基于近红外漫反射光谱和 PCA-SVM 算法快速鉴别炉甘石

陈龙[1,2]，张晓冬[2]，孙扬波[2]，陈科力[2*]
（1. 襄阳市中心医院，湖北文理学院附属医院，湖北 襄阳 441021；
2. 湖北中医药大学 中药资源和中药复方教育部重点实验室，武汉 430065）

［摘要］ **目的**：利用主成分分析（PCA）和支持向量机（SVM）算法，建立炉甘石生品、伪品及炮制品的近红外漫反射光谱（NIRS）鉴别模型。**方法**：采集炉甘石生品、伪品及炮制品的 NIRS，选取特征谱段，优选光谱预处理方法及最佳主成分数，建立 PCA-SVM 鉴别模型。**结果**：在 7 500～4 000 cm^{-1} 谱段，以一阶导数法（FD）为最佳光谱预处理方法，PCA 提取的光谱前 5 个主成分为最佳主成分，并经网格搜索算法确定惩罚因子 $c=0.25$，核函数参数 $g=8$ 为最佳 SVM 内部参数，建立炉甘石 PCA-SVM 鉴别模型。该模型五折交叉验证准确率 100%，且模型对训练集和测试集样品预测正确率亦均达 100%。**结论**：基于 PCA-SVM 算法所建立的炉甘石 NIRS 鉴别模型预测准确率高，结合固体粉末漫反射技术无损、快速的优点，该模型可用于炉甘石生品、伪品及炮制品的无损、快速鉴别。

［关键词］ 炉甘石；近红外漫反射光谱；主成分分析；支持向量机；一阶导数法；网格搜索算法；五折交叉验证
［中图分类号］ R22；R28；R9；O657；O433；C37 ［文献标识码］ A ［文章编号］ 1005-9903（2019）18-0116-08
［doi］ 10.13422/j.cnki.syfjx.20190747
［网络出版地址］ http://kns.cnki.net/kcms/detail/11.3495.R.20181210.0943.001.html
［网络出版时间］ 2018-12-11 15:47

Rapid Identification of Calamina Based on Near-infrared Diffuse Reflectance Spectroscopy and PCA-SVM Algorithm

CHEN Long[1,2], ZHANG Xiao-dong[2], SUN Yang-bo[2], CHEN Ke-li[2*]
(1. *Xiangyang Central Hospital*, *Affiliated Hospital of Hubei University of Arts and Science*, *Xiangyang* 441021, *China*;
2. *Key Laboratory of Traditional Chinese Medicine Resource and Compound Prescription*, *Hubei University of Chinese Medicine*, *Wuhan* 430065, *China*)

［**Abstract**］ **Objective**: To establish a near-infrared diffuse reflectance spectroscopy (NIRS) identification model for crude products, counterfeit products and processed products of Calamina by principal component analysis (PCA) and support vector machine (SVM) algorithm. **Method**: NIRS of crude products, counterfeit products and processed products of Calamina were collected, the characteristic spectrum segments were selected, the preprocessing method and the optimum principal component number were optimized, and the PCA-SVM qualitative model was established. **Result**: The characteristic spectrum segment of analysis model was 7 500-4 000 cm^{-1}. Spectra were preprocessed by the first-order derivative method (FD). The optimum principal component number was 5. And the optimum internal parameters of SVM [penalty factor (c) =0.25 and kernel function parameter (g) =8] were screened by applying the grid search algorithm. In the PCA-SVM qualitative model, the prediction accuracy rate was 100% for the 5-fold cross validation, and the prediction accuracy rates

［收稿日期］ 20181019（015）
［基金项目］ 国家中药标准化项目（1399）；湖北中医药大学教育部重点实验室 2017 年度开放基金课题
［第一作者］ 陈龙，硕士，药师，从事中药资源及其品质研究，E-mail：chenlong2435@163.com
［通信作者］ *陈科力，教授，从事中药资源及其品质研究，Tel：027-68890106，E-mail：kelichen@126.com

also were 100% both for training set and test set. **Conclusion**: PCA-SVM analysis model of NIRS for Calamina samples has a high prediction accuracy rate, and it can be used for the rapid and nondestructive identification of crude products, counterfeit products and processed products of Calamina by combining the diffuse reflection technique on solid powder.

[**Key words**] Calamina; near-infrared diffuse reflectance spectroscopy; principal component analysis; support vector machine; first-order derivative method; grid search algorithm; 5-fold cross validation

炉甘石来源于碳酸盐类矿物菱锌矿($ZnCO_3$)或水锌矿[$Zn_5(CO_3)_2(OH)_6$][1],经高温煅烧,主要成分分解为氧化锌(ZnO),具有解毒明目退翳、收湿止痒敛疮的功效,是一种常用的外用药[2]。市售炉甘石质量问题突出,常见以主要成分为方解石($CaCO_3$)的矿石充伪炉甘石[3-4]。此外,炉甘石高温炮制后入药,炮制前后药效差异较大,不可混用[5]。因此,迫切需求用于炉甘石及其炮制品的真伪快检方法。

