基于近红外光谱的安胎丸生产年份预测方法

陈 蓓[1]，郑恩让[1*]，马晋芳[2]，葛发欢[3]，肖环贤[4]

1. 陕西科技大学电气与控制工程学院，陕西 西安 710021
2. 广州谱民信息科技有限公司，广东 广州 510006
3. 中山大学药学院，广东 广州 510006
4. 江西保利制药有限公司，江西 赣州 341900

摘 要 随着中药制剂存储时间的延长，其有效成分含量逐渐降低。化学检测手段损耗样品、检测时间长、成本高，利用近红外光谱对不同年份的经典名方安胎丸进行年份鉴别。为探讨这种无损、快速质量控制方式的可行性，采集了三年的105粒样本在1 000～1 799 nm波段近红外光谱吸光度数据，随机选择80个作为训练集，25个作为测试集。首先采用连续投影算法(SPA)，消除原始光谱数据中的冗余信息，对输入全光谱进行优化降维，根据测试集的内部交叉验证均方根误差值，从输入的800个波长中提取出11个特征波长，分别是：(1 692，1 714，1 405，1 001，1 114，1 478，1 514，1 788，1 202，1 014，1 164) nm；然后建立支持向量机(SVM)分类模型，由于SVM模型中的参数选取对分类正确率影响很大，利用粒子群优化(PSO)算法，对SVM模型中惩罚参数C和核函数参数进行寻优，形成PSOSVM分类模型；最后将SPA降维后的特征波长输入到PSOSVM分类算法中。用Matlab软件进行仿真测试，分别构建SVM，SPA-SVM和本文的SPA-PSOSVM三种方法分类模型，分类测试正确率分别达到了76%，92%和100%。从仿真结果可以看出，SPA波长优选可有效地降低光谱信息中存在的冗余信息，减少建模所需的时间，结合PSOSVM分类模型降低了模型的复杂度，提高分类精度。结果证实，依照所建立的利用近红外算法，可以准确无损区分中药制剂安胎丸生产的年份，该研究可为中药制剂年份间差异评价提供一种思路。

关键词 近红外光谱；安胎丸；年份预测；连续投影算法；粒子群优化结合支持向量机

中图分类号：R286.0 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)08-2592-06

引 言

安胎丸是经典古方，由当归、白芍、白术、川穹、黄芩等五味药材加工制成，具有安胎养血的功效[1]。随着人们经济和生活水平的提高，需求日益增加。市售安胎丸由于药品的原材料差异，并且随着药物生产和保存的时间、储存环境不同等都使药效差别较大，会对患者带来一定的损失，制剂的批间差异是其质量考量的重要指标。卫生部药品标准中采用薄层色谱法对其中三味药材进行定性鉴别，王雪丽等[2-3]主要采用高效液相色谱(HPLC)化学检测方法对部分质控成分的进行定量检测，这些质控方法不仅会损耗样品，且检测时间长，检测成本高。中药制剂成分多样，发生反应机制复杂，部分成分尚不明确，单一或者几种成分含量测定具有片面性。近红外光谱(near infrared spectroscopy，NIRS)是利用近红外光对化学物质中不同含氢基团的吸收信息来进行分析、检测的一种新型无损、快速、无污染的分析技术[4]，可作为一种整体质量评价的手段，在中医药质控领域，主要应用在品种鉴定、产地鉴别、品质分级和定量分析等方面[5-6]，对于药品生产年份质控方式的相关研究未见报道。

近红外光谱通常包含大量的波长变量，一定程度上还会引入噪声，通过筛选特征波长建立模型，可以降低模型的复杂程度，提高预测能力。常用的波长提取方法有遗传算法(GA)、无信息变量消除法(UVE)、连续投影算法(successive projection algorithm，SPA)等。王涛等[7]在预测胡杨叶含水量中建模对比，采用SPA波长选择算法预测精度和相关度都优于GA；经UVE筛选之后的波长变量数目仍然过于庞大，不能达到最终的简化目标，SPA算法能大大减少建模所

