高光谱成像技术结合特征波长优化对苍术颗粒剂生产厂家的可视化判别研究

黄 晔[1]，刘 丽[2]，梁 晶[2]，杨红霞[3]，李晓丽[4]，徐 宁[2*]

1. 浙江省皮肤病防治研究所药剂科，浙江 德清 313200
2. 浙江工业大学药学院，浙江 杭州 310014
3. 湖州市食品药品检验研究院，浙江 湖州 313000
4. 浙江大学食品及生物工程学院，浙江 杭州 310058

摘 要 为了给苍术颗粒剂基于高光谱成像的可视化区分提供理论指导，选用竞争性自适应重加权采样法(CARS)和相关性分析(CA)进行两次特征波长选择，提出了利用近红外高光谱成像技术对苍术颗粒剂产品溯源的新方法。874～1 734 nm波段范围内采集150个来自三个生产厂家的苍术颗粒剂高光谱图像，提取感兴趣区域(ROI)的光谱反射率值作为鉴别模型的输入变量，采用邻近算法(KNN)、误差反向传输神经网络(BPNN)、偏最小二乘法判别分析(PLS-DA)、最小二乘支持向量机(LS-SVM)建立四种算法(分类器)的判别模型。通过对模型效果的评价标准(预测集总体判别率以及kappa系数)来判别三个不同厂家苍术颗粒剂的区分效果。除KNN模型外，预测集的判别率都是100%，kappa系数均为1。为了加快运算速度，研究通过CARS、随机蛙跳算法(RF)、连续投影算法(SPA)和序列前向选择(SFS)算法初步选择特征波长；采用CARS，RF，SFS和SPA结合CA算法取得了4组最优波长。分别得到4个(975，1 220，1 419，1 476 nm)、2个(1 005，1 442 nm)、4个(924，1 005，1 419，1 584 nm)和3个(948，1 146，1 412 nm)最优波长，并分别建立了KNN，BPNN，PLS-DA和LS-SVM判别模型。在筛选三种最优算法的情况下，能够以较少的特征波长个数获得的最好建模效果为：CARS-CA-LS-SVM模型中预测集总体判别率是100%，kappa系数为1。将CARS-CA筛选出波长变量的每个像素点光谱数据输入到LS-SVM模型中，将判别结果用不同颜色直观显示。该研究为快速无损进行苍术颗粒剂产品溯源提供了思路，为今后开发相关机构的快速监管提供了技术支持。

关键词 高光谱成像；苍术颗粒剂；化学计量学；特征波长

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)11-3567-06

引 言

中药配方颗粒因便于服用、保存和携带等优点，已经广泛流通于全国医院。目前已获得各类中药配方颗粒试点资质的企业高达57家，然而颗粒剂的生产工艺及质量控制还没有建立统一的国家标准。同一品种因生产企业不同仍存在较大差异。苍术为菊科植物茅苍术或北苍术的干燥根茎，具有燥湿健脾、祛风散寒、明目等功效[1]。苍术颗粒剂无法像饮片一样通过性状、显微鉴别区分品种和产地信息，亦无法通过高效液相色谱法对苍术素有效定量。主要采用薄层色谱法(thin layer chromatography，TLC)、水分测定法、浸出物测定法等，对其有效成分与含量进行初步的质量检测，因此开拓新的有效的质量控制方法具有重要意义。

高光谱成像技术具有高效准确，无损、无污染的优点[2-3]。国内外学者已尝试通过高光谱成像技术应用于中药材的质量控制，包括中药年份鉴别、品种鉴别等。有报道研究了高光谱技术结合化学计量学对不同年份及放置方式的陈皮建立鉴别模型，预测准确率达98.33%，为陈皮年份无损鉴别提供了新的技术参考。Tankeu等[4]利用高光谱成像技术结合偏最小二乘判别分析，区分出粉防己和广防己这两种外观形态相似、实则来源不同科属、化学成分截然不同的两

