·资源与鉴定·

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

魏从师，雷福汉，艾伟霞，冯晶，郑虹，马丹，石新华*
（武汉市中医医院 药学部，武汉 430014）

［摘要］ **目的**：利用近红外漫反射光谱（NIRS）法，结合主成分分析（PCA）和支持向量机（SVM）联用算法，建立 6 种树脂及其他类中药安息香（Benzoinum），琥珀（Succinum），没药（Myrrha），乳香（Olibanum），松香（Colophonium），天竺黄（Bambusaen Concretio Silicea）的 NIR 模式识别模型，用于该 6 味中药的快速鉴别。**方法**：收集上述 6 种中药样品，经性状鉴别和理化分析确定正品药材 55 批，粉碎成均匀粉末，在 4 000 ~ 12 000 cm^{-1} 光谱区，采集各样品粉末的 NIR 光谱，选取特征谱段 9 000 ~ 5 400，5 000 ~ 4 000 cm^{-1} 为建模谱段，分别采用矢量归一化法（vector normalization，VN），一阶导数法（first derivative，FD），二阶导数法（second derivative，SD）3 种不同光谱预处理方法进行预处理并分别进行 PCA 降维。根据主成分空间散点图，优选最佳预处理方法。利用最佳预处理方法处理后光谱的 PCA 降维数据，建立 SVM 模式识别模型，SVM 模型参数 c 和 g 采用网格搜索法结合五折交叉验证进行寻优。对比不同主成分数所建 PCA-SVM 模型的预测准确率，确定最佳的主成分数，最终建立 6 种中药 NIR 快速鉴别模型。**结果**：在 9 000 ~ 5 400，5 000 ~ 4 000 cm^{-1} 建模谱段，确定最佳光谱预处理方法为 SD，SD 预处理光谱 PCA 降维后，确定最佳主成分数为 3 个，累计贡献率达 93.57%。经网格搜索法确定最佳 SVM 建模参数组为 c = 65 536，g = 512。所建 PCA-SVM 模型对训练集和验证集样品预测正确率均达 100%，模型五折交叉验证准确率亦达 100%。**结论**：所建的 6 种中药 NIR 光谱 PCA-SVM 鉴别模型，预测准确率高，模型预测能力强，结合 NIRS 技术无损、快速的优点，该模型可用于上述 6 种中药的无损、快速鉴别。

［关键词］ 近红外漫反射光谱；主成分分析；支持向量机；树脂及其他类中药；模式识别；空间散点图；网格搜索；快速鉴别
［中图分类号］ R282.2 ［文献标识码］ A ［文章编号］ 1005-9903(2017)09-0017-07
［doi］ 10.13422/j.cnki.syfjx.2017090017
［网络出版地址］ http://www.cnki.net/kcms/detail/11.3495.R.20170214.1609.050.html
［网络出版时间］ 2017-02-14 16:09

Rapid Identification of 6 Kinds of Traditional Chinese Medicines Containing Resins and Other Components Based on Near Infrared Refectance Spectroscopy and PCA-SVM Algorithm

WEI Cong-shi, LEI Fu-han, AI Wei-xia, FENG Jing, ZHENG Hong, MA Dan, SHI Xin-hua*
(*Department of Pharmacy*, *Wuhan Hospital of Traditional Chinese Medicine*, *Wuhan* 430014, *China*)

［**Abstract**］ **Objective**: To establish the pattern discernment model with the use of near-infrared reflectance spectroscopy (NIRS), principal component analysis (PCA) and support vector machine (SVM) algorithms for rapid identification of Benzoinum, Succinum, Myrrha, Olibanum, Colophonium and Bambusae Concretio Silicea, all of which are the traditional Chinese medicines (TCMs) containing resins and other

［收稿日期］ 20161205(037)
［基金项目］ 武汉市卫计委科研项目(WZ16Z03)
［第一作者］ 魏从师，副主任中药师，从事中药鉴定研究，Tel：027-83087217，E-mail：2673508232@qq.com
［通讯作者］ *石新华，主任中药师，从事中药制剂开发研究，Tel：027-83087205，E-mail：xhshi@qq.com