收稿日期：2019-09-08，**修订日期**：2019-12-30

基金项目：国家自然科学基金项目(31670596)和陕西科技大学博士科研启动基金项目(2019BJ-06)资助

作者简介：陈 蓓，1982年生，陕西科技大学电气与控制工程学院讲师 e-mail：chenbei@sust.edu.cn

*通讯联系人 e-mail：zhenger@sust.edu.cn

需变量的数目，比 GA 和 UVE 算法得到的变量数目更少，可提高建模的效率和速度。有研究利用 SPA 结合支持向量机(support vector machine，SVM)有效地对玉米的霉变程度进行了判别，预测准确率达到了 91.11%，但文中 SVM 参数是凭经验选取的合适参数，并不能保证参数是最佳的。鉴于此，本文提出一种基于近红外光谱技术的安胎丸生产年份预测方法，以某药厂三年的 105 粒安胎丸为研究对象，实验采集其近红外光谱，应用 SPA 算法去除光谱冗余信息，优选出样本的特征波长，结合粒子群优化(particle swarm optimization，PSO)算法对 SVM 分类模型进行参数寻优，建立 PSOSVM 分类预测模型。通过该方法可以区分安胎丸生产的不同年份，对药物的质量评价提供一种方法。

1　实验部分

1.1　仪器

SupNIR1500 近红外光谱仪(聚光科技(杭州)有限公司)，Matlab2018(美国 MathWorks 公司)，Ultimate3000 高效液相色谱仪(美国 Thermo 公司)。

1.2　样本采集

从 2013 年—2015 年安胎丸中随机抽取样品 15 个批次，批号为：130501，130502，130601，130602，130701，131101，131201，140401，140402，140501，140502，141103，151001，151002 和 151101，总共 105 丸样品。

利用美国 Thermo 公司的 Ultimate3000 高效液相色谱仪，测定安胎丸中关键质量控制成分的含量，列表统计如表 1 所示。

利用聚光科技有限公司生产的 SupNIR1500 近红外光谱仪采集安胎丸光谱，漫反射模式，波长扫描范围是 1 000～1 799 nm。每丸样品重复扫描三次，得到平均的光谱数据保存。105 个样品的光谱如图 1 所示。从图 1 可以看出来，近红外光谱信息重叠严重，特别是 2013 年和 2014 年的光谱，很难从峰值位置直观鉴别各样品的特征信息。因此，必须采用合适的特征提取办法，才能对安胎丸样品进行年份的鉴别。

表 1　安胎丸中关键质控成分的含量测定统计表(mg·pill^{-1})

Table 1　Statistical table for determination of key quality control indicator components in Antaipills (mg·pill^{-1})

年份	统计量	阿魏酸	黄芩苷	汉黄芩苷	黄芩素	汉黄芩素	洋川芎内酯 A
2013	最小值	0.129 6	9.727 2	1.050 6	0.678 6	0.96	0.053 4
	最大值	0.222 6	16.702 8	2.023 8	1.944 6	1.861 2	1.225 2
	平均值	0.172 3	13.493 9	1.607 0	1.226 4	1.485 0	0.575 4
2014	最小值	0.645 0	30.163 8	3.712 8	4.957 2	2.598 0	0.720 0
	最大值	0.774 6	49.539	6.308 4	9.080 4	4.179 6	1.293 6
	平均值	0.718 4	37.280 3	4.920 5	7.205 4	3.509 3	0.946 5
2015	最小值	0.448 3	52.465 8	7.289 4	4.571 4	2.647 8	0.979 2
	最大值	0.708 0	64.145 4	8.613 0	6.765 0	3.567 6	1.507 8
	平均值	0.660 0	59.632 9	8.123 5	5.694 7	3.271 3	1.299 7

