6,7]，严重影响用药安全，因此快速有效鉴别当归真伪成为当务之急。

已有许多报道采用性状鉴别法、紫外光谱法、薄层色谱法、分子技术等对当归进行鉴别[7-10]，取得了一定的成效，但是仍存在易受人为主观因素影响、样品制备复杂、特征信息较少、结果难量化、成本高、溶剂污染等局限。近红外光谱（Near Infrared Spectrum；NIRS）技术是正在迅速发展的一种绿色分析技术，具有快速、价廉、无损等特点，与化学计量学结合，广泛用于农业、食品、化学和石油化工、制药等领域的定性和定量分析[11]，被美、欧、日、韩、澳大利亚等国家药典纳入附录内容[12]。近红外光谱既能全面地反映中药的整体信息，在苦参[13]、大黄[14]、板蓝根[15]、三七[16]等中药真伪鉴别中近红外技术已经得到了很好的应用。此

收稿日期：2019-08-20

修回日期：2019-09-26

*　首都中医药研究专项重点课题（17ZY05）：不同产区当归功效组分特异性及其形成机制研究，负责人：王大仟；北京中医药大学在读研究生项目（2016-JYB-XS057）：基于中药变质多因素筛选及相关性评估构建中药加速模型，负责人：拱健婷。

**　通讯作者：李莉，研究员，主要研究方向：中药材质量评价与中药资源开发；刘长利，副教授，主要研究方向：中药材规范化生产及药材质量调控研究。

表1 样品信息表

样本编号	样本名	样本基原	产地	样本量
1-10	当归	Angelica sinensis	青海	10
11-15	当归	Angelica sinensis	甘肃	5
16-44	当归	Angelica sinensis	四川	29
45-66,76-112	当归	Angelica sinensis	云南	59
67-75	当归	Angelica sinensis	湖北	9
113-128	云南野当归	Angelica sp.	云南	16
129-144	欧当归	Levisticum officinale Koch	山西	16
145-159	华中前胡	Peucedanum medicum	云南	15

外，近红外技术结合化学计量方法PCA[12,13]、偏最小二乘判别分析[13,17]、ANN[14]、SVM[15]等能够实现当归产地、产期[18]及不同部位[17]的精细鉴别。本研究将NIRS技术与GBDT结合应用于当归药材真伪的快速鉴别，旨在建立一种快速、准确、便捷的当归真伪判别模型，规范当归药材市场。

1 材料与方法

1.1 样品来源

本试验中的样品分别于2017年10月至2017年12月期间采集，其中来自不同产地的正品当归5批次共112份，伪品当归云南野当归、欧当归、华中前胡各1批次，共计47份。

全部样品均由北京市临床药学研究所李莉研究员鉴定，其编号、基原植物、产地信息如表1所示，样品均保存于北京市临床药学研究所202实验室。

1.2 仪器与方法

1.2.1 仪器

NIR-M-R2型近红外光谱仪（扬光绿能），配备InGaAs检测器光谱采集范围900-1 700 nm，共228个变量，采用Hadamard模式进行扫描。

1.2.2 光谱采集

样品从头部1 cm处进行切割获得断面，将光纤探头垂直于样品头部断面获取近红外光谱信息。为减小实验误差，每个样品测定3次，将各次测定得到的光谱曲线加和取平均得到各样品的数据曲线用于后续数据分析。

1.2.3 方法学考察

按照样品测定方法操作，在同一背景下对同一样品进行6次扫描，求得228个波长下吸光度的标准偏差为0.000 56-0.004 3，光谱的均方差0.000 051-0.003 9，仪器稳定性良好。

取同一份样品，按照样品测定方法操作，采集其切割0、1、2、3、4、5、6 h后样品断面的光谱数据，求得不同吸收波长下RSD为0.007 9-0.87%，样品在6 h内相对稳定。

1.3 数据处理

1.3.1 预处理

由于仪器首尾噪声较大使得光谱前端和后端有较明显的噪声，为避免低信噪比对分析的影响，本实验截取961-1 655 nm波长范围内的光谱数据作为下一步分析的数据，以初步去除两端噪声干扰。