为解决上述问题,本课题组将近红外漫反射光谱(NIRS)技术与化学计量学算法结合,应用于炉甘石生品、伪品及炮制品的快速鉴别。如 SUN 等[6]采用反向传播人工神经网络(BP-ANN)建立炉甘石 NIRS 鉴别模型,获得 95% 的鉴别准确率。在此基础上,张晓冬等[7]进一步扩大样本量,并分别利用主成分分析(PCA)判别法和聚类分析法建立 NIRS 鉴别模型,其鉴别准确率分别为 94.34% 和 96.26%。上述算法均获得了较高的判别准确率,但在应用中均存在一定的缺点[7]。如 BP-ANN 建模中,网络参数优化随机性强[6];又如 PCA 为传统线性算法,对实际的非线性问题拟合能力有限。因此,具有非线性拟合能力且优化过程相对稳定的支持向量机(SVM)算法受到本课题组关注。明晶等[8]以 SVM 算法建立了琥珀的 NIRS 鉴别模型。此外,利用 NIRS 数据进行 SVM 建模时,常需对光谱数据进行降维,PCA 即为常用的降维算法。二者结合的 PCA-SVM 算法具有较好的建模能力,魏从师等[9]利用 PCA-SVM 算法建立了 6 种树脂类中药鉴别模型。基于此,在前期研究基础上[7],本实验拟利用 PCA-SVM 算法建立炉甘石的 NIRS 鉴别模型,探索进一步优化炉甘石生品、伪品和炮制品粉末的 NIRS 鉴别方法;同时,基于固体漫反射技术具有检测速度快、对粉末样品进行非接触式扫描而不破坏样品的优点,所建立的 NIRS 将实现对各类炉甘石粉末样品的快速、无损、准确鉴别。

1 材料

MPA 型傅里叶变换近红外光谱仪(德国 Bruker 公司,配备固体积分球漫反射附件和 OPUS 7.5 光谱采集和处理软件)。Matlab R2014a 软件(美国 MathWorks 公司)和 Unscrambler 9.7 软件(挪威 CAMO 公司)用于光谱数据分析和建模。炉甘石样品来源于湖南、广西、云南、四川、贵州等主要矿产区,以及亳州、禹州、安国等各大药材市场,部分样品由马应龙药业集团股份有限公司和湖北中医药大学矿物药标本馆提供;所有的样品均由湖北中医药大学生药教研室陈科力教授依据 2015 年版《中国药典》(一部)[2]进行鉴别,并结合 X 射线衍射(XRD)物相分析,判断样品真伪,具体样品信息见文献[7],共得到合格炉甘石生品样品(以下简称生品)42 批次(随机编号 1 ~ 22,63 ~ 82),主要成分为水锌矿,少数含菱锌矿;炉甘石伪品样品(以下简称伪品)35 批次(随机编号 23 ~ 42,83 ~ 97),主要成分为方解石、石英、白云石等,不含水锌矿或菱锌矿。从生品中随机挑选若干批次,按 2015 年版《中国药典》(一部)中炉甘石的炮制方法进行炮制,炮制后的样品亦进行 XRD 物相分析,共获得合格炉甘石炮制品(以下简称炮制品)38 批(随机编号 43 ~ 62,98 ~ 115),主要成分为 ZnO。将上述生品、伪品和炮制品(共计 115 批)随机分为训练集(62 批,编号1 ~ 62)和测试集(53 批,编号 63 ~ 115)。样品分类信息见表 1。

2 方法与结果

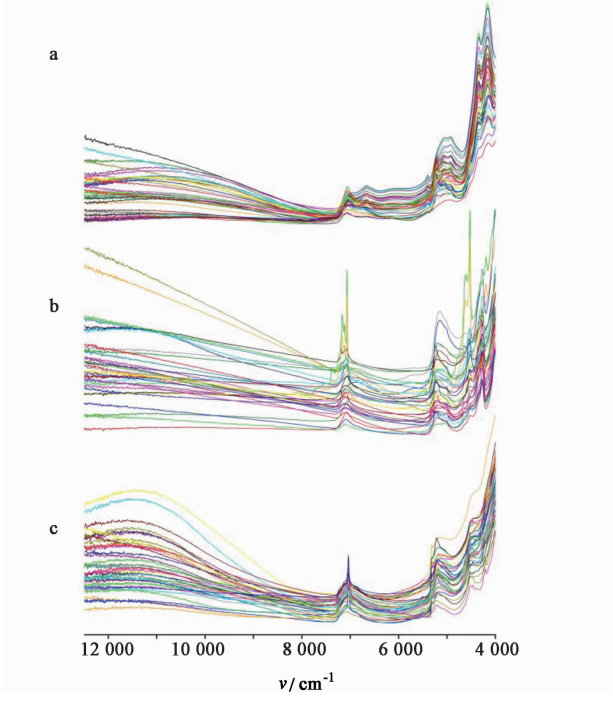
2.1 近红外光谱采集 将所有炉甘石样品粉碎,过 60 目筛,制成均匀粉末。分别取各批样品 2 g 置于样品杯中,在傅里叶变换近红外光谱仪上,采用积分球漫反射测试模式进行光谱扫描。在 OPUS 7.5 软件上设置光谱扫描范围12 500 ~ 4 000 cm^{-1},扫描数 32 次,分辨率 8 cm^{-1}。每个样品重复扫描 3 次,导入 OPUS 7.5 软件并取平均图谱用于分析。各类样品图谱见图 1。