components. **Method**: According to morphological identification and conventional physical and chemical analysis on the above 6 kinds of samples collected form major national medicinal materials markets, a total of 55 batches of samples were verified as genuine medicinal herbs. These samples were smashed into uniform powders, and the NIR spectra of sample powder were scanned at 4 000-12 000 cm^{-1}. The characteristic bands at 9 000-5 400, 5 000-4 000 cm^{-1} were used to establish models for pre-treatment by vector normalization (VN), first derivative (FD) and second derivative (SD) respectively, and in addition, PCA dimension reduction was also conducted. The best pre-treatment method was selected according to the three-dimensional scatter diagram of principal components, and the PCA dimension reduction data after best pre-treatment method were used to establish the pattern discernment models based on the SVM algorithm. The SVM model parameters c and g were optimized by using grid search method combined with 5-fold cross validation method. By comparing the forecast accuracy of PCA-SVM models established with different number of principal components, the best number was optimized and finally, NIR rapid identification model was established for 6 Chinese herbs. **Result**: At 9 000-5 400, 5 000-4 000 cm^{-1}, SD was determined as the best pre-treatment method, and after SD pre-treatment for PCA dimension reduction, the best number of principal components was determined as 3, with an accumulative contribution rate of 93.57%. According to the grid search method, $c = 65\ 536$, $g = 512$, which were the optimum parameters of the SVM model. In the established PCA-SVM models, the forecast accuracy was 100% for both training set and verification set, and the forecast accuracy was also 100% for 5-fold cross validation method. **Conclusion**: The NIRS identification model based on PCA-SVM had high forecast accuracy and strong prediction ability, which can be used for rapid, nondestructive and accurate identification of above 6 kinds of TCMs when it was used in combination with the advantages of NIRS.

[**Key words**] near-infrared reflectance spectroscopy (NIRS); principal component analysis (PCA); support vector machine (SVM); traditional Chinese medicines containing resins and other components; pattern discernment; three-dimensional scatter diagram; grid search method; rapid identification

树脂类中药,如安息香、没药、乳香、松香是一类较常用的药物,大多来源于植物体的渗出物、分泌物,或通过简单的切割所得[1];此外,其他类中药中,琥珀为树脂的化石,天竺黄为植物分泌液干燥后的块状物,二者与树脂类中药类似,来源均与植物的分泌物相关。上述药材具有良好的防腐、抗菌、消炎、活血、祛痰、消肿等功效,临床上用于芳香开窍、调气活血、舒筋止痛、消积杀虫等,对许多常见病疗效显著[2]。如:琥珀具有镇惊安神、活血散瘀、利水通淋之功效[3],为临床常用的贵重药材之一;安息香为中医常用为开窍药,具有开窍辟秽、行气活血、镇咳祛痰之功[4]。6 种树脂及其他类中药来源及成分较复杂[5],多数为固体或半固体,少数为液体,性状特征相近且无典型的显微特征,在中药鉴定中,对他们的研究较少,除常规性状及理化鉴别[1]外,尚无较好的鉴别方法[6]。

近红外漫反射光谱(NIRS)技术是中药鉴定领域发展较为迅速的光谱分析技术之一,其具有无损、快速、绿色环保等优点[7]。NIRS 主要反映 C-H, O-H 和 N-H 等含氢基团的倍频与合频吸收,而上述 6 种树脂及其他类中药富有含氢基团,在近红外光谱区具明显特征吸收。此外 NIRS 技术随着化学计量学的发展而兴起[8],二者联用,可以实现中药真伪的快速、准确鉴别。故本文采用 NIRS 技术结合主成分分析(PCA)和支持向量机(SVM)算法等多种化学计量学方法建立 6 种中药安息香、琥珀、没药、乳香、松香、天竺黄的 NIRS 模式识别模型,用于这 6 种药材的快速鉴别。该研究中,采用 PCA 算法对高维光谱数据进行降维,实现 SVM 建模过程的输入数据压缩,简化模型,提高模型的预测能力和稳健性。此外,本研究通过绘制光谱 PCA 降维所得的前 3 个主成分得分三维相关散点图,对不同预处理方法进行优选,该方法直观、简便且有效。

1 材料

MPA 型傅里叶变换近红外光谱仪(配备固体积分球漫反射附件),OPUS 7.5 型采集和处理软件(德国布鲁克光学仪器公司);Matlab 2014a 软件(美国 Math Works 公司)。

从亳州、安国、禹州等各大药材市场购入安息香、琥珀、没药、乳香、松香、天竺黄 6 种中药若干批。

由武汉市中医医院药学部张义生主任中药师，经过性状鉴别及理化鉴别获得正品安息香 8 批，正品琥珀8 批，正品没药 9 批，正品乳香 10 批，正品松香 9 批，正品天竺黄 11 批。随机将上述 55 批药材分为训练集和验证集，并以 1 ~6 的整数赋值为各类样品的类别标签。样品鉴别信息及样品集划分见表 1。校正集样本用于建立模式识别模型，验证集样本用于模型预测性能的验证。