收稿日期：2019-11-11，**修订日期**：2020-03-05

基金项目：国家自然科学基金项目(31771676)和湖州市科技计划项目(2018GY44)资助

作者简介：黄 晔，1982年生，浙江省皮肤病防治研究所副主任药师 e-mail：497042175@qq.com

*通讯联系人 e-mail：xuning@zjut.edu.cn

种植物，可以有效防止广防己混入粉防己。大多数的研究集中于简单区分与防伪，探索高光谱成像技术应用于中药现代化快速无损分析的过程中，如何优选样品光谱特征波段，建立准确度高、预测效果好的模型仍是亟需解决的问题。仅李超等[5](Fourier transform infrared spectroscopy，FTIR)对国内 8 省份 18 产区的苍术样品建立了红外指纹图谱，而高光谱成像的苍术配方颗粒研究尚无报道。

本工作研究了在高光谱 874～1 734 nm 区域 3 个不同生产厂家的苍术配方颗粒图谱信息，结合 9 种波段数据分别建立 4 种判别模型来寻找潜在的信息，对不同厂家的苍术颗粒剂进行快速区分，并将结果可视化。

1 实验部分

1.1 材料和仪器

浙江惠松制药有限公司(厂家 A，浙江杭州)、江阴天江药业有限公司(厂家 B，江苏江阴)、华润三九医药股份有限公司(厂家 C，广东深圳)的苍术配方颗粒各 50 份，共 150 份样本。

高光谱成像系统主要包括分辨率为 672×512 的 CCD 相机(C8484-05，Hamamatsu Photonics，Hamamatsu City，Japan)，分辨率为 2.8 nm 的光谱仪(ImSpector N17E；Spectral Imaging Ltd，Oulu，Finland)，线光源(Fiber-Lite DC950，Dolan Jenner Industries Inc，Boxborough，MA)，计算机，暗箱和电控移动平台。高光谱图像采集前，首先获取暗电流和参考板的高光谱图像数据，用于数据处理前对原始高光谱图像的校正。电控移动平台移动速度为 17 mm·s^{-1}，工作距离为 20.5 cm，曝光时间 2.4 ms，采集在 874～1 734 nm 范围样本的高光谱信息。

1.2 薄层色谱

薄层色谱法步骤参照《中国药典》2015 年版四部 通则 0502。

1.3 光谱数据及图像处理

苍术颗粒剂样本分别置于 96 孔板中，保证每个样本在同一高度。黑白校正后设置样本区域作为感兴趣区域(region of interest，ROI)，计算出每个样本 ROI 范围内 874～1 734 nm 的平均光谱。采用偏最小二乘判别分析(partial least square discrimination analysis，PLS-DA)、最小二乘支持向量机(least-squares support vector machine，LS-SVM)、反向人工神经网络(back propagation neural network，BPNN)、邻近算法(k-nearest neighbor，KNN)、竞争性自适应重加权采样法(competitive adaptive reweighted sampling，CARS)、随机蛙跳算法(random frog，RF)[6]、连续投影算法(successive projections algorithm，SPA)、序列前向选择算法(sequential forward selection，SFS)[7]以及相关性分析(correlation analysis，CA)采用 Matlab R2018a(The Math Works，Natick，USA)处理。

2 结果与讨论

2.1 苍术颗粒剂的薄层色谱鉴定及平均高光谱

对应苍术对照药材薄层色谱的相同位置，各样品的荧光斑点颜色一致，见图 1(a)。虽然三个样品的薄层色谱有些许差别，但不能确认样品的生产商。

苍术颗粒剂高光谱敏感波段大都集中在 1 100～1 650 nm 附近，见图 1(b)。1 100～1 300 nm 归属于 C—H 伸缩振动的二级倍频[8]，1 300～1 400 nm 归属于 C—H 伸缩振动的组合带[9]，1 450 nm 归属于 O—H 伸缩振动的一级倍频，和苍术颗粒剂中存在的水分有关[8]，1 480 nm 附近归属于 O—H 伸缩振动的二级倍频[9]。1 250～1 680 nm 含有的信息和苍术颗粒剂中的氨基酸有关。