由图 1 可知,炉甘石生品、伪品及炮制品的 NIRS 主要特征区域为 7 500 ~ 4 000 cm^{-1}。该谱段内,同类样品的 NIRS 较为一致,而异类样品 NIRS 差异较大。说明 7 500 ~ 4 000 cm^{-1}谱段具有用于

表 1　炉甘石的样品信息

Table 1　Information of Calamina samples　批

来源	训练集			测试集			合计
	生品	伪品	炮制品	生品	伪品	炮制品	
产于湖南	2	0	–	1	0	–	3
产于广西	3	1	–	3	0	–	7
产于贵州	2	0	–	1	0	–	3
产于四川	1	0	–	1	0	–	2
产于云南	2	0	–	1	0	–	3
购于亳州	2	4	–	2	1	–	9
购于安国	2	0	–	1	1	–	4
购于深圳	2	0	–	4	1	–	7
购于禹州	0	3	–	0	2	–	5
购于玉林	0	1	–	0	0	–	1
购于樟树	0	3	–	0	2	–	5
购于西安	0	3	–	0	0	–	3
马应龙药业集团股份有限公司	3	0	–	3	0	–	6
自制	3	5	20	3	8	18	57



a. 生品；b. 伪品；c. 炮制品

图 1　炉甘石样品的 NIRS

Fig. 1　NIRS of Calamina samples

鉴别炉甘石生品、伪品和炮制品的潜力。另根据前期研究，炉甘石生品主要成分中的 CO_3^{2-} 在 4 360 cm^{-1}和 4 160 cm^{-1}附近具有 C—O 键特征振动峰[10]。而由于阳离子的影响，伪品（主要含 $CaCO_3$）的 C—O 键振动峰相对偏移，表现在 4 275 cm^{-1}附近；同时，样品中的吸附水、结晶水及结构水含量和结合方式不同，其在 NIRS 上的特征亦不同，主要表现在 7 000 cm^{-1}附近和 5 500 ~ 5 000 cm^{-1}。这从理论上说明 NIRS 可用于炉甘石真伪鉴别。对比图 1（a）和（c），炮制品经高温煅烧，$ZnCO_3$ 分解为 ZnO，C—O 键破坏，生品在 4 360 cm^{-1}和 4 160 cm^{-1}附近的特征减弱或消失，据此可用于炮制品的鉴别。同时，煅烧中部分杂质发生变化，7 000 cm^{-1}附近的尖锐的结构水的特征峰增强，XRD 初步研究表明该特征可能与样品中常见的杂质异极矿有关，具体原因及其相关性尚待进一步探索。上述研究已证明了 NIRS 技术可用于炉甘石的鉴别。但受样品来源、伴生矿等影响，同类药材之间图谱亦存在差异，且 NIRS 峰多展宽，重叠严重，不便于以直观对比的方法鉴别炉甘石。故选择 NIRS 技术结合化学计量学算法，建立模式识别模型，用于炉甘石鉴别。

2.2　NIRS 预处理及降维

2.2.1　主成分分析（PCA）降维　根据上述光谱特征分析，确定建模谱段为 7 500 ~ 4 000 cm^{-1}，在 OPUS 7.5 软件中导出此谱段内的各样品原始光谱数据，每个样品含有 909 个数据点。这些数据组成了具有较高维度的样品集光谱矩阵。该光谱矩阵信息重叠严重，具有严重的多重共线性，若直接用于建模，则易造成模型过拟合，模型稳定性及分析精度也随之降低。故需对光谱进行降维处理[7,11]。在 Unscrambler 9.7 软件中，对训练集原始光谱数据（7 500 ~ 4 000 cm^{-1}）进行 PCA 降维，主成分累计贡献率变化见图 2。前 2 个主成分（PC1 和 PC2）累计贡献率 >95%，可代表原数据的主要信息，故以训练集各样品 PC1 和 PC2 的得分分别为横坐标和纵坐标，在平面直角坐标系中绘制样品主成分得分散点图，见图 3，据此对炉甘石进行 PCA 判别分析。结果发现训练集中的生品样品点与伪品、炮制品样品点具有分离趋势，分布区域重叠较少；但伪品和炮制品的样品点所处区域彼此重叠，二者不能通过图 3 进行区分。提示原始光谱直接进行 PCA 降维，用于炉甘石鉴别的效果不佳，尚需对原始光谱数据进行预处理。

2.2.2　光谱预处理方法筛选　NIRS 建模时需对原始光谱进行预处理，以消除无关信息和噪声的干扰[12]。常用的光谱预处理方法有平滑、归一化、求导等。合适的光谱预处理可增强模型性能。在