1.3.2 软件

本研究中PCA在PAST 3.0软件上完成，其余算法GBDT、SVM、ANN、随机森林（Random forest；RF）均由Python 2.7实现。

2 结果与分析

2.1 当归及伪品的近红外光谱曲线

本试验采集到的近红外光谱数据如图1所示，整体来看，当归及伪品的NIRS曲线变化趋势和特征吸收基本一致，无法直接鉴别当归真伪。从平均光谱图（图1b）可以看出，当归与华中前胡、欧当归、云南野当归样品的光谱曲线吸收率存在区别，平均吸收率大小依次为华中前胡＞云南野当归＞欧当归＞当归，这一差异为当归真伪的鉴别奠定了数据基础。但是由于仪器测量的波长范围局限，更多地得到的是芳烃、甲基、亚甲基、次甲基、水、胺等的合频和倍频吸收峰，信号强度低且峰谱宽，同一波段是样品多种信息的叠加，谱峰重叠严重使得对当归及伪品的光谱进行直接分辨较为困难，因此有必要借助化学计量学算法做进一步分析。

2.2 PCA定性分析

961-1 655 nm间有198个变量，数据量大且相邻波段之间的相关性强，造成信息的冗余，选用适当的方法剔除不相关变量十分必要。PCA目的是降维，消除相互重叠的信息部分，实现用少数关键变量代替全光谱，降低模型运算量和复杂度、提高模型稳定性和预测准确性[19]。PCA通过提取198个指标相关矩阵内部相关信息，剔除原始数据中高度冗余的变量，使数目较少的新变量成为原变量的线性组合，而且新变量能最大限度的表征原变量的数据结构特征[20]。

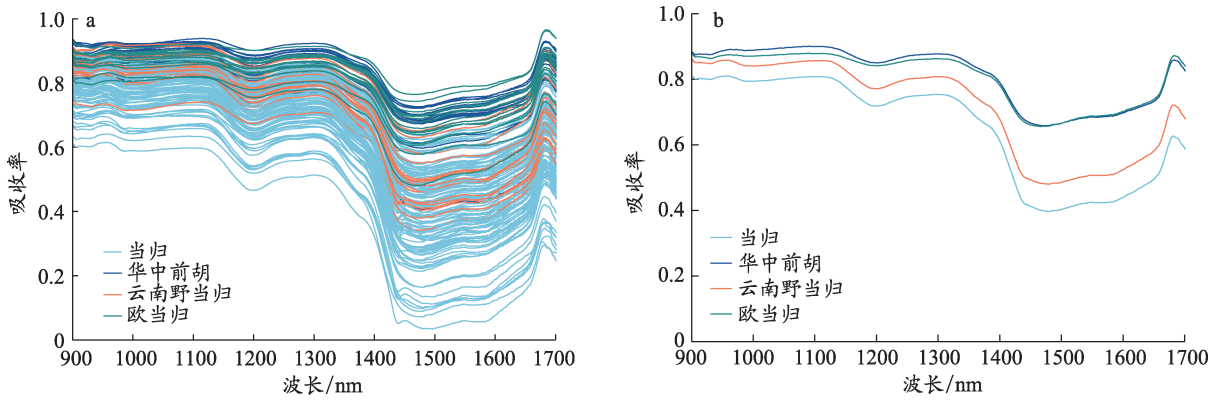

图1　当归、华中前胡、欧当归、云南野当归的近红外原始光谱(1a)和平均光谱(1b)