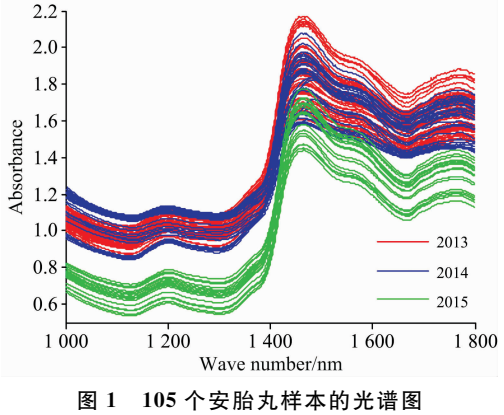

图 1　105 个安胎丸样本的光谱图

Fig. 1　The spectra of 105Antai pills samples

1.3　数据预处理

根据年份不同，对安胎丸样品对应的光谱数据进行分类，可以将样品分为三类，得到的具体分类表如表 2。

表 2　样本根据年份分类表

Table 2　The table of sample classification according to year

年份	样本数量	分类类别
2013	50	1
2014	34	2
2015	21	3

从表 1 中数据可以看出，生产年份不同，存放的时间不同，关键质控指标成分里的含量也不同，体现了生产年份与安胎丸的质量有密切关系，进一步说明按年份分类研究对药物的质量控制有一定的意义。

1.4　建模方法与模型评价

近红外光谱测量得到的数据数量庞大，光谱波长较多，相邻波长间存在较多的冗余信息和相关性，如果直接用全光谱建模，必然会使得建模的时间和模型的复杂度增加，模型的预测正确率和稳定性降低。

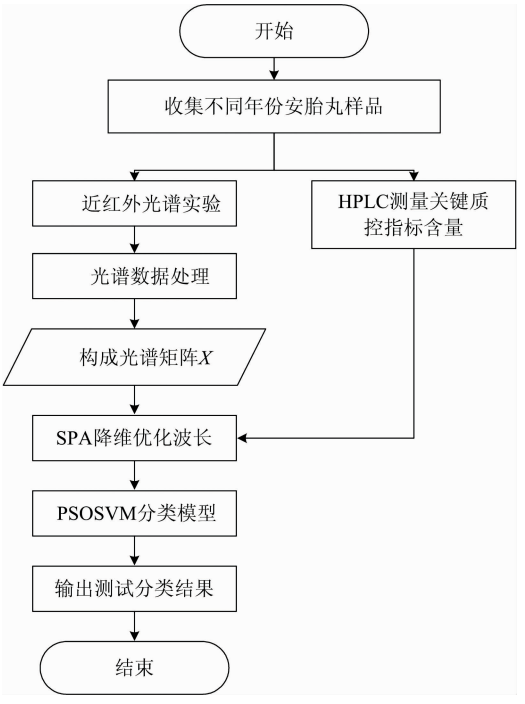

图 2 实验数据处理流程图

Fig. 2 The flow chart of processing experimental data

具体的实验建模分析过程如图 2 所示。先使用连续投影算法(SPA)对采集到的波长进行优化，对输入进行降维，最大程度的消除干扰。得到降维后的光谱数据输入到 PSOSVM 分类模型中，进行分类预测。利用分类正确率作为模型的评价准则，其定义为

$$\text{Accuary} = \sum_{i=1}^{3} N_i / N \tag{1}$$

在式(1)中 N 为测试集样本数量，N_i 为识别第 i 类分类正确的样本数量。

1.4.1 波长优化选择算法 SPA

连续投影算法[8](successive projection algorithm，SPA)属于前向选择变量算法，首先选择一个波长变量作为初始值，计算该变量在未选变量上的投影，将最大投影向量对应的波长作为新的待选变量，依次迭代，直到内部交互验证均方根误差(RMSECV)达到最小，选出最佳波长变量数 N 及波长变量集合 Y。SPA 算法实现步骤如下：

(1) 光谱矩阵 $\boldsymbol{X}^{n\times p}$(其中 n 为样本数，p 为待选波长变量)标准化；

(2) 随机选取初始迭代波长：从 $\boldsymbol{X}$ 中随机选择一个列向量 X_j，记为列向量 $X_{k(0)}$，$k(0)=1,2,3\cdots,p$；