表 1　样品信息及样品集划分

Table 1　Sample information and partitioning of sample sets

No.	来源	批号	鉴别结果	药材名	标签	样品集
AXX-01	安徽亳州中药材市场	20140901	安息香(树脂)	Benzoinum	1	验证集
AXX-02	武汉天济中药饮片公司	20150501	安息香(树脂)	Benzoinum	1	训练集
AXX-03	武汉市中医医院	20150301	安息香(树脂)	Benzoinum	1	验证集
AXX-04	安徽亳州中药材市场	20140902	安息香(树脂)	Benzoinum	1	训练集
AXX-05	河南禹州中药材市场	20151001	安息香(树脂)	Benzoinum	1	训练集
AXX-06	江西樟树中药材市场	20130401	安息香(树脂)	Benzoinum	1	验证集
AXX-07	河北安国中药材市场	20130501	安息香(树脂)	Benzoinum	1	训练集
AXX-08	安徽亳州中药材市场	20140903	安息香(树脂)	Benzoinum	1	训练集
HP-01	安徽亳州中药材市场	20140904	琥珀(其他)	Succinum	2	验证集
HP-02	安徽亳州中药材市场	20140905	琥珀(其他)	Succinum	2	训练集
HP-03	河南禹州中药材市场	20151002	琥珀(其他)	Succinum	2	训练集
HP-04	河北安国中药材市场	20130502	琥珀(其他)	Succinum	2	验证集
HP-05	成都荷花池中药材市场	20121101	琥珀(其他)	Succinum	2	训练集
HP-06	武汉天济中药饮片公司	20150601	琥珀(其他)	Succinum	2	训练集
HP-07	北京同仁堂药店(武汉)	20151101	琥珀(其他)	Succinum	2	验证集
HP-08	马应龙药业(武汉)	20160301	琥珀(其他)	Succinum	2	训练集
MY-01	安徽亳州中药材市场	20140906	没药(树脂)	Myrrha	3	训练集
MY-02	武汉市中医医院	20120901	没药(树脂)	Myrrha	3	训练集
MY-03	武汉市中医医院	20140601	没药(树脂)	Myrrha	3	验证集
MY-04	武汉市中医医院	20151101	没药(树脂)	Myrrha	3	训练集
MY-05	广西玉林中药材市场	20130401	没药(树脂)	Myrrha	3	训练集
MY-06	广西玉林中药材市场	20130402	没药(树脂)	Myrrha	3	验证集
MY-07	河南禹州中药材市场	20151003	没药(树脂)	Myrrha	3	训练集
MY-08	武汉天济中药饮片公司	20151101	没药(树脂)	Myrrha	3	训练集
MY-09	河北安国中药材市场	20130503	没药(树脂)	Myrrha	3	验证集
RX-01	安徽亳州中药材市场	20140907	乳香(树脂)	Olibanum	4	训练集
RX-02	安徽亳州中药材市场	20140908	乳香(树脂)	Olibanum	4	验证集
RX-03	武汉市中医医院	20120901	乳香(树脂)	Olibanum	4	训练集
RX-04	武汉市中医医院	20140601	乳香(树脂)	Olibanum	4	验证集
RX-05	广西玉林中药材市场	20130403	乳香(树脂)	Olibanum	4	训练集
RX-06	河南禹州中药材市场	20151004	乳香(树脂)	Olibanum	4	训练集
RX-07	成都荷花池中药材市场	20121102	乳香(树脂)	Olibanum	4	验证集
RX-08	河北安国中药材市场	20130504	乳香(树脂)	Olibanum	4	训练集
RX-09	河北安国中药材市场	20130505	乳香(树脂)	Olibanum	4	验证集
RX-10	武汉天济中药饮片公司	20151101	乳香(树脂)	Olibanum	4	训练集
SX-01	安徽亳州中药材市场	20140909	松香(树脂)	Colophonium	5	训练集
SX-02	安徽亳州中药材市场	20140910	松香(树脂)	Colophonium	5	验证集
SX-03	安徽亳州中药材市场	20140911	松香(树脂)	Colophonium	5	训练集
SX-04	安徽亳州中药材市场	20140912	松香(树脂)	Colophonium	5	训练集
SX-05	河南禹州中药材市场	20151005	松香(树脂)	Colophonium	5	验证集
SX-06	河南禹州中药材市场	20151006	松香(树脂)	Colophonium	5	训练集
SX-07	广西玉林中药材市场	20130404	松香(树脂)	Colophonium	5	训练集
SX-08	江西樟树中药材市场	20130402	松香(树脂)	Colophonium	5	验证集
SX-09	成都荷花池中药材市场	20121103	松香(树脂)	Colophonium	5	训练集
TZH-01	安徽亳州中药材市场	20140913	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-02	安徽亳州中药材市场	20140914	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-03	安徽亳州中药材市场	20140915	天竺黄(其他)	Bambusae Concretio Silicea	6	验证集
TZH-04	广西玉林中药材市场	20130405	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-05	广西玉林中药材市场	20130406	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-06	河北安国中药材市场	20130504	天竺黄(其他)	Bambusae Concretio Silicea	6	验证集
TZH-07	武汉市中医医院	20130901	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-08	北京同仁堂药店(武汉)	20151101	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集
TZH-09	安徽亳州中药材市场	20140916	天竺黄(其他)	Bambusae Concretio Silicea	6	验证集
TZH-10	安徽亳州中药材市场	20140917	天竺黄(其他)	Bambusae Concretio Silicea	6	验证集
TZH-11	河北安国中药材市场	20130505	天竺黄(其他)	Bambusae Concretio Silicea	6	训练集