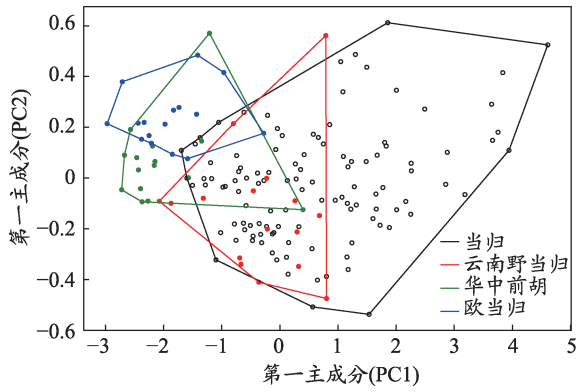

图2　当归、华中前胡、欧当归、云南野当归主成分分析得分图

表2　三种算法的性能比较

算法	准确率	召回率	*F*度量	建模时间/s
GBDT	0.94	0.85	0.89	65
SVM	0.90	0.82	0.86	55
ANN	0.88	0.78	0.85	83

对所有样品的光谱数据进行主成分分析可知，第一主成分(Principle Component 1，PC1)和第二主成分(PC2)的贡献率分别为97.78%、1.99%，表明PC1和PC2已能够表达99%以上的原始光谱信息。由图2可知，当归与华中前胡、欧当归、云南野当归主成分二维投影图中存在交叉无法有效分类，这是因为当归与伪品均为伞形科植物，化学成分较相似，其中当归与云南野当归重叠部分最多，与二者同属当归属亲缘关系较华中前胡、欧当归近有关。若要进一步准确鉴别当归及伪品需对光谱数据进行进一步处理。

2.3　建立GBDT判别模型

2.3.1　GBDT、SVM、ANN模型比较

GBDT算法由Jerome Friedman [21,22]于2001年提出，可用于分类和回归，通过集成多个弱学习器CART回归树最终组合成一个强学习器。GBDT每一次迭代是为了减少上一个模型的拟合残差，并在残差减少的梯度方向上建立新的CART回归树。本文所建判别模型基于Python－Sklearn工具包实现，实验中所用的计算机配置为Intel Core-i3处理器，2.2GHz主频，4 GB内存。GBDT参数采用网格搜索方法最终寻找到的最优参数为：树数量为1500、学习率0.01、最大深度为6、一阶正则项系数为0.3、二阶正则项系数为0.4、损失函数为交叉熵损失函数。

为验证提出的分类模型的优越性，将该模型与SVM、ANN构建的模型进行比较。同样采用网格搜索方法寻找最优参数，最终SVM分类模型参数为：核函数为RBF(径向基核函数)，核函数系数为0.1，惩罚项系数C为100，最大迭代次数为120次；ANN分类模型参数为：输入层神经元228个、隐藏层神经元10个、输出层神经元4个、学习率0.4、激活函数为sigmoid函数。

将正品当归作为正类，其它为负类，分类器在测试数据集上的预测或正确或错误，设置4种情况出现的样本数量如下：TP将正类预测为正类的数量，FN为将正类预测为负类的数量，FP为将负类预测为正类的数量，TN为将负类预测为负类的数量。这四个量可以导出几个重要的量化评估指标——准确率、召回率以及F度量，用于评价分类算法的性能。

$$\text{精确度：}P=\frac{TP}{TP+FP}$$

$$\text{准确率：}A=\frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{召回率：}R=\frac{TP}{TP+FN}$$

表3 GBDT分类模型准确率

实际分组	判别分组									
	训练集					测试集				
	当归	华中前胡	欧当归	云南野当归	准确率/%	当归	华中前胡	欧当归	云南野当归	准确率/%
当归	75	0	0	1	98.68	35	0	2	0	94.59
华中前胡	0	10	0	0	100.00	0	5	0	0	100.00
欧当归	0	3	7	0	70.00	0	1	4	0	80.00
云南野当归	0	2	0	9	81.82	2	0	0	3	60.00
总体准确率%	94.39%					90.38%				

表4 SVM分类模型准确率

实际分组	判别分组									
	训练集					测试集				
	当归	华中前胡	欧当归	云南野当归	准确率/%	当归	华中前胡	欧当归	云南野当归	准确率/%
当归	71	0	1	4	93.42	34	0	0	3	91.89
华中前胡	1	9	0	0	90.00	1	4	0	0	80.00
欧当归	0	0	9	1	90.00	0	1	3	1	60.00
云南野当归	3	0	0	8	72.72	1	0	1	3	60.00
总体准确率%	90.65%					84.61%				