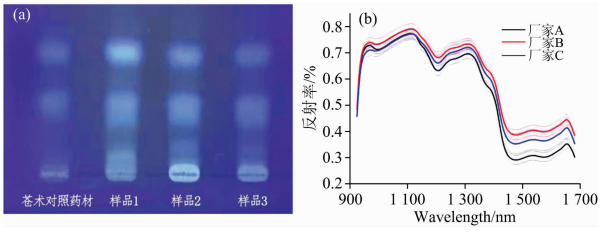

图 1 苍术颗粒剂(a)薄层色谱鉴定结果，(b)平均高光谱

样品 1，2 和 3 生产商分别为华润三九、江阴天江、浙江惠松

Fig. 1 (a) The TLC result and (b) average hyperspectral of *Atractylodes Lancea* granules

Sample 1，2，and 3 were manufactured from Huarun Sanjiu，Jiangyin Tianjiang，and Zhejiang Huisong respectively

2.2 不同厂家苍术颗粒剂全波段建模区分及特征波长优化

2.2.1 CARS 及 RF

基于全波段所建立的判别模型，KNN 模型的总体判别率为 96%，Kappa 系数为 0.937 8。BPNN，LS-SVM 以及 PLS-DA 模型判别率都为 100%，Kappa 系数为 1。

随样本运行次数增加，所选特征波长数目先迅速减少，随后趋于平缓，见图 2(a)，表明在 CARS 中已经实现了快速选择、双阶段选择及精选选择。由图 2(b)可以看见，随样本运行次数增加，消除了部分冗余信息后 RMSECV 开始缓慢减少；当样本运行次数为 25 次后，消除了部分关键波长下的光谱信息 RMSECV 开始缓慢增长；图 2(c)中“*”线表示 RMSECV 达最低值之最佳点，当样本运行 25 次时，RMSECV 值最小，获得 19 个特征波长。RF 方法可检测每个波长下高光谱信息对不同厂家苍术颗粒剂区分的重要性。波长被选择几率越大，表明该波长下高光谱信息与不同厂家的苍术颗粒剂样本区分的相关性可能较大[见图 2(d)]。基于 RF 方法将波长被选择的概率从大到小排列，筛选出前 10 个波长组成波长数目从 1～10 的 10 组数据，并建立判别模型[图 2(e)和图 2(f)]。随着波长数目的增加，模型的总体判别率以及 Kappa 系数总体均呈上升趋势。当波长数目大于等于 5 时，除 KNN 外的另外三种模型总体判别率均达到了 100%、Kappa 系数达到了 1 且保持恒定。

2.2.2　相关性分析及优化的波段

继续计算选择出的两个敏感波长之间的皮尔森相关系数，两个波长相关系数的值高于 0.9 的，只保留一个。

经 CARS-CA，RF-CA，SFS-CA 以及 SPA-CA 分别筛选得到了 4 个、2 个、4 个以及 3 个最佳波长(表 1)。

954 nm 归属于 C—H，N—H，O—H 伸缩振动的三级倍频[10]，975，1 476 和 1 483 nm 归属于 O—H 伸缩振动的二级倍频[11]，1 005 nm 归属于 N—H 伸缩振动的二级倍频[9]，1 122 nm 归属于 C—H 伸缩振动[12]，1 220，1 126，1 146，1 237，1 294，1 348，1 365 和 1 368 nm 为 C—H 的伸缩振动的二级倍频[13]，1 372 nm 归属于 1 412 nm，1 415 nm 归属于芳香烃的 C—H 拉伸振动[14]。

CARS 选择的特征波长，在 1 100～1 300，1 360～1 420 以及 1 430～1 480 nm 位置，和相关性分析结果一致。CARS-CA，RF-CA，SFS-CA 和 SPA-CA 得到的最优波长分别有 2 个(1 220 和 1 476 nm)，1 个(1 442 nm)，1 个(1 584 nm)、1 个(1 146 nm)，均分布在对应的平均光谱差异度较大的区域，见图 3(a,b,c)。其中 1 476，1 442 与 1 584 nm 的信息都和苍术颗粒剂中的氨基酸有关。