2 方法与结果

2.1 近红外光谱采集 共计 55 批样品，粉碎并过 60 目筛，分别取 5 g 置于样品瓶中，采用积分球漫反射测试模式扫描 NIR 光谱。光谱扫描范围 4 000 ~ 12 000 cm^{-1}，扫描次数 32 次，仪器分辨率为 8 cm^{-1}。每个样品重复扫描 3 次，取平均值作为该样品的分析光谱。所有样品 NIR 光谱见图 1。

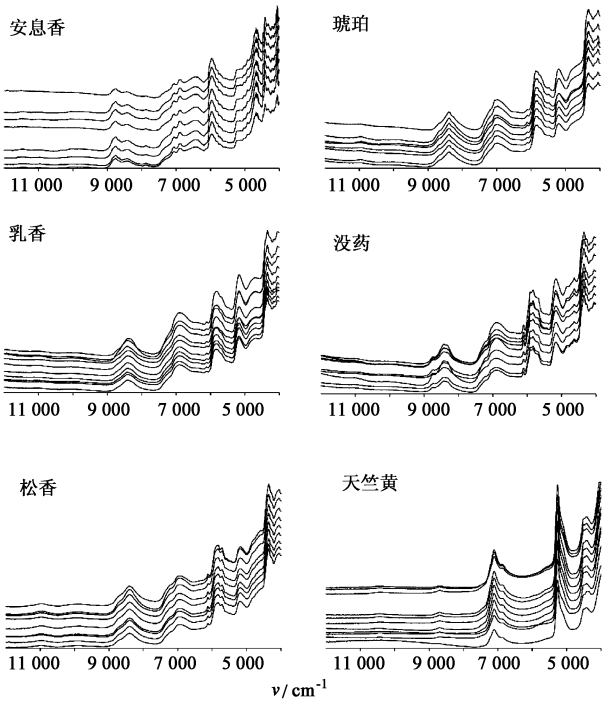


图 1 6 种树脂及其他类中药 NIR 光谱

Fig. 1 NIR spectra of 6 kinds traditional Chinese medicines containing resins and other components

对比分析各类药材 NIR 特征可得 6 种树脂及其他类中药 NIR 主要特征谱段在 9 000 ~ 4 000 cm^{-1}，表现出丰富的 C-H 特征：4 000 ~ 5 000 cm^{-1} 为 C-H 键第一组合峰，谱带较强；5 400 ~ 6 200 cm^{-1} 为 C-H 键一级倍频峰；6 500 ~ 7 500 cm^{-1} 为 C-H 键第二组合峰；8 000 ~ 9 000 cm^{-1} 为 C-H 键二级倍频峰；天竺黄仅表现 2 组组合峰，无倍频峰。但由于各种药材所含 C-H 键数量和所处的化学环境不同，各特征峰在峰形和相对强度上具有一定差异。整体而言，除天竺黄外，其他药材 NIR 特征类似。故通过寻找特征吸收峰，可从该 6 种中药中，直接鉴别出天竺黄。但其他药材难以区分，需要结合化学计量学方法，对原始光谱进行预处理，提取特征信息，建立模式识别模型。同时，由于水分未完全丢失，部分药材在 5 400 ~ 5 000 cm^{-1} 谱段存在水的特征峰[9]。因此，为消除水分对模型的干扰，剔除水的特征谱段，本研究以 9 000 ~ 5 400，5 000 ~ 4 000 cm^{-1} 为建模谱段，进行定性分析。