表5 ANN分类模型准确率

实际分组	判别分组									
	训练集					测试集				
	当归	华中前胡	欧当归	云南野当归	准确率/%	当归	华中前胡	欧当归	云南野当归	准确率/%
当归	71	0	3	2	93.42	33	1	0	3	89.19
华中前胡	0	8	0	2	80.00	0	4	0	1	80.00
欧当归	0	0	8	2	80.00	0	1	3	1	60.00
云南野当归	2	0	1	8	72.72	0	0	2	3	60.00
总体准确率%	88.78%					82.69%				

$$F\text{度量}:F1=\frac{2PR}{P+R}$$

3种算法的预测结果见表2，梯度提升决策树无论是精确度、准确率、召回率还是F值都比另外两种算法要好，由于需要迭代生成很多棵树，所以训练模型的时间略长于SVM；SVM效果次之，并且在训练模型的过程中只需要寻找惩罚项系数、核函数类型、核函数系数、迭代次数等几个参数，所以耗时比另外两个模型要短；ANN得到的效果最差，这是因为神经网络模型需要大量的数据样本做支撑，从而用来训练模型参数、学习各个特征之间的相关关系，而本课题的样本量较小，所以导致训练效果最差，并且ANN需要寻找的网络参数量比较大，导致耗时最长。

随机抽取76份当归、10份华中前胡、10份欧当归、11份云南野归作为训练集建立判别模型，剩余的37份当归、5份华中前胡、5份欧当归、5份云南野归作为测试集以评估所建模型的性能，GBDT鉴别结果如表3所示，所建立的分类模型对当归的真伪有较好的鉴别效果。训练集与测试集总体判别率分别为94.39%、90.38%，其中GBDT模型对华中前胡的识别率达到100%。GBDT识别模式下当归部分样品发生误判，但单组判别准确率也大于90%，可见利用GBDT建立的分类模型能够有效鉴别当归与其混伪品。

2.3.2 RF优化模型

特征选择是指从原始特征集中选择使某种评估标准最优的特征子集，以使在该最优特征子集上所构建的分类或回归模型达到与特征选择前近似甚至更好的预测精度，同时筛除低信息量的冗余特征，达到降低训练模型所需时间、增强模型鲁棒性的目的。RF是Leo Breiman于2001年将Bagging集成学习理论与随机子空间方法相结合而提出的一种机器学习算

表6　特征光谱判别模型准确率

实际分组	判别分组									
	训练集					测试集				
	当归	华中前胡	欧当归	云南野当归	准确率/%	当归	华中前胡	欧当归	云南野当归	准确率/%
当归	74	1	1	0	97.37	34	1	0	2	91.89
华中前胡	0	9	0	1	90.00	0	4	0	1	80.00
欧当归	0	1	7	2	70.00	1	1	3	0	60.00
云南野当归	1	2	0	8	72.73	1	0	0	4	80.00
总体准确率%	91.59%					86.54%				

法[23-25]。RF具有准确度高、学习速度快、对噪声和异常值有较好的容忍性，对高维数据分类问题具有良好的可扩展性和并行性[26]。它集成多棵决策树的预测，在决策树构建过程中，树的每个结点都是以一定原则度量变量重要性，这一过程实际上就是一个特征选择过程[27]。

采用RF来度量各个特征波长的重要性，步骤如下：①从159个样本中随机有放回抽取N（本文设置为全量样本的70%，即112个）个样本，并且随机从198个特征波长中随机选择M（本文设置为总特征的40%，即79个）个特征波长，构成一个样本子集。重复此过程100次，得到100个样本子集；②对100个样本子集单独训练决策树模型，设置每棵决策树深度为6，不做任何剪枝操作，按照Gini指数最小原则进行特征分裂，直到该节点下的所有样本都属于同一类或者达到设置的最大深度；③将生成的100个决策树组成随机森林，按照多棵树分类器投票决定最终的分类结果。同时，统计生成每棵树时所使用的特征波长频次，累加求取均值后得到每个特征波长使用的频次，按照使用频次对198个特征进行从大到小排序。