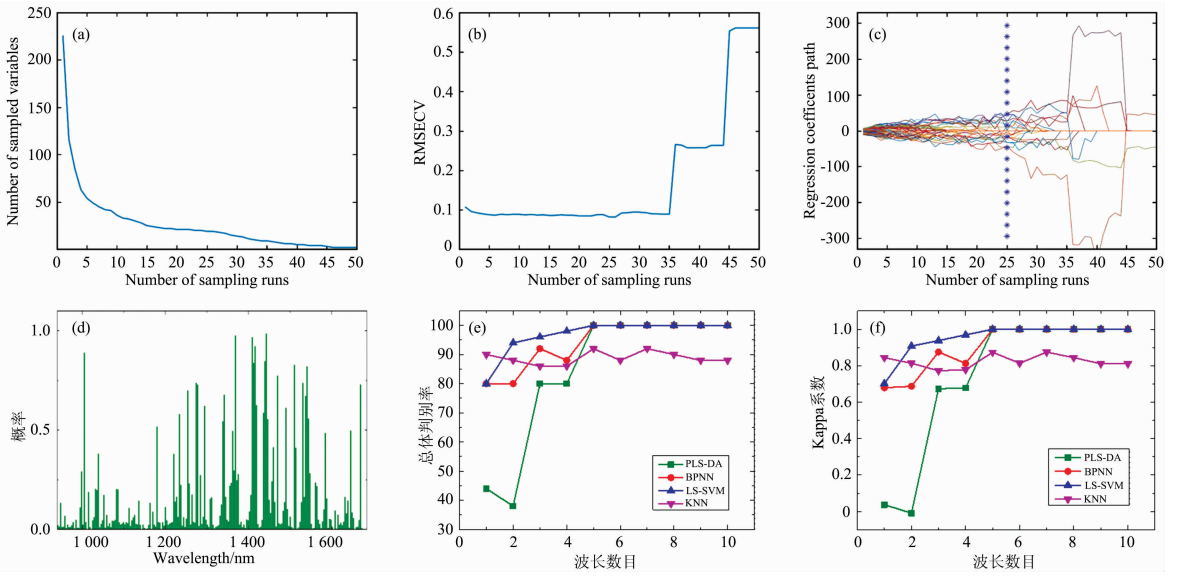

图 2　(a)CARS 采样变量数量的变化趋势，(b)RMSECV 值，(c)随着采样运行的增加每个变量的回归系数，(d)通过 RF 选择特征波长的结果，不同波长数目下不同模型的(e)总体判别率和(f)Kappa 系数

Fig. 2　(a) The changing trend of the number of sampled variables, (b) RMSECV values, (c) regression coefficients of each variable with the increasing of sampling runs of CARS, (d) results of selected characteristic wavelengths by RF, (e) discriminant rate and (f) kappa coefficient in different models based on different wavelengths