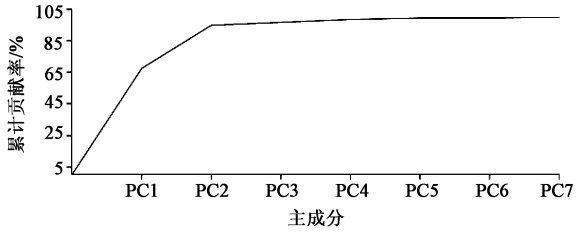


图 2　原始 NIRS 数据(7 500 ~ 4 000 cm^{-1})PCA 降维所得各主成分的累计贡献率变化

Fig. 2　Change curve of cumulative contribution rate of each principal component obtained from PCA dimension reduction for original NIRS data(7 500-4 000 cm^{-1})

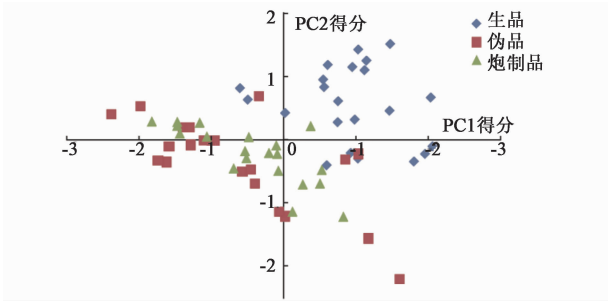
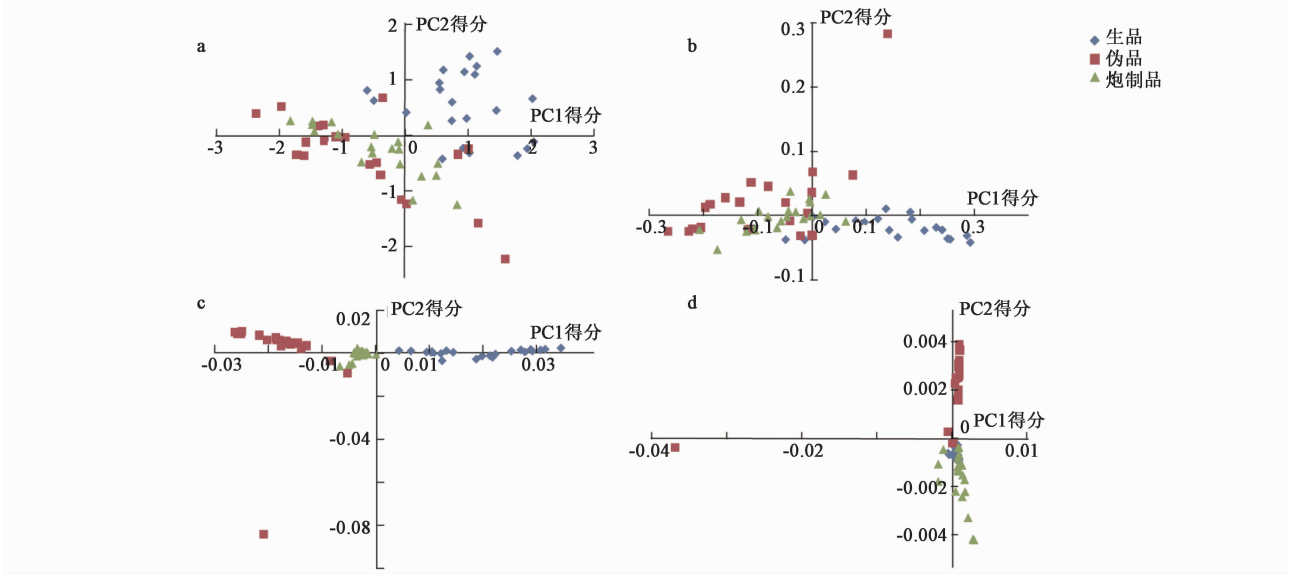


图 3　训练集各样品原始 NIRS 的 PC1 和 PC2 得分散点

Fig. 3　Scatter diagram of PC1 and PC2 scores of original NIRS from training set samples

Unscrambler 9.7 软件中,分别以 Savitzky-Golay 平滑法,矢量归一化法(VN),一阶导数法(FD),二阶导数法(SD)对原始光谱数据(7 500 ~ 4 000 cm^{-1})进行预处理,并分别进行 PCA 降维。训练集样品不同预处理后光谱的前 2 个主成分得分散点图见图 4。



a. Savitzky-Golay 平滑法;b. VN;c. FD;d. SD

图 4　炉甘石训练集样品在 7 500 ~ 4 000 cm^{-1} 谱段下 PC1 和 PC2 的得分散点

Fig. 4　Scatter diagrams of PC1 and PC2 scores of Calamina training set samples at spectrum segment of 7 500-4 000 cm^{-1}