2.2 光谱预处理及 PCA 降维 NIR 除包含样品自身信息外，还包括了其他无关信息和噪声，如电噪声、样品背景和杂散光等。在建立模型时，可通过多种光谱预处理方法消除这些噪声及干扰。常用的光谱预处理方法有：矢量归一化法（VN），一阶导数法（FD），二阶导数（SD）法等[10]。

此外，定性分析时涉及高维光谱数据，高维数据会明显增加分析的复杂性，且每个样品的光谱数据存在谱峰重叠问题，导致光谱数据信息冗余，因此在建模时，为了简化模型，提高模型的预测能力和稳健性，需要对光谱数据进行降维[11]。PCA 是常用的光谱降维方法。它可将输入变量（光谱矩阵）进行转换，使少数几个新变量（主因子得分矩阵）是原变量的线性组合，同时，这些新变量要尽可能多地表征原变量的数据特征而不丢失信息。经转换得到的新变量互不相关，可消除众多信息共存中相互重叠的信息部分[12]。

为确定最佳的光谱预处理方法并提取有效的主成分，本文利用 Matlab 2014a 软件，在训练集样品的 9 000 ~ 5 400，5 000 ~ 4 000 cm^{-1} 谱段，分别对未处理光谱及 VN，FD 或 SD 预处理后的光谱进行 PCA 降维，并进行对比分析。未预处理及不同预处理方法下，主成分贡献率及累计贡献率见表 2。

由表 2 可得，无预处理及不同预处理条件下，相同主成分数的贡献率不同；不同方法预处理后的光谱 PCA 降维，前 6 个主成分（PC1 ~ PC6）累计贡献率均达到 99%，未预处理光谱烦的前 4 个主成分的累计贡献率达到 99%；4 种情况下，光谱矩阵 PCA 降维，前 3 个主成分（PC1，PC2，PC3）累计贡献率均超过 90%。因此，分别以训练集各样品的 PC1，PC2 和 PC3 的主成分得分值为该样品在三维空间的坐标值，绘制空间散点图。训练集样品未预处理及不同预处理后光谱前 3 个主成分得分散点图见图 2。

由图 2 可知，训练集样品的光谱经 SD 预处理后，其主成分得分的 3 维空间散点图上，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，其分类效果最佳。因此，本文选取 9 000 ~ 5 400，5 000 ~ 4 000 cm^{-1} 谱段，以 SD 为最佳预处理方法进行光谱预处理，以 PCA 提取主成分，获得各样品光谱的主成分得分，作为 SVM 模型的输入变量，建立 6 种树脂及其他类中药近红外光谱 PCA-SVM 定性分析模型。

表 2　不同预处理方法下主成分贡献率及累计贡献率

Table 2　Contribution rate and cumulative contribution rate of principal components in different pre-treatment methods　%

预处理方法	参数	PC1	PC2	PC3	PC4	PC5	PC6
未处理	贡献率	88.64	7.44	1.83	1.61	0.25	0.14
	累计贡献率	88.64	96.08	97.91	99.52	99.77	99.91
VN	贡献率	62.49	23.90	8.15	3.11	0.983	0.52
	累计贡献率	62.49	86.39	94.53	97.64	98.63	99.15
FD	贡献率	52.33	33.16	8.46	2.97	1.48	0.69
	累计贡献率	52.33	85.49	93.95	96.92	98.4	99.09
SD	贡献率	71.18	16.70	5.69	3.14	2.05	0.44
	累计贡献率	71.18	87.88	93.57	96.71	98.77	99.21