最终，选择前20个频次高的特征波长作为最重要的特征子集，它们分别是：976 nm、1 016 nm、1 492 nm、1 511 nm、1 521 nm、1 528 nm、1 550 nm、1 573 nm、1 576 nm、1 580 nm、1 586 nm、1 598 nm、1 611 nm、1 621 nm、1 624 nm、1 636 nm、1 640 nm、1 646 nm、1 649 nm、1 655 nm。

近红外光谱振动倍频区有丰富的基团结构信息，一些含有C-H、N-H、O-H和S-H化学键的化合物会产生吸收，除在1 400 nm-1 800 nm之间产生一级倍频，往往还会分别在900 nm-1 200 nm和780 nm-900 nm谱带内产生二级倍频和三级倍频，反映的是中药化学成分的综合信息[28]。如本文中970 nm和1 450 nm附近的吸收峰主要是由于样本细胞中水对光谱吸收引起的，分别为O-H伸缩振动的二级倍频和一级倍频；在1 200 nm附近的吸收峰与N-H键有关。RF筛选出的特征波长均处于一级倍频、二级倍频区，且从图1可以看出，1 400 nm-1 655 nm范围内当归的吸光度与华中前胡、欧当归、云南野当归吸光度的差异较大有利于当归真伪的鉴别，因此将RF所筛选的20个特征波长用于建立特征光谱判别模型。

2.3.3　近红外特征光谱判别模型的建立

为了建立基于近红外特征光谱的当归真伪判别模型，将所分析出的20个特征波长作为GBDT的输入，所得模型的判别效果见表6。相比于原始光谱，特征光谱判别模型所用到的光谱变量大大减少，建模过程得到了简化，特征光谱所建模型判别准确率虽有所下降，但训练集与预测集的正确率仍均高于85%。对4类样本进行分析，当归单组训练集和测试集判别准确率分别达到了97.37%和91.89%，因此所建立的特征光谱判别模型也能够较好地实现当归的真伪鉴别。

3　讨论

在当归真伪判别研究中发现，PCA判别分析时区分效果不佳，这表明传统的线性模式识别方法PCA难以满足鉴别准确性的要求，需要采用更先进的模式识别相关理论和算法来提高近红外光谱技术的识别能力。本文采用GBDT、SVM、ANN三种非线性方法进行建模分析，识别准确率在训练集和测试集上均大于80%优于PCA。本研究结果显示近红外光谱技术能有效地识别当归及伪品光谱特征差异，并结合GBDT、SVM、ANN模式识别理论建立了判别模型，为当归及伪品鉴别提供了一种准确而快速的新方法。

近红外光谱虽然信息量大，但由于当归及其混伪品为近缘植物具多种相同成分，使得NIR光谱图非常相似，不能简单以峰位、峰形进行直接分类，选择合适的数据处理方法提取到特征信息也是分类鉴别的关

键。相较于NIR分析常用的建模方法ANN和SVM，本文尝试引入一种基于多特征 GBDT的分类方法，利用RF筛选变量并调整参数训练出最佳预测模型。通过3种模型判别结果的对比，可以看出，GBDT模型性能优良，具有较高的预测准确率和很好的适用性，可应用于当归的定性判别分析。然而，针对当归伪品，所建立的判别模型存在误判现象，分析其原因可能是由于伪品较难收集，本研究建立模型的伪品数量有限。有待于在今后实践中扩大校正集和预测集样本容量，完善数据库以优化模型。