(3)将光谱矩阵 $\boldsymbol{X}$ 中剩余的列向量记为 S，$S=\{j,1\leqslant j\leqslant p,j\notin k(0),k(1),\cdots,k(n-1)\}$；

(4) 分别计算 x_j 对剩余列向量 S 的投影 Px_j，

$$Px_j = x_j - (x_j^{\mathrm{T}} x_{k(p-1)}) x_{k(p-1)} (x_{k(p-1)}^{\mathrm{T}} x_{k(p-1)})^{-1} \tag{2}$$

(5)定义 $k(n)=\arg[\max(\| Px_j \|, k\in s)]$为 $N-1$ 个投影值中的最大；

(6)将最大投影值对应的波长变量作为下次迭代的初始值：

$$x_j = Px_j,\ j \in S$$

(7) 令 $i=i+1$，如果 $i<N$，则返回到步骤 3 循环进行计算；

(8) 将优化降维后得到的所有波长组合在一起，表示为集合 Y：

$$Y = \{x_{k(n)}: n = 1,2,\cdots,N-1\}$$

SPA 算法可以从全部波段里提取出特征波长，能够几乎消除原始光谱矩阵中的冗余信息，将优化降维后的特征波长输入到后面的建模中，能够显著增加模型正确率和运算速度。

1.4.2 分类建模算法 PSOSVM

支持向量机(support vector machine，SVM)是在统计学的 VC 维理论基础上发展起来的机器学习方法，在 1995 年由 Vapnik 首先提出[9]，具有理论完备、分类准确率高、泛化性能好等优点，能够解决小样本、非线性和高维数据划分的问题，主要用于模式识别和非线性回归，它的思想是建立一个分类超平面作为决策曲面，使正例和反例之间的隔离边缘最大化。SVM 引入核函数 $K(x,x_i)$巧妙地解决了非线性分类问题，SVM 有很多不同的核函数，由于径向基 RBF 核函数能够逼近任何非线性函数，具有很好的学习能力，因此选择 RBF 核函数，表达式如式(3)

$$K(x,x_i) = \exp(-\gamma \| x - x_i \|^2),\ \gamma > 0 \tag{3}$$

但是，SVM 模型中惩罚参数 C 和核函数参数 γ 的选取对分类的正确率影响很大。仅依赖于经验值的试凑是不可行的。

粒子群优化(particle swarm optimization，PSO)算法是计算智能领域基于群体智能的优化算法之一，算法概念源于对人工生命和鸟群捕食行为的研究[10]，采用仿生智能算法进行参数寻优，不用遍历所有参数组，目前已经广泛应用于神经网络、函数优化等其他算法中。

PSOSVM 算法的流程如图3所示，将PSO算法用于支

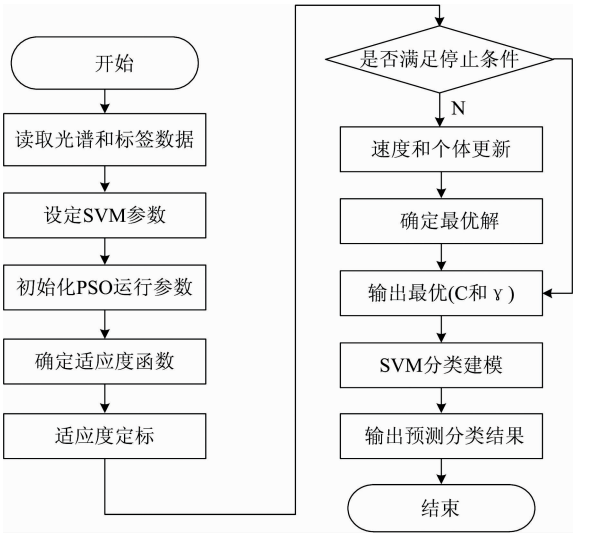

图 3 PSOSVM 算法的流程图

Fig. 3 Flow chart of the PSOSVM algorithm

持向量机的参数优化，可以降低优化过程的计算代价，提高分类的正确率[11]。

2　结果与讨论

2.1　优选特征波长

实验采集的近红外光谱波长扫描范围 1 000～1 799 nm，共有 800 个波长变量，如果直接作为分类模型的输入，输入量过大，训练时间过长。SPA 通过最小化变量之间的共线性，实现了最优波长的选择，通过 SPA 对近红外光谱数据进行降维。