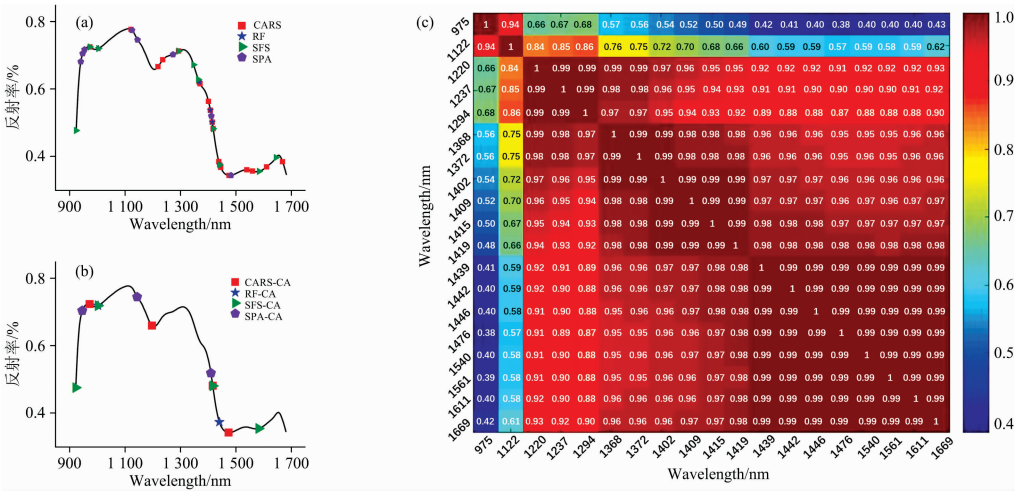

图 3　苍术颗粒剂厂家区分研究中(a)初步筛选特征波长，(b)CA 筛选波长选择结果，(c)基于 CARS 选择的敏感波长之间的相关性分析

Fig. 3　Characteristic wavelength selection results based on (a) preliminary selection and (b) CA method and (c) correlation analysis among characteristic wavelengths selected by CARS in the distinguish study of *Atractylodes Lancea* granules from different manufactures

表 1 基于高光谱技术的苍术颗粒剂厂家区分特征波段选择

Table 1 Selected characteristic wavelengths in the distinguish study of *Atractylodes Lancea* granules from different manufactures based on hyperspectral technology

波长选择方法	数量	特征波长/nm
CARS	19	975，1 122，1 220，1 237，1 294，1 368，1 372，1 402，1 409，1 415，1 419，1 439，1 442，1 446，1 476，1 540，1 561，1 611，1 669
CARS-CA	4	975，1 220，1 419，1 476
RF	5	1 005，1 368，1 409，1 415，1 442
RF-CA	2	1 005，1 442
SFS	10	924，975，1 005，1 294，1 348，1 365，1 419，1 442，1 584，1 645
SFS-CA	4	924，1 005，1 419，1 584
SPA	8	941，948，954，1 126，1 146，1 274，1 412，1 483
SPA-CA	3	948，1 146，1 412

2.3 基于特征波长建立的判别模型

基于 CARS-CA，RF-CA，SFS-CA 和 SPA-CA 选择的最优波长，和原始数据的 256 个波长变量相比，分别减小了 98.44%，99.22%，98.44%和 98.83%的变量，大大增加了模型的运算效率。

从表 2 可知，基于四种最佳波段建立的 KNN 以及 BPNN 模型的总体判别率均没有达到 100%，Kappa 系数也没有达到 1。基于四组最佳特征波长建立的 PLS-DA 以及 LS-SVM 模型的总体判别率为 100%以及 Kappa 系数为 1 的占比分别为 25%以及 50%，可以得出 LS-SVM 模型判别效果优于其他三种。基于 CARS-CA 所建立的四种模型，总体判别率为 100%以及 Kappa 系数为 1 的占比为 50%，优于其他三组最佳特征波长。综上所述，CARS-CA-LS-SVM 模型在总体判别率为 100%以及 Kappa 系数为 1 的情况下，大大减少了模型的输入变量，提高了运算效率，为区分不同厂家苍术颗粒剂的最优模型。

表 2 基于特征波长建立的区分不同厂家苍术颗粒剂的模型判别

Table 2 Model discrimination based on characteristic wavelengths in the distinguish study of *Atractylodes Lancea* granules from different manufactures

波长选择方法	波长数目	KNN		BPNN		PLS-DA		LS-SVM			
		总体判别率/%	Kappa 系数	总体判别率/%	Kappa 系数	总体叛别率/%	Kappa 系数	γ	δ^2	总体判别率/%	Kappa 系数
CARS	19	90	0.84	100	1	100	1	1.904×10^2	0.289 3	100	1
CARS-CA	4	94	0.91	98	0.97	100	1	1.198×10^3	0.289	100	1
RF	5	88	0.68	100	1	100	1	4.083×10^3	0.025 05	100	1
RF-CA	2	86	0.78	78	0.67	82	0.72	4.08×10^3	0.025 0	86	0.79
SFS	10	90	0.84	100	1	100	1	3.664×10^5	1.412	100	1
SFS-CA	4	90	0.84	98	0.97	68	0.50	3.66×10^5	1.41	100	1
SPA	8	94	0.91	98	0.97	100	1	6.889×10^3	0.509 0	100	1
SPA-CA	3	94	0.91	64	0.42	86	0.78	6.89×10^3	0.509	94	0.91