本文旨在建立一个快速、简便、无损的当归定性判别模型，以药材断面进行光谱的采集较选择粉末简便、耗时短。目前已有文献报道采集枸杞子表面近红外漫反射光谱实现产地快速识别，基于主根横断面近红外光谱实现西洋参和人参的快速筛查[29,30]，以上证明采集药材的断面、表面光谱进行定性鉴别是可行的。此外，市场上存在一些贵重药材掺伪现象，如冬虫夏草、野山参、鹿茸、西洋参、川贝等，此类药材价格昂贵，充分发挥 NIR非破坏性的优势进行直接鉴别研究具有现实意义。我国幅员辽阔，中药品种繁多，“同名异物”和“同物异名”的现象依然存在，即使在今天，中药品种混淆的问题亦有出现，如“关木通”导致马兜铃酸肾病事件[31]，香港“白英”和“寻骨风”混淆导致病人患上肾衰竭和尿道癌[32]。如何运用现代科学的理论知识和技术方法来快速、简便、准确地鉴定中药品种，保证临床疗效，是一个迫切的课题。近红外光谱技术能够提中药品种识别的速度和识别能力，满足基层现场快速鉴别的需要。充分发挥近红外自身优势，通过对中药材的大样本量分析，建立稳健的近红外模型，结合云计算和互联网等现代手段，以在全国范围内建立近红外中药品种识别模型网络系统，应用于产地、加工炮制、运输、储存、流通各个环节，从而解决目前存在的品种混乱问题，对中药规范化管理具有重要意义。中药鉴定是一门与时俱进的学问，应在传统经验鉴别的基础上引入现代科学的理论知识和技术方法使中药鉴定更为快捷、科学，推动我国中药现代化进程。

参考文献

1 国家药典委员会. 中华人民共和国药典(一部). 北京: 中国医药科技出版社, 2015:133.

2 张瑛, 王亚丽, 潘新波. 当归历史资源分布本草考证. 中药材, 2016, 39(8): 1908-1910.

3 张山雷. 本草正义. 福州: 福建科学技术出版社, 2006: 55.

4 中华人民共和国国家卫生健康委员会. 卫生部关于进一步规范保健食品原料管理的通知, 2002.

5 桂镜生. 中药商品学. 昆明: 云南大学出版社, 2015: 137.

6 温子帅, 刘爱朋, 景松松, 等. 当归和欧当归的定性与定量鉴别. 中成药, 2018, 40(12): 2719-2723.

7 王翰华, 胡双丰. 当归及其常见伪品东当归和欧当归的鉴别. 中国医药指南, 2012, 10(5): 4-5.

8 李明. 当归及其混伪品的过氧化物同工酶电泳鉴别. 中药材, 2000 (4): 200.

9 罗沛宜, 罗茂, 朱烨, 等. 基于叶绿体trnL-F和rpoC1序列对当归及其混伪品的分子鉴定研究. 中国药学杂志, 2015, 50(10): 840-845.

10 Magwaza L S, Naidoo S I M, Laurie S M, *et al*. Development of NIRS models for rapid quantification of protein content in sweetpotato [Ipomoea batatas (L.) *LAM.J. LWT-Food Science and Technology*, 2016, 72:63-70.

11 胡咏川, 田晓鑫, 刘蕾, 等. 近红外光谱技术鉴定中药的进展. 中国中药志, 2012, 37(8): 1066-1071.

12 巩晓宇, 邱双凤, 彭炜, 等. 近红外光谱法快速鉴别苦参饮片的真伪. 中国医院用药评价与分析, 2016, 16(7): 883-885.

13 相翠玉, 宋晓勇. 中草药大黄的近红外光谱和人工神经网络鉴别分析. 时珍国医国药, 2012, 23(12): 3168-3169.

14 张丹雁, 刘家水, 陈奕龙, 等. 近红外光谱一致性检验及相关系数法快速鉴别南板蓝根真伪优劣. 时珍国医国药, 2014, 25(3): 627-629.

15 钟玉兰, 乐智勇. 利用近红外光谱技术快速鉴别三七粉及其伪品. 江西中医药大学学报, 2018, 30(3): 70-73,76.

16 张亚亚, 顾志荣, 丁军霞, 等. 当归不同药用部位近红外漫反射光谱指纹图谱研究. 中药材, 2015, 38(7): 1413-1416.

17 李波霞, 魏玉辉, 席莉莉, 等. 近红外光谱和化学计量学对不同产地不同产期当归的定性研究. 光谱实验室, 2011, 28(4): 2128-2134.