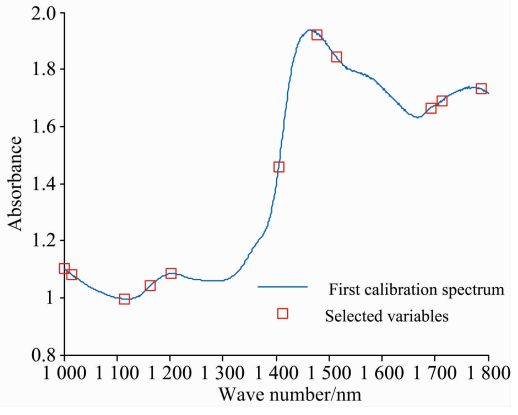

图 4　优选特征波长分布图

Fig. 4　The distribution map of preferred characteristic wavelength

针对上述 105 丸样品，随机选择 80 个作为训练集，25 个为测试集，数据随机分布，符合建模要求。选取质控成分中含量最高，且含量聚类与年份分类吻合的黄芩苷的含量为基准，根据测试集的内部交叉验证均方根误差值最小，由图 4 可看出，提取最佳的特征波长有 11 个，分别是：{1 692，1 714，1 405，1 001，1 114，1 478，1 514，1 788，1 202，1 014，1 164} nm，其重要程度依次递减。

2.2　分类模型分析对比

利用训练集 80 个样本的光谱数据及年份分类，采用 SPA-PSOSVM 算法建立安胎丸生产年份鉴别的分析模型，以分类的正确率作为评价准则。随机给定 SVM 分类模型参数，其中惩罚参数 $C=1$ 和核函数参数 $\gamma=1$，同时将建模结果与 SVM 和 SPA-SVM 两种算法作对比，得到的测试集的分类结果对比如图 5—图 7 所示。

为更清楚地对比三种方法的效果，汇总仿真中部分变量、参数和结果，如表 3 所示(其中黑体字表示本文所用方法)。

从表 3 对比可知，近红外光谱数据通过三种方法分类建模测试，第一种方法，单一的 SVM 建模参与变量数目最多，正确率最低；第二种方法，全光谱数据经过 SPA 降维后，变量数目从 800 个降到 11 个，再利用 SVM 建模，正确率提高到 92%，模型耗时大大缩短，体现了 SPA 算法可有效提高正确率和降低建模时间；第三种方法，即本方法，光谱经 SPA 降维后，再通过 PSO 寻优，SVM 分类算法的最佳惩罚参数 C 和核函数参数 γ(Best $C=16.406\ 4$，$\gamma=0.773\ 39$)，正确率达到了 100%，由于寻优过程消耗时间，模型耗时高于单一的 SVM 分类方法。综合考虑，基于近红外光谱结合 SPA-PSOSVM 建立安胎丸的年份分类预测模型正确率最高，性能较好(见图 7)。