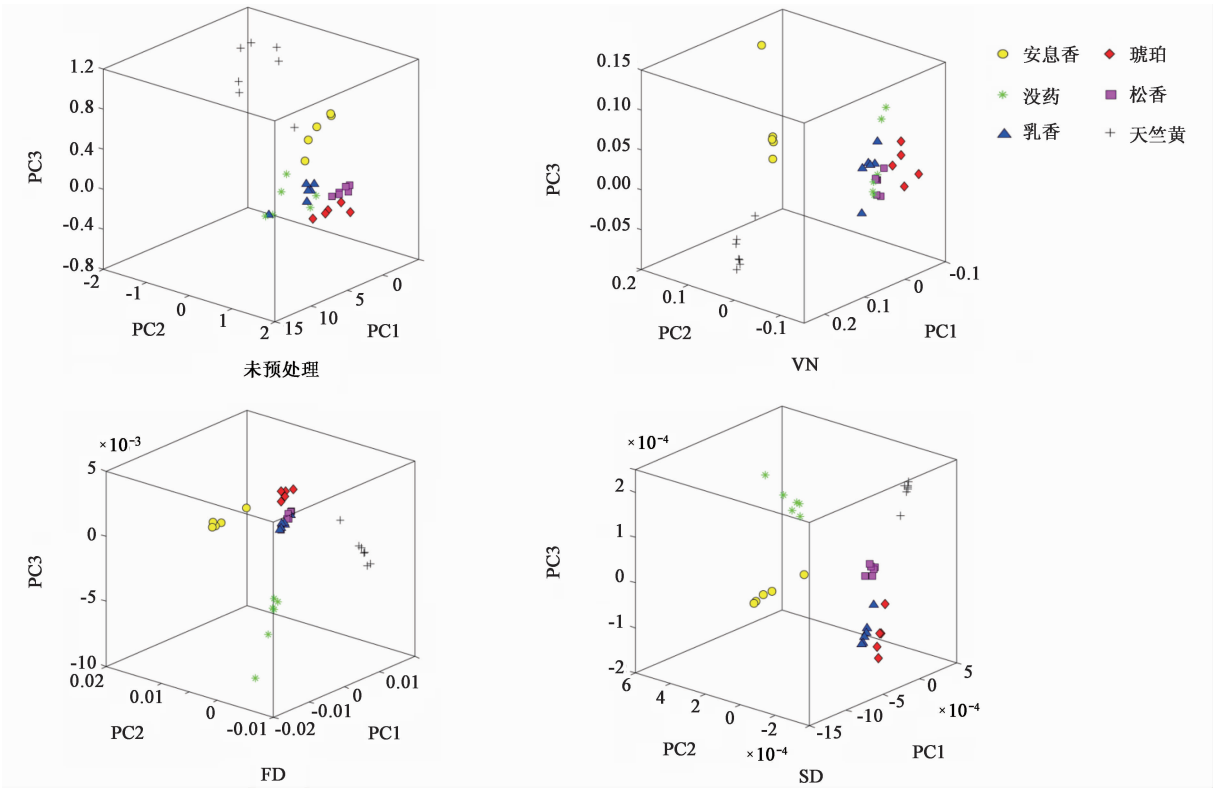


图 2　不同预处理方法下主成分(PC1, PC2, PC3)散点

Fig. 2　Scatter diagram of principal components (PC1, PC2, PC3) with different pre-treatment methods

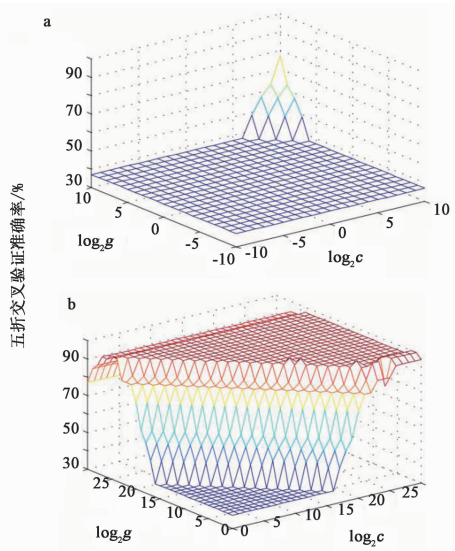
2.3　PCA-SVM 模型的建立及评价

2.3.1　SVM 算法　SVM 算法[13-14]是 1990 年代由 Vapnik 提出的一种基于统计学习理论的新的机器学习方法,属于一种有监督的学习算法。SVM 通过寻求结构化风险最小来提高学习机泛化能力,实现经验风险和置信范围的最小化,达到在统计样本量较少的情况下,亦能获得良好统计规律的目的,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 常用核函数有多项式核,Sigmoid 感知核和高斯径向基核(Radial Basis function, RBF)等,其中 RBF 性能最佳,应用最广泛,其对非线性问题有较好的处理能力[15]。

2.3.2　参数优化方法及模型评价指标　在 RBF 为核函数的 SVM 算法中有 2 个重要的参数:惩罚因子 c 和核函数参数 g,二者大小均需在模型优化过程中确定。网格搜索法是 SVM 问题上应用最为普遍的参数寻优算法,它是将参数(c, g)在一定的空间范围中划分成网格,从网格中全部的点中找到最优参数[16]。基于上述原理,本文使用了 RBF 核函数建立 6 种树脂及其他类中药的 SVM 模式分类模型。模型以 SD 预处理的样品光谱(9 000 ~ 5 400,