18 Eriksson L, Johansson E, Kettaneh-Wold N, *et al*. Multi-and megavariate data analysis part 1: Basic principles and applications. Umea, *Sweden: Umetrics AB*, 2006.

19 Costache G N, Coloran P. Combining pca-based datasets without retraining of the basis vector set. *Pattern Recogn Lett*, 2009, 30(16): 1441-1447.

20 Friendman J H. Stochastic gradient boosting. *Comput Stat Data An*, 2002, 38(4): 367-378.

21 Friendman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Stat*, 2000, 28(2): 337-374.

22 Breiman L. Random forests. *Mach Learn*, 2001, 45(1): 5-32.

23 Kwok S W, Carter C. Multiple decision trees. *Machine Intelligence and*

Pattern Recognition, 1990, 9: 327-335.
24 Ho T K. The random subspace method for constructing decision forests. *IEEE T Pattern Anal*, 1998, 20(8): 832-44
25 王奕森, 夏树涛. 集成学习之随机森林算法综述. 信息通信技术, 2018, 12(1): 49-55.
26 Genuer R, Poggi J M, Tuleaumalot C. Variable selection using random forests. *Pattern Recogn Lett*, 2010, 31(14): 2225-2236.
27 陆婉珍. 现代近红外光谱分析技术. 北京: 中国石化出版社, 2000.
28 王灵灵, 黄亚伟, 戚淑叶, 等. 基于主根横断面近红外光谱的西洋参和人参鉴别研究. 光谱学与光谱分析, 2012, 32(4): 925-929.
29 杜敏, 巩颖, 林兆洲, 等. 样品表面近红外光谱结合多类支持向量机快速鉴别枸杞子产地. 光谱学与光谱分析, 2013, 33(5): 1211-1214.
30 高月, 肖小河, 朱晓新, 等. 马兜铃酸的毒性研究及思考. 中国中药杂志, 2017, 42(21): 4049-4053.
31 Liang Z T, Jiang Z H, Leung K S Y, *et al*. Authentication and differentiation of two easily confusable Chinese material medica: Herba Solani Lyrati and Herba Aristolochiae Mollissimae. *J Food Drug Anal*, 2006(141): 36.

Identification of *Angelica Sinensis* and Its Adulterants by NIRS and GBDT

Gong Jianting[1,2], Li Li[1,2], Zou Huiqin[3], Xu Dong[3], Wang Daqian[1,2], Cong Yue[1,2], Liu Changli[4]

(1. Beijing Institute of Clinical Pharmacy, Beijing 100035, China; 2. Beijing Hospital of Traditional Chinese Medicine Affiliated to Capital Medical University, Beijing 100010, China; 3. School of Chinese Pharmacy, Beijing University of Chinese Medicine, Beijing 102488, China; 4. School of Traditional Chinese Medicine, Capital Medical University, Beijing 100069, China)

Abstract: Objective To identify *Angelica sinensis* and its adulterants by NIRS. Methods The near-infrared spectra of *Angelica sinensis* were collected and further analyzed by PCA which is one kind of pattern recognition. Then three different classifiers, namely GBDT, SVM and ANN, were employed to establish discriminative models. Afterwards, an optimized model was screened out by using RF filter characteristic wavelength optimization on the basis of GBDT mode. Results PCA could not distinguish *Angelica sinensis* and its adulterants effectively. Compared to SVM and ANN, GBDT creates better identification model. It showed higher accuracy, and the overall accuracy rate of training set and prediction set was 94.39% and 90.38%, respectively. Furthermore, 20 characteristic wavelengths were extracted by RF and re-establish the *Angelica sinensis* authenticity characteristic identification model, the overall accuracy rate of the training set and prediction set was 91.59% and 86.54%. Conclusion An identification model of *Angelica sinensis* and its adulterants is built based on NIR and GBDT, which provides a new method for traditional Chinese medicine non-invasive identification.

Keywords: *Angelica sinensis*, Gradient Boosting Decision Tree, Near infrared spectroscopy, Pattern recognition, Discriminative model, Identification

（责任编辑：周阿剑，责任译审：邹建华）