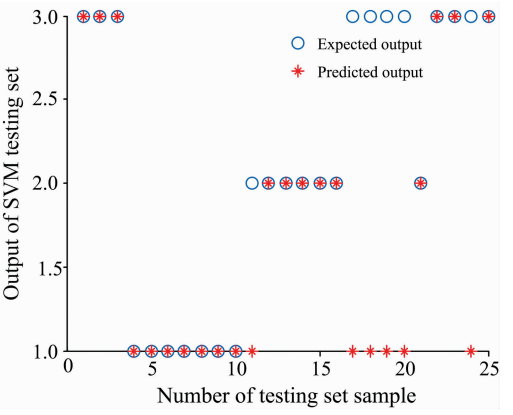

图 5　基于 SVM 的测试集分类结果

Fig. 5　The graph of test classification result based on SVM

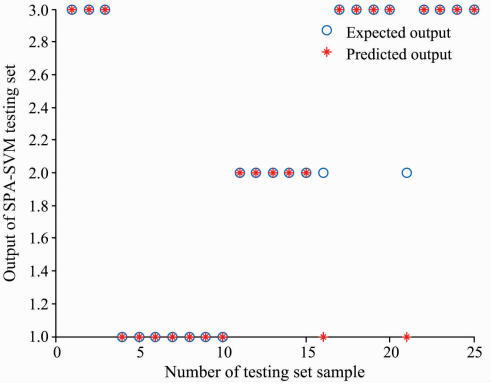

图 6　基于 SPA-SVM 的测试集分类结果

Fig. 6　The graph of test classification result based on SPA-SVM

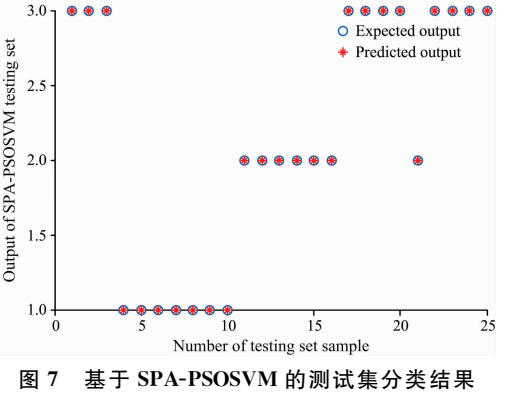

图 7　基于 SPA-PSOSVM 的测试集分类结果

Fig. 7　The graph of test classification result based on SPA-PSOSVM

表 3 三种方法测试效果对比表

Table 3 Table of testcomparison of three methods

方法	C	γ	建模波长数/个	分类错误数/个	模型耗时/s	正确率/%
SVM	1	1	800	6	13.987 1	76
SPA-SVM	1	1	11	2	0.566 9	92
SPA-POSSVM	**16.406 4**	**0.773 39**	**11**	**0**	**18.616 2**	**100**

3 结 论

对安胎丸近红外光谱数据进行 SPA 优化降维，在全波段提取了 11 个特征波长，占全部波长的 1.375%，建立了 PSOSVM 分类模型，预测模型正确率达到了 100%，表明 SPA 算法是一种比较有效的特征波长提取办法，建立的 SPA-PSOSVM 分类模型正确率明显高于 SVM 和 SPA-SVM，可以有效地用于安胎丸的生产年份分类预测，为药材原材料的优选、厂家生产工艺的革新和伪劣过期药品的判别提供参考和依据。根据中药的主要质控成分含量随存储时间的变化特点，本方法可为中药的质量评价提供一种快速无损的判定方式。

References

[1] Drug Specifications Promulgated by the Ministry of Public Health, PR China(卫生部药品标准). Traditional Chinese Medicine Preparation(中药成方制剂). Pharmacopoeia Commission of the Peoples Repbulic of China Press(中华人民共和国药典委员会编). Vol Ⅰ, 1989: 76.

[2] WANG Xue-li, MA Jin-fang, PENG Yin, et al(王雪利, 马晋芳, 彭 银, 等). Journal of Chinese Medicinal Materials(中药材), 2018, 41(1): 155.

[3] CHEN Can-wen, SONG Fen-yun, LI Hua, et al(陈粲文, 宋粉云, 李 华, 等). Chinese Journal of Pharmaceutical Analysis(药物分析杂志), 2016, 36(3): 465.

[4] YAN Yan-lu, CHEN Bin, ZHU Da-zhou, et al(严衍禄, 陈 斌, 朱大洲, 等). Near Infrared Spectroscopy-Principle, Technology and Application(近红外光谱分析的原理、技术与应用). Beijing: China Light Industry Press(北京: 中国轻工业出版社), 2013.