由图 4 可知,训练集样品的光谱经 FD 预处理后,在 PC1 和 PC2(累计贡献率 78.23%)得分散点图上,同类样品彼此靠近,异类样品彼此分离。相比于其他预处理方法,FD 预处理后样品分类效果最佳,故选择 FD。但由于伪品之间差异较大,其样品点在得分散点图上的分布较离散。该结果与前期研究[7]以全谱段(12 500 ~ 4 000 cm^{-1})进行 PCA 判别分析结果一致。由张晓冬等[7]的前期研究可知,为进行准确鉴别,需确定各类样品点在主成分得分散点图上的阈值区域。因此,借鉴支持向量机(SVM)超平面的概念[13],确定线性分界线,进行炉甘石鉴别。但实际问题多为非线性问题,选择基于非线性超平面将具有更优的分类效果。基于此,为提高炉甘石 NIRS 鉴别模型的准确率,将样品光谱矩阵经 PCA 降维所提取的主成分得分矩阵设为 SVM 输入变量,优化 SVM 内部参数,建立 PCA-SVM 分类模型。

2.3　PCA-SVM 模型的建立与评价　SVM 算法[13]可在高维空间中利用核函数构造线性判别函数来解决非线性问题,在解决小样本、非线性、高维数据时具有很大优势,在很大程度上能够克服“维数灾难”和“过学习”等问题。SVM 分类模型常用于解决分类判别问题,是常用的定性模型。

2.3.1　SVM 核函数及内部参数优化方法　SVM

性能优劣主要取决于核函数及其参数的选择，一般而言，满足 Mercer 条件的函数都可以作为核函数，常用的有多项式核函数，Sigmoid 核函数和高斯径向基核函数（RBF）等[14]。其中应用最广泛的是 RBF，其在低维、高维、小样本或大样本等情况均适用。在 RBF 的 SVM 算法中，有 2 个重要参数——惩罚因子 c 和核函数参数 g，二者大小均需进行优化。常用的优化方法包括网格搜索算法、粒子群优化（particle swarm optimization，PSO）算法、遗传算法（genetic algorithm，GA）等[15]。其中网格搜索算法比较直观，可以穷尽搜索直到得到 SVM 最优化的参数，同时可以并行计算，节约参数优化时间[16]。PSO 算法模拟鸟群飞行觅食的行为，通过鸟之间的集体协作使群体达到最优目的。在 PSO 算法系统中，每个备选解被称为一个粒子，多个粒子共存、合作寻优，每个粒子根据其自身的经验和相邻粒子群的最佳经验在问题空间中向更好的位置飞行，搜索最优解[17]。GA 是借鉴生物界自然选择和遗传机制，利用选择、交换和突变等算法的操作，随着不断的遗传迭代，保留目标数值较优的变量，最终达到最优结果的一种方法[18]。GA 和 PSO 算法均为智能寻优算法。

2.3.2 模型验证与评价指标　将所有样品分为训练集和测试集，其中训练集样品用于 SVM 模型的建模学习，并通过训练集内部五折交叉验证，指导 SVM 参数组合（c，g）的寻优过程。一般取五折交叉验证准确率最大时所对应的参数组合即为最优。同时，考察模型对训练集的预测准确率，用于模型性能的对比。测试集样品用于模型预测能力的评价，即采用建模学习获得的最佳 SVM 模型对测试集样品进行预测，并计算预测准确率。模型对测试集预测准确率越高，说明其预测能力越强。

直接将训练集样品的光谱矩阵（62 × 909 个数据点）作为 SVM 输入变量，由于数据点过多，使得模型异常庞大，计算速度慢，多重共线性严重，模型易过拟合。因此需对光谱数据进行 PCA 降维处理。即将原始光谱矩阵进行预处理，然后通过 PCA 提取主成分，获得主成分得分矩阵，最后将主成分得分矩阵作输入变量，进行 SVM 建模。选择 FD 为光谱预处理方法，以 PCA 提取的 PC1 和 PC2 得分矩阵（62 × 2个数据点）为 SVM 输入变量，建立 PCA-SVM 模型。同时，以各样品经 2015 年版《中国药典》所载方法和 XRD 共同确定的样品类别（生品、伪品或炮制品）为模型的参考目标输出。为适应模型学习需要，分别以自然数 1，2，3 代表样品的 3 个类别，依次为正品、伪品和炮制品。确定模型输入和参考输出数据后，导入 Matlab R2014a 软件中，并对输入数据进行标准化处理。在该软件中采用 LibSVM 代码工具包建立初始化的 PCA-SVM 模型，并分别采用网格搜索算法，PSO 算法和 GA 对 SVM 内部参数进行寻优，其中网格搜索算法的寻优范围为 $\log_2 c \in [-10, 10]$，$\log_2 g \in [-10, 10]$，寻优步数均为 1，寻优过程见图 5；PSO 算法寻优范围 $c \in [0.001, 1\,000]$，$g \in [0.001, 1\,000]$，寻优步数均为 0.6，寻优过程见图 6；GA 寻优范围 $c \in [0.001, 1\,000]$，$g \in [0.001, 1\,000]$，寻优步数均为 0.9，寻优过程见图 7。