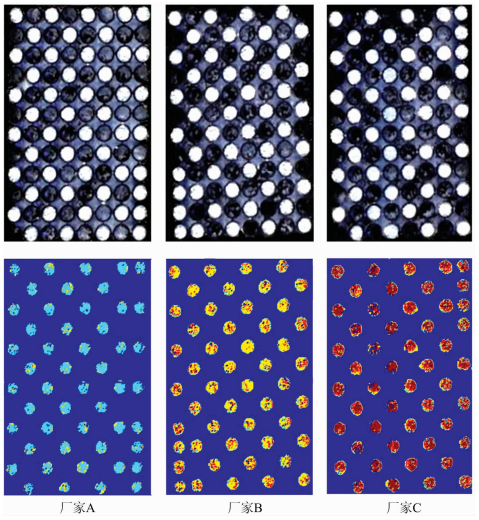

图 4 基于 CARS-CA-LS-SVM 模型的不同厂家苍术颗粒剂区分结果图

Fig. 4 The distinguish result map of *Atractylodes Lancea* granules from different manufacturers based on CARS-CA-LS-SVM model

2.4 基于最优模型建立的判别模型结果可视化

所有的苍术颗粒剂样本都能被正确识别(如图 4 所示)，并且很容易与其他厂家区分。然而，厂家 B 的苍术颗粒剂样本有 4 个样本的一些像素点被预测成了厂家 C，本来应该被预测为黄色的一些像素点被预测成了红色。其原因可能是该像素点的光谱携带了超出厂家 C 范围的苍术颗粒剂信息，当使用基于样本平均谱的模型来预测相应的像素谱时，这些光谱特征会偏离样本平均谱的预测集。整体来看，4 个样本所有像素点的颜色，还是预测为厂家 B 黄色的像素点较多，是准确的。

不同厂家苍术颗粒剂的区分可视化是基于最优模型和特征波长建立的鲁棒性和代表性的判别模型，结果证明是可行的。为今后开发苍术颗粒剂和其他中药的综合质量实时监测系统提供了可能。

3 结 论

高光谱技术结合 CARS 和 CA 法进行二次特征波长选择，可有效实现不同厂家的苍术颗粒剂可视化判别，实现了

三个不同厂家苍术颗粒剂的区分。剔除不相关或非线性变量的输入变量，降低了计算量，提高了运算效率，为解决苍术颗粒剂的无损质量控制和产品溯源问题提供了参考。为开发基于特征波长的中药制剂产品溯源多光谱检测系统提供了科学支持。

下一步将进行更多生产厂家区分以研究参数的有效性，并拟扩大样本数做进一步验证和完善以建立更稳定、更普遍适用的判别模型。

References

[1] Wei Y, Sui D J, Xu H M, et al. Chinese Journal of Natural Medicines, 2017, 15(12): 905.
[2] Shi C, Qian J P, Zhu W Y, et al. Food Chemistry 2019, 275: 497.
[3] Zhou X, Sun J, Wu X H, et al. Spectrochimica Acta Part A: Molecular & Biomolecular Spectroscopy, 2019, 206: 378.
[4] Tankeu S, Vermaak I, Chen W Y, et al. Phytochemistry, 2016, 122: 213.
[5] LI Chao, HUANG Xian-zhang, ZHANG Chao-yun, et al(李 超, 黄显章, 张超云, 等). Journal of Chinese Medicinal Materials(中药材), 2019, 42(1): 51.
[6] Chu B Q, Yu K Q, Zhao Y R, et al. Sensors, 2018, 18(4): 1259.
[7] Liang J, Li X L, Zhu P P, et al. Applied Sciences, 2019, 9 (10): 2092.
[8] Zhao Y Y, Zhang C, Zhu S S, et al. Molecules, 2018, 23 (6): 1352.
[9] Zhang C, Liu F, He Y. Scientific Reports, 2018, 8: 2166.
[10] Weng H Y, Lv J W, Cen H Y, et al. Sensors and Actuators B: Chemical, 2018, 275: 50.
[11] Zhang C, Jiang H, Liu F, et al. Food and Bioprocess Technology, 2017, 10(1): 213.
[12] Zhao Y Y, Zhu S S, Zhang C, et al. RSC Advances, 2018, 8(3): 1337.
[13] Wu N, Zhang C, Bai X L, et al. Molecules, 2018, 23(11): 2831.
[14] Feng X P, Yu C L, Shu Z Y, et al. Fuel, 2018, 228: 197.