[5] LI Wen-long, QU Hai-bin(李文龙, 瞿海斌). Journal of Zhejiang University·Medical Sciences(浙江大学学报·医学版), 2017, 46(1): 80.

[6] XIAO Xue, LI Jun-shan, ZHANG Bo, et al(肖 雪, 李军山, 张 博, 等). Acta Scientiarum Naturalium Universitatis Nankaiensis(南开大学学报), 2017, 50(3): 44.

[7] WANG Tao, BAI Tie-cheng, YU Cai-li, et al(王 涛, 白铁成, 喻彩丽, 等). Jiangsu Agricultural Sciences(江苏农业科学), 2018, 46(19): 269.

[8] Gabriela Krepper, Florencia Romeo, David Douglas de Sousa Fernandes, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 189: 300.

[9] Vapnik V N. IEEE Transactions on Neural Networks, 1999, 10(5): 988.

[10] Ardjani F, Sadouni K, Benyettou M. International Workshop on Database Technology and Applications, IEEE, 2010, 2: 32.

[11] WANG Dao-ming, LU Chang-hua, JIANG Wei-wei, et al(王道明, 鲁昌华, 蒋薇薇, 等). Journal of Electronic Measurement and Instrumentation(电子测量与仪器学报), 2015, (4): 611.

Prediction Method for Production Year of Antai Pills Based on Near Infrared Spectroscopy

CHEN Bei[1], ZHENG En-rang[1*], MA Jin-fang[2], GE Fa-huan[3], XIAO Huan-xian[4]

1. School of Electrical and Control Engineering, Shaanxi University of Science & Technology, Xi'an 710021, China
2. Guangzhou Pumin Information Technology Co., Ltd., Guangzhou 510006, China
3. School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510006, China
4. Jiangxi Poly Pharmaceutical Co., Ltd., Ganzhou 341900, China

Abstract With the increase of the storage time of traditional Chinese medicine, the content of its effective components gradually decreases. Chemical detection means to consume the samples, with a long period and a high cost. In this paper, near infrared spectroscopy was used to identify the years of Antai pills of the classical prescriptions with different years. In order to explore the feasibility of this nondestructive and rapid quality control method, the absorbance data of 105 samples in 1 000～1 799 nm band near infrared spectroscopy of three years were collected, 80 samples were randomly selected as training sets and 25 samples as test sets. Firstly, the Successive Projection Algorithm (SPA) was adopted to eliminate the redundant information in the original spectral data, and the full input spectrum was optimized and the dimensionality was reduced. According to the internal of the test sets, the error value of the root mean square was cross-verified, 11 characteristic wavelengths were extracted from 800 wavelengths, respectively: (1 692, 1 714, 1 405, 1 001, 1 114, 1 478, 1 514, 1 788, 1 202, 1 014, 1 164) nm. Then the Support Vector Machine (SVM) classification model was established. Since the selection of the parameters in SVM model has a great influence on the classification accuracy, the Particle Swarm Optimization (PSO) algorithm was used to optimize the penalty parameter C and the kernel function parametersin SVM model to form PSOSVM classification model. Finally, after SPA dimension reduction, the characteristic wavelength was input into PSOSVM classification algorithm. Matlab software was used in the simulation test, and SVM, SPA-SVM classification models and SPA-PSOSVM classification model in this paper were respectively constructed. The classification test accuracy reached 76%, 92% and 100% respectively. From the simulation results, it can be seen that the SPA wavelength optimization could effectively reduce the redundant spectral information and reduce the time required for modeling. The PSO-SVM classification model, the complexity of the model was reduced and the classification accuracy was improved. The results show that the near infrared algorithm established in this paper could accurately and nondestructively distinguish the production years of the traditional Chinese medicine Antai pills, and this study could provide a way of thinking for the evaluation of the differences of the years of traditional Chinese medicine.

Keywords Near infrared spectroscopy; Antai pills; Year classification; Successive projection algorithm (SPA); Particle swarm optimization combined with support vector (PSOSVM)

(Received Sep. 8, 2019; accepted Dec. 30, 2019)

* Corresponding author