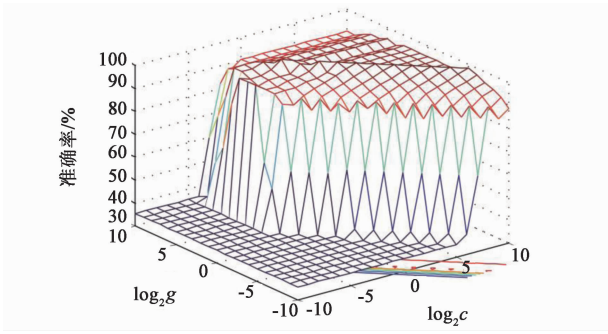


图 5　基于网格搜索算法的 SVM 内部参数寻优过程

Fig. 5　Optimization process for internal parameters of SVM based on grid search algorithm

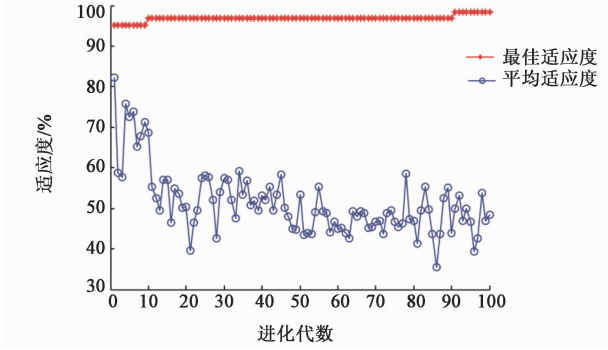


图 6　基于 PSO 算法的 SVM 内部参数寻优过程

Fig. 6　Optimization process for internal parameters of SVM based on PSO algorithm

依次利用 3 种寻优法所获得的参数组建立 3 个模型（模型 1 ~ 3）。3 个模型的内部参数及验证、评价效果见表 2。结果发现 3 种寻优法确定的内部参数虽不同，但效果一致，预测能力相同，且模型对训练集和测试集预测准确率均较高（ > 96%，3 个模型五折交叉验证、训练集和测试集的准确率均依次为 98.39%，96.77% 和 96.23%）。说明 3 个 PCA-SVM 模型均可用于炉甘石鉴别。

在寻优过程中发现，网格搜索算法寻优原理简单，具有可重复性，而 PSO 算法和 GA 寻优均为新兴

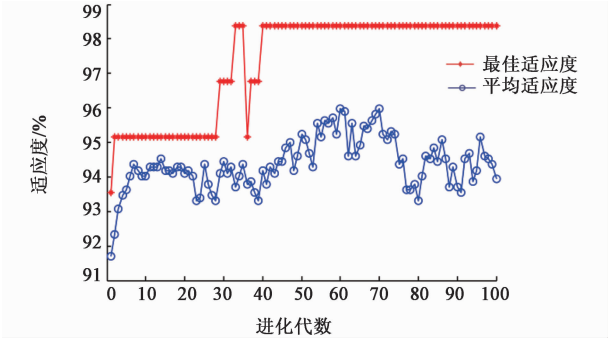


图 7　基于 GA 的 SVM 内部参数寻优过程

Fig. 7　Optimization process for internal parameters of SVM based on GA

表 2　主成分 PC1 和 PC2 所建 PCA-SVM 模型的建模参数

Table 2　Modeling parameters for PCA-SVM model from PC1 and PC2

模型	参数寻优方法	c	g
1	网格搜索算法	1.000 0	32.000 0
2	PSO 算法	2.526 6	19.641 4
3	GA	11.546 1	2.492 9

的智能算法，其运算过程具有一定的随机性，对复杂问题的解决能力更强。建模光谱数据经 FD 预处理后进行 PCA 降维，使得建模数据相对简化，复杂度降低，且不同类样品具有较强的分离趋势。故确定最佳内部寻优方法为网格搜索算法。以网格搜索算法确定的模型 1 对训练集和测试集样品的预测效果见图 8。

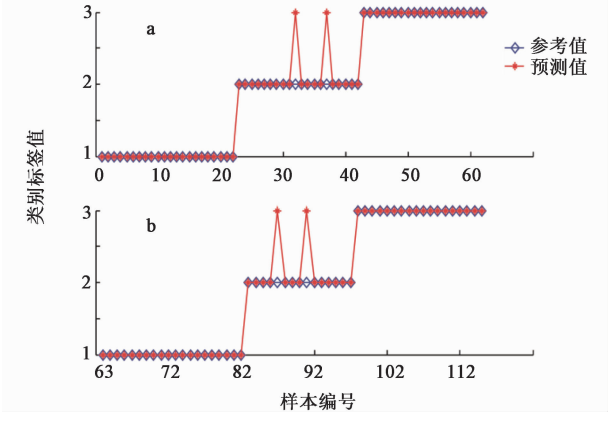


图 8　模型 1 对训练集(a)和测试集(b)样品的预测效果

Fig. 8　Predicting effect of model 1 on samples of training set(a) and test set(b)