Research on the Visualization Differentiation of *Atractylodes Lancea* Granule Manufactures Based on Hyperspectral Imaging Technology Combined With the Selection of Characteristic Wavelengths

HUANG Ye[1], LIU Li[2], LIANG Jing[2], YANG Hong-xia[3], LI Xiao-li[4], XU Ning[2*]

1. Department of Pharmacy, Zhejiang Skin Disease Prevention and Treatment Center, Deqing 313200, China
2. College of Pharmacy, Zhejiang University of Technology, Hangzhou 310014, China
3. Huzhou Institute of Food and Drug Inspection, Huzhou 313000, China
4. College of Food and Bioengineering, Zhejiang University, Hangzhou 310058, China

Abstract In order to provide theoretical guidance for the visualization differentiation of *Atractylodes Lancea* granules based on hyperspectral imaging, competitive adaptive reweighted sampling (CARS) and correlation analysis (CA) was used to select two characteristic wavelengths. A new method for traceability of *Atractylodes Lancea* granules using near-infrared hyperspectral imaging technology was proposed. Hyperspectral image of 150 *Atractylodes Lancea* granules from three manufacturers in the range of 874～1 734 nm, extracting the spectral reflectance value of the region of interest (ROI) as the input variables for the identification model, and using the proximity algorithm (k-nearest neighbor, KNN), back-propagation neural networks (BPNN), partial least squares-discrimination analysis (PLS-DA) and least square support vector machine (LS-SVM) to establish discriminant models of four algorithms (classifiers). The discrimination effect of three different manufacturers of *Atractylodes Lancea* granules was discriminated by the evaluation criteria of the model effect (predictive set overall discriminant rate and kappa coefficient). Except for the KNN model, the discriminant rate of the prediction set was 100%, and the kappa coefficient was 1. In order to speed up the operation, this study selected the characteristic wavelengths by CARS, random frog (RF), successive projections algorithm (SPA) and sequential forward selection (SFS) algorithm, and used CARS, RF, SFS, and SPA combined with the CA algorithm to achieve four sets of optimal wavelengths. Four (975, 1 220, 1 419, 1 476 nm), two (1 005, 1 442 nm), four (924, 1 005, 1 419, 1 584 nm) and three (948, 1 146, 1 412 nm) optimal wavelengths were obtained respectively, and KNN, BPNN, PLS-DA, and LS-SVM discriminant models were established. Therefore, in the case of screening three optimal algorithms, the best modeling effect that can be obtained with fewer feature wavelengths was: the

overall discriminant rate of the prediction set in the CARS-CA-LS-SVM model was 100%, the kappa coefficient was 1. Finally, the spectral data of each pixel of the wavelength variables selected by CARS-CA were input into the LS-SVM model, and the discrimination results were visually displayed in different colors. This study provides a method for the rapid and lossless traceability of *Atractylodes Lancea* granules product, and provides technical support for the rapid supervision of related organizations in the future.

Keywords Hyperspectral imaging; *Atractylodes lancea* granules; Chemometrics; Characteristic wavelengths

(Received Nov. 11, 2019; accepted Mar. 5, 2020)

* Corresponding author

敬告读者——《光谱学与光谱分析》已全文上网

从 2008 年第 7 期开始在《光谱学与光谱分析》网站(www.gpxygpfx.com)“在线期刊”栏内发布《光谱学与光谱分析》期刊全文，读者可方便地免费下载摘要和 PDF 全文，欢迎浏览、检索本刊当期的全部内容；并陆续刊出自 2004 年以后出版的各期摘要和 PDF 全文内容。2009 年起《光谱学与光谱分析》每期出版日期改为每月 1 日。

《光谱学与光谱分析》期刊社