5 000 ~ 4 000 cm^{-1}) 经 PCA 提取的主成分得分为 SVM 输入变量,以各类样品的类别标签值为输出,采用网格搜索优化法,以五折交叉验证准确率为指标,对 SVM 模型参数组合(c, g)进行寻,并用最佳参数所建模型对训练集和验证集样品进行预测,分别计算预测准确率。综合考虑五折交叉验证准确率、训练集预测准确率、验证集预测准确率,对 PCA-SVM 模型进行评价。

进行网格搜索寻优时,需首先确定搜索范围。搜索范围不恰当,可能导致寻优失败。扩大搜索范围,虽可增加寻优的成功率,但同时增加了计算量。因此本文采 2 步网格搜索寻优策略。以前 3 个主成分得分作 PCA-SVM 输入变量为例,首先在 $\log_2 c \in [-10, 10]$, $\log_2 g \in [-10, 10]$ 范围进行初步的网格搜索,寻优结果的三维见图 3a。据此可得,初步搜索范围不恰当,训练集样品五折交叉验证准确率在 $\log_2 c > 5$, $\log_2 g > 5$ 位置持续增大,搜索范围内未出现峰值,搜索失败。故进一步在 $\log_2 c \in [0, 30]$, $\log_2 g \in [0, 30]$ 范围进行搜索,寻优结果的三维见图 3b,此时五折交叉验证准确率峰值出现,据此判断最佳参数组合(c, g)分别为:$\log_2 c = 16$, $\log_2 g = 9$,即 $c = 65\ 536$, $g = 512$。



a. 初步搜索;b. 二次搜索

图 3　网格搜索寻优结果的三维

Fig. 3　Three-dimensional figure of optimization results by grid search method

2.3.3　主成分数优选　由表 2,SD 预处理光谱进行 PCA 降维后,前 6 个主成分的累计贡献率达 99%,且由图 2,SD 预处理光谱的 PCA 提取的前 3 个主成分即表现出分类能力,但尚不能确定前 3 个主成分即为最佳主成分数。若在建立模型中使用的主成分数过少,将会出现所建模型不能反映样本被测性质与光谱信息之间的关系,出现欠拟合现象,而使用过多的主成分数,则会影响模型的预测精度及模型的泛化能力,出现过拟合现象。故应对不同主成分数的建模效果进行考察[17]。本文分别将 SD 预处理光谱经 PCA 提取的前 1 个、前 2 个、前 3 个、前 4 个、前 5 个和前 6 个主成分得分作为 SVM 输入变量,依据上述 SVM 建模及参数寻优方法,建立 6 个 PCA-SVM 模型,并依次进行评价,对比分析,确定最佳主成分数,结果见表 3。

表 3　不同主成分数所建 PCA-SVM 模型对比

Table 3　Comparison for PCA-SVM models established with different number of principal components

主成分数	c	g	准确率/%		
			五折交叉验证	训练集	验证集
1	$4.295\ 0 \times 10^9$	131 072	85.71	74.29	80.00
2	97.142 9	2 965 800	97.14	94.29	95.00
3	65 536	512	100.00	100.00	100.00
4	1	8 388 608	100.00	100.00	100.00
5	1	4 194 304	100.00	100.00	100.00
6	1	4 194 304	100.00	100.00	100.00

2.3.4　模型评价　见表 3,当以前 3 个主成分得分为输入变量,6 种树脂及其他类中药 NIR 光谱 PCA-SVM 分类模型即达到最佳效果,训练集和验证集预测准确率均为 100%。增加主成分数,模型效果无显著改善,且 g 值偏大。而减少主成分数,模型预测准确率显著降低,主成分数为 1 时,预测能力最差,光谱信息丢失严重。故确定 6 种树脂及其他类中药 NIR 光谱 PCA-SVM 分类模型的最佳输入主成分数为 3,即以各样品 SD 预处理后光谱的 PC1, PC2 和 PC3 得分值为建模输入数据。所建模型对训练集和验证集的预测效果见图 4。

3　讨论

本文利用不同来源的 55 树脂及其他类中药药材的 NIR 光谱,建立了 6 种树脂及其他类中药 PCA-SVM 模式识别模型,该模型对训练集和验证集样品的预测准确率均达到 100%,预测准确率高,模型具有较强的预测能力,可用于安息香、琥珀、没药、乳香、松香、天竺黄的快速鉴别。试验表明了 SVM 算法在近红外光谱定性鉴别中的较强的适用性,亦验证了 SVM 算法适用于小样本量建模。

