

基于粒子群算法的最小二乘支持向量机 在红花提取液近红外定量分析中的应用

金 叶 杨 凯 吴永江 刘雪松 陈 勇*

(浙江大学药学院 杭州 310058)

摘 要 提出一种基于粒子群算法的最小二乘支持向量机(PSO-LS-SVM) 方法,用于建立红花提取过程关键质控指标的定量分析模型。近红外光谱数据经波段选择、预处理和主成分分析(降维) 后,利用粒子群优化(PSO) 算法对最小二乘支持向量机算法中的参数进行优化,然后使用最优参数建立固含量和羟基红花黄色素 A(HSYA) 浓度的定量校正模型。将校正结果与偏最小二乘法回归(PLSR) 和 BP 神经网络(BP-ANN) 比较,并将所建的 3 个模型用于红花提取过程未知样品的预测。结果表明,BP-ANN 校正结果优于 PSO-LS-SVM 和 PLSR,但是对验证集和未知样品集的预测能力较差,而 PSO-LS-SVM 和 PLSR 模型的校正、验证结果相近,相关系数均大于 0.987, RMSEC 和 RMSEP 值相近且小于 0.074, RPD 值均大于 6.26, RSEP 均小于 5.70%。对于未知样品集, PSO-LS-SVM 模型的 RPD 值大于 8.06, RMSEP 和 RSEP 值分别小于 0.07% 和 5.84%, 较 BP-ANN 和 PLSR 模型更低。本研究所建立的 PSO-LS-SVM 模型表现出较好的模型稳定性和预测精度,具有一定的实践意义和应用价值,可推广用于红花提取过程的近红外光谱定量分析。

关键词 近红外光谱; 粒子群优化; 最小二乘支持向量机; 红花提取液

1 引 言

红花注射液主要成分为红花提取物,具有抗凝、防栓、扩张血管和有效防治心脑血管疾病的作用。羟基红花黄色素 A(Hydroxysafflor yellow A, HSYA) 是红花的主要活性成分^[1],药理实验证明: HSYA 能明显提高缺氧耐受力,使冠脉扩张,增加冠脉流量,并有明显抑制血小板聚集作用^[2-3]。提取工艺是红花注射液生产过程的起点,直接关系到红花药材的利用率。目前,提取工艺的质量控制主要依靠经验和传统质量分析方法(HPLC 等),耗时繁琐。故开发红花提取过程中关键质控指标的快速测定方法,有助于解决红花提取过程中关键的质量控制问题,对于中药工业技术进步和产品质量升级具有重要的意义。

近红外(Near infrared, NIR) 光谱技术作为一种快速无损的绿色分析技术,具有分析快速、样品处理简单、无需消耗试剂等特点。近年来,近红外光谱技术已经越来越多的被应用于中药研究,包括药材产地鉴别、有效组分含量测定和制药过程的在线检测和监控^[4-7]。使用近红外进行定量分析时必须建立定量校正模型。目前比较常用的近红外定量分析建模方法有偏最小二乘回归法(Partial least square regression, PLSR)、人工神经网络法(ANN) 和支持向量机法(SVM) 等。SVM 是建立在统计学习理论(Statistical learning theory, SLT) 基础上的一种机器学习算法,已在 NIR 光谱分析中得到广泛应用^[8-9]。最小二乘支持向量机(LS-SVM) 是经典 SVM 的一种改进,以等式约束代替标准 SVM 算法中的不等式约束,并将误差平方和损失函数作为训练集的经验损失,将求解二次规划问题转化为求解线性方程组问题,提高了求解的速度和收敛精度^[10]。针对 LS-SVM 的参数选取会对结果产生较大影响,参数的选取具有一定的“盲目性”的问题,采用试凑法或遍历优化进行参数选择费时而未必能找到全局最优解,提出了利用粒子群优化(PSO) 算法优化其模型参数的方法^[11-13]。

本研究基于粒子群算法的最小二乘支持向量机(PSO-LS-SVM),建立了红花提取过程中质控指标——固含量和 HSYA 浓度的定量分析模型,并将预测结果与近红外光谱分析常用的 PLSR 和 ANN 模型进行比较,实现红花提取过程未知样品中固含量和 HSYA 浓度的快速检测。

2011-08-16 收稿; 2011-12-09 接受

本文系浙江省重大科技计划项目(No. 2008C03005) 和国家“十一五”科技支撑计划项目(No. 2006BAI06A08) 资助

© 1994-2012 China Electronic Journal Electronic Publishing House. All rights reserved. <http://www.cnki.net>