2.4　PCA-SVM 建模最佳主成分数优选　由于 FD 预处理 NIRS 数据(7 500 ~ 4 000 cm^{-1})经 PCA 降维所得的前 2 个组成分累计贡献率(78.23%)较低，其对原始光谱数据信息的代表性不强，导致所建 PCA-SVM 模型(模型 1)对部分样品鉴别有误。故需在模型 1 基础上，增加建模主成分个数，避免数据信息丢失。但增加主成分，则使数据维度增加，模型稳定性可能降低，造成模型过拟合。因此需对建模主成分数进行优选。将 FD 预处理光谱经 PCA 提取的前 3 个，前 4 个，前 5 个和前 6 个主成分的得分矩阵分别作为 SVM 输入变量，依据上述 SVM 建模及参数寻优方法(网格搜索算法)依次建立 4 个 PCA-SVM 分类模型(模型 4 ~ 7)，并依次进行验证和评价，见表 3。

表 3　不同主成分数的 PCA-SVM 模型的建模参数及验证、评价效果

Table 3　Modeling parameters and effect of verification and evaluation for PCA-SVM model from different principal component numbers

模型	主成分数/个	累计贡献率/%	c	g	准确率/%		
					五折交叉验证	训练集	测试集
4	3	90.85	0.25	16	98.39	98.39	98.39
5	4	94.44	4.00	4	98.39	96.77	96.23
6	5	96.16	0.25	8	100.00	100.00	100.00
7	6	97.43	1.00	2	100.00	100.00	100.00

对比分析表 2 和表 3 可知，当以前 5 个主成分的得分矩阵为输入变量，利用 NIRS 建立的 3 种炉甘石样品 PCA-SVM 分类模型(模型 6)达到最佳效果。此时，模型 6 的五折交叉验证准确率及其对训练集和测试集的预测准确率均为 100%；增加主成分数，模型效果无显著改善；而减少主成分数，模型预测准确率显著降低。故确定炉甘石 PCA-SVM 分类模型的最佳输入主成分数为 5(PC1，PC2，PC3，PC4 和 PC5)，其累积贡献率达 96.16%，可代表原始光谱的绝大多数信息。模型 6 对训练集和测试集的预测效果见图 9。

3　讨论

本研究通过收集大量炉甘石样品，并依据 2015 年版《中国药典》所载方法及 XRD 分析，对市场上销售的炉甘石样品进行了较为全面的质量考察。在收集的 58 批样品中，有 22 批为不合格的伪品，不合格率达 37.93%，说明目前市场上炉甘石药材质量问题严重，探索快速有效的炉甘石真伪鉴别和质量评价方法是非常必要的。基于此，本文在大量补充代表性样品的基础上，利用 PCA 结合 SVM 算法，建立炉甘石生品、伪品和炮制品的 NIRS 定性模型，

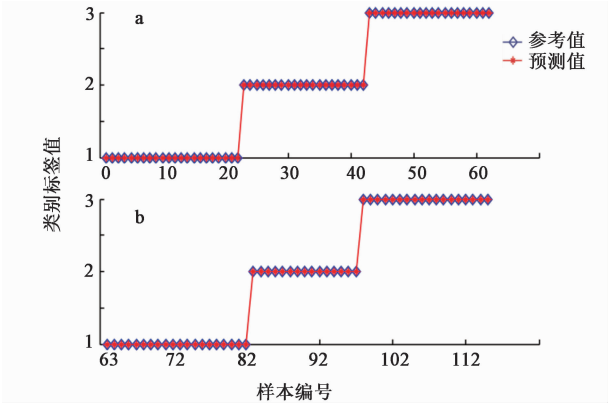


图 9　模型 6 对训练集(a)和测试集(b)样品的预测效果

Fig. 9　Predicting effect of model 6 on samples of training set(a) and test set(b)

用于炉甘石的快速准确鉴别。该 PCA-SVM 模型以 FD 对建模谱段(7 500 ~ 4 000 cm^{-1})的光谱数据进行预处理,然后以 PCA 提取预处理后光谱数据的前 5 个主成分(累计贡献率 96.16%),并以这 5 个主成分得分矩阵为 SVM 输入变量,以样品类别标签参考值为输出,SVM 算法核函数选择 RBF,SVM 内部参数优化选择网格搜索算法。最终确定 $c = 0.25$, $g = 8$,模型训练集五折交叉验证准确率及其对训练集和测试集预测准确率均达到 100%,模型建模性能优异,预测能力强,可用于炉甘石生品、伪品及炮制品粉末的快速、无损、准确鉴别。

目前,本课题组已尝试多种化学计量学算法用于炉甘石 NIRS 鉴别分析,分别建立了 BP-ANN 模型,PCA 判别模型,聚类分析模型以及本文所建立的 PCA-SVM 模型。这些模型各具优势,同时亦存在不足。前期研究已对 BP-ANN 算法,PCA 判别法和聚类分析法进行了详细对比分析[7],常规的 PCA 判别法和聚类分析法为传统线性算法,建模方法成熟,所建模型性能稳定,可推广性强;BP-ANN 算法的非线性拟合能力强,但优化过程存在较大的随机性,建模难度大,算法不成熟,尚需进一步的应用探索。本文采用 PCA-SVM 算法,该方法充分利用 PCA 在数据降维上的优势,并结合了以 BRF 为核函数的 SVM 算法所具有的非线性拟合能力,相比于 PCA 判别法,其分析准确率更高,更适合分析较复杂的实际问题。此外,基于散点图的 PCA 判别法,在分析二维和三维数据时简单方便,但此时可能存在信息缺失,因此模型的预测能力难以进一步提升;而 PCA-SVM 模型通过增加维度,并进行主成分数优选,可以获得更高的预测准确率。