建模过程中,本文采用光谱 PCA 降维所得的

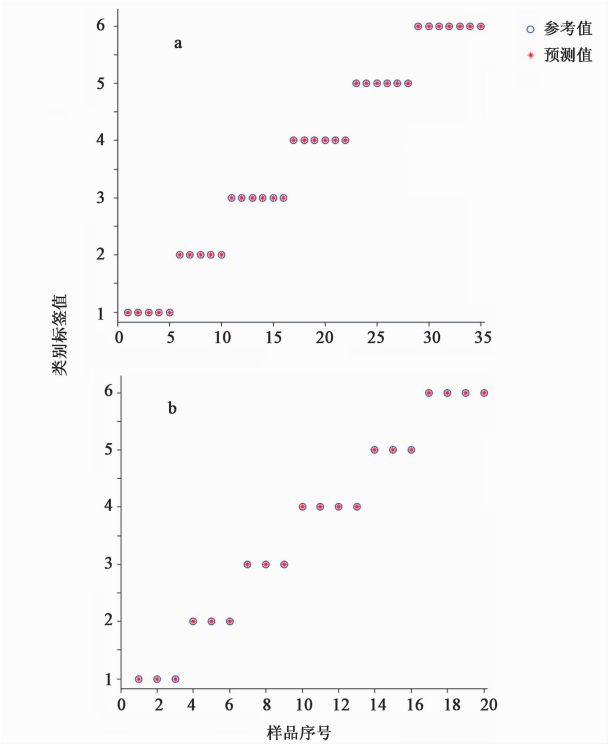


图 4　最佳 PCA-SVM 模型对训练集(a)和验证集(b)样品的预测效果

Fig. 4　Forecasting effect of best PCA-SVM model for training set (a) and validation set (b)

主成分得分空间散点图,对不同光谱预处理方法进行优选。根据空间散点图上各类别样品点的分类情况,确定光谱的最佳预处理方法。上述方法在建立 PCA-SVM 模型时,对预处理方法的优选简单、快速,且可直观考察预处理和降维的效果。

本文首次将 NIRS 技术应用于树脂及其他类中药的鉴别,证明了其具有可行性,为树脂及其他类中药的鉴别提供了新的方法。但由于样本种类和样本量的限制,本文仅对 6 种树脂及其他类中药进行了 NIRS 分析。后期,尚需对模型进行补充,扩充样品种类,增加代表性样品量,使更多树脂及其他类中药可利用该方法快速鉴别,以扩大 NIRS 技术的适用范围,提高分析的准确性。该过程即为模型维护过程。

[参考文献]

[1] 陈孔荣,冯璟. 常用树脂及其他类中药的鉴别[J]. 浙江中医药大学学报,1998,22(2):51-52.
[2] 贾欣珠,孙云庆. 部分树脂及其他类中药的临床应用[J]. 山西医学教育,2002,6(5):55-56.
[3] 国家中医药管理局《中华本草》编委会. 中华本草. 第 1 册[M]. 上海:上海科学技术出版社,1999:426.
[4] 董重,夏厚林,石战英,等. 不同品种安息香红外光谱特征研究[J]. 四川中医,2010,28(7):37-39.
[5] 杨宏昕,郑敏. 树脂类中药及其化学鉴别[J]. 内蒙古中医药,2001,17(3):38-39.
[6] 李峰. 树脂及其他类中药的荧光光谱鉴别[J]. 山东中医药大学学报,1998,22(1):66-67.
[7] 赵中振,梁之桃. 近红外光谱技术在中药鉴定中的应用与优势[J]. 中国中药杂志,2012,37(8):86-87.
[8] 陆婉珍. 现代近红外光谱分析技术[M]. 北京:中国石化出版社, 2007:1.
[9] 李勇,魏益民,张波,等. 近红外水分稳健分析模型研究[J]. 光谱学与光谱分析, 2005, 25(12):1963-1967.
[10] 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):824-829.
[11] 雷雨,何东健,周兆永,等. 苹果霉心病可见/近红外透射能量光谱识别方法[J]. 农业机械学报,2016, 47(4):193-200.
[12] 陆婉珍. 近红外光谱仪器[M]. 北京:化学工业出版社,2010:35.
[13] HU L Q, YIN C L, ZENG Z P. Detection of adulteration in acetonitrile using near infrared spectroscopy coupled with pattern recognition techniques[J]. Spectrochim Acta A, 2015, 151(6):34-39.
[14] Vapnik V N. The nature of statistical learning theory[M]. New York:Springer-Verlag, 1995:35.
[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] 郑剑,周竹,仲山民,等. 基于近红外光谱的褐变板栗识别建模方法研究[J]. 食品科技,2016,42(1):285-290.

[责任编辑　邹晓翠]