此外,SVM 算法与 BP-ANN 算法均是新兴的人工智能算法[9,19]。在炉甘石 NIRS 鉴别分析中,BP-ANN 算法优化存在较大的随机性,建模过程稳定性差;而 SVM 模型优化过程更稳定,更适合小样本量分析。前期 SUN 等[6]所建立的炉甘石 BP-ANN 模型预测准确率达 95%,但其预测性能尚可通过增加代表性样品或进行特征谱段优选进行提升。

本研究虽实现了炉甘石的快速鉴别,但这并不能完全满足炉甘石质量控制的需求,尚需寻找适于炉甘石定量分析的快检方法。而 NIRS 技术在定量分析上亦具有优势[20],其用于炉甘石定量分析的潜力较大。基于此,本课题组已尝试利用 NIRS 技术对炉甘石进行定量分析,并进行了初步探索[21],以期构建完善的炉甘石 NIRS 快速质量控制体系。

[参考文献]

[1] 国家中医药管理局《中华本草》编委会. 中华本草[M]. 上海:上海科技出版社,1999:382-383.
[2] 国家药典委员会. 中华人民共和国药典. 一部[M]. 北京:中国医药科技出版社,2015:227.
[3] 周灵君,张丽,路长珍,等. 市售生、煅炉甘石的成分分析及质量评价[J]. 中国药房,2010,21(27):2534-2536.
[4] 张杰红,刘友平,施学娇,等. 市售炉甘石的化学成分及抑菌活性研究[J]. 中药与临床,2011,2(6):16-18.
[5] 周灵君,徐春蕾,张丽,等. 炉甘石炮制机制研究[J]. 中国中药杂志,2010,35(12):1556-1559.
[6] SUN Y B, CHEN L, HUANG B S, et al. A rapid identification method for calamine using near-infrared spectroscopy based on multi-reference correlation coefficient method and back propagation artificial neural network[J]. Appl Spectrosc,2017,71(4):1447-1456.
[7] 张晓冬,陈龙,白玉,等. 近红外光谱结合主成分分析和聚类分析鉴别炉甘石生品、伪品和炮制品[J]. 中国实验方剂学杂志,2018,24(12):1-8.
[8] 明晶,陈龙,陈科力,等. 基于近红外光谱和 SVM 算法对琥珀掺伪的定性鉴别与定量分析[J]. 中药材,2017,40(1):32-37.
[9] 魏从师,雷福汉,艾伟霞,等. 基于 NIRS 技术和 PCA-SVM 算法 6 种树脂及其他类中药的快速鉴别[J]. 中国实验方剂学杂志,2017,23(9):17-23.
[10] 陈龙,袁明洋,陈科力. 常见矿物药近红外漫反射光谱特征归纳与分析[J]. 中国中药杂志,2016,41(19):3528-3536.
[11] 陆婉珍,袁洪福,褚小立. 近红外光谱仪器[M]. 北京:化学工业出版社,2010:35-37.
[12] 尼珍,胡昌勤,冯芳. 近红外光谱分析中光谱预处理

方法的作用及其发展[J]. 药物分析杂志, 2008, 28(5): 824-827.

[13] Vapnik V N. *The Nature of Statistical Learning Theory* [M]. New York: Springer, 1999: 988-999.

[14] 李盼池, 许少华. 支持向量机在模式识别中的核函数特性分析[J]. 计算机工程与设计, 2005, 26(2): 302-304.

[15] 林升梁, 刘志. 基于 RBF 核函数的支持向量机参数选择[J]. 浙江工业大学学报, 2007, 35(2): 163-167.

[16] 王健峰, 张磊, 陈国兴, 等. 基于改进的网格搜索法的 SVM 参数优化[J]. 应用科技, 2012, 39(3): 28-31.

[17] 姚全珠, 蔡婕. 基于 PSO 的 LS-SVM 特征选择与参数优化算法[J]. 计算机工程与应用, 2010, 46(1): 134-136.

[18] 杨旭, 纪玉波, 田雪. 基于遗传算法的 SVM 参数选取[J]. 辽宁石油化工大学学报, 2004, 24(1): 54-58.

[19] 徐子杰, 陈龙, 刘义梅, 等. 基于多参考相关系数法和 BP-ANN 建立紫石英的近红外光谱定性模型[J]. 中国实验方剂学杂志, 2017, 23(22): 37-42.

[20] 史晶晶, 张迪文, 白雁, 等. 近红外光谱法快速测定酒女贞子中女贞苷含量[J]. 中国实验方剂学杂志, 2016, 22(8): 69-73.

[21] ZHANG X D, CHEN L, SUN Y B, et al. Determination of zinc oxide content of mineral medicine calamine using near-infrared spectroscopy based on MIV and BP-ANN algorithm [J]. Spectrochim Acta A Mol Biomol Spectrosc, 2017, 193: 133-140.

[责任编辑　刘德文]