2 建模方法与原理

LS-SVM 是一种新型支持向量机,为标准 SVM 的一种扩展。LS-SVM 能够进行线性和非线性的多元建模,降低计算的复杂程度,加快求解速度,已在 NIR 光谱分析中得到广泛的应用^[14~16]。LS-SVM 的算法原理及具体建模过程参见文献[10,17,18]。由于径向基核函数(Radial basis function, RBF)作为非线性函数能够减少训练过程中计算的复杂性,本研究选择 RBF 核函数进行 LS-SVM 计算。影响 LS-SVM 模型性能的因素通常有两个,即核函数参数 σ (核宽度)和正则化参数 C 的取值。正则化参数 C 控制对样本超出计算误差的惩罚程度,而 σ 则控制函数回归误差,并且直接影响初始的特征值和特征向量。 σ 过小会导致过拟合,相反; σ 过大则模型过于简单,从而影响预测精度^[19]。因此,为了提高 LS-SVM 的学习和泛化能力,需要对核函数参数 σ 和正则化参数 C 进行优化。

粒子群优化(PSO)算法,最早是 Kennedy 等于 1995 年提出的全局随机优化技术^[12],其基本概念源于对鸟群捕食行为的研究。PSO 具有不易陷入局部极小、易实现和调整参数较少的优点,从而用于求解非线性、不可微和多模态问题等。PSO 中,每个优化问题的潜在解称之为“粒子”。每个粒子都有自己的位置和速度(决定它的飞行方向和距离)以及一个由被优化的函数决定的适应值。PSO 初始化为一群随机粒子(随机解),在每一次迭代中,粒子通过跟踪两个极值更新自己。第一个就是粒子本身所找到的最优解,这个解称为个体极值(pbest)。另一个极值是整个种群目前找到的最优解,这个极值是全局极值(gbest)。本研究中,设定粒子群中包含 l 个粒子,粒子 i ($i \in \{1, 2, \dots, l\}$) 的信息可以用 D 维向量(即算法所优化的参数个数)表示,空间位置为 $x^i = (x_1^i, \dots, x_d^i, \dots, x_D^i)$,速度为 $v_i = (v_1^i, \dots, v_d^i, \dots, v_D^i)$,在求得 pbest 和 gbest 这两个最优解后,粒子根据式(1)和式(2)来更新自己的速度和位置。

$$v_d^i(t+1) = (\omega v_d^i(t) + c_1 r_1(t)(p_{\text{best}}^i(t) - x_d^i(t)) + c_2 r_2(t)(g_{\text{best}}^i(t) - x_d^i(t))) \quad (1)$$

$$x_d^i(t+1) = x_d^i(t) + v_d^i(t+1) \quad (2)$$

其中 $v_d^i(t)$ 表示 t 时刻粒子 i 在 d 维空间的速度, $x_d^i(t)$ 表示 t 时刻粒子 i 在 d 维的空间位置。 $p_{\text{best}}^i(t)$ 为单个粒子 i 在所优化的第 d 个参数中的历史最优解, $g_{\text{best}}^i(t)$ 是所有粒子在所优化的第 d 个参数中的历史最优解。 c_1 和 c_2 为加速因子,通常取值范围为 $(0, 2)$ 。 r_1 和 r_2 为两个 $(0, 1)$ 之间变化的相对独立的随机函数。 ω 为权重因子,其值非负,值的大小影响整体寻优能力。

在每一次迭代过程中,每个粒子都需要根据目标函数计算其适应值,然后根据适应值确定当前粒子最优位置 $p_{\text{best}}^i(t)$ 及群体最优位置 $g_{\text{best}}^i(t)$,再通过式(1)和式(2)调整各个粒子的速度及位置。其结束条件为迭代次数达到设定值或者群体迄今为止搜索到的最优位置满足预设最小适应值。本研究中设定最大迭代次数为 100 代,预设适应值 = 0。在实验中通过 LS-SVM 建立定量分析模型,利用粒子群算法对 LS-SVM 的参数寻优,并以均方误差(Mean squared error, MSE)作为目标函数^[20]。

3 实验部分

3.1 仪器与试剂

Agilent 1200-DAD 高效液相色谱仪; Antaris 傅立叶变换近红外光谱仪(Thermo Nicolet, USA),配有透射检测器、采样系统以及 Result、TQ Analyst 等数据处理软件。

红花药材(山西亚宝药业股份有限公司提供);羟基红花黄色素 A 标准品(成都曼斯特生物科技有限公司);甲醇、乙腈(色谱纯, Merck 公司); H_3PO_4 (分析纯); Millipore 超纯水。

3.2 近红外光谱采集

称取红花药材 200 g 于三颈烧瓶内,加入 15 倍量(3000 mL)水,煎煮 1 h。煎煮前 30 min 每隔 2.5 min 收集一次提取液 8 mL,并补 8 mL 水。后 30 min,每隔 5 min 收集一次提取液 8 mL,并补 8 mL 水。不同批次红花药材的提取实验重复 7 次,共获得 126 个样品。建完模型后重复一次提取实验,将采集到的数据(样品数量为 18)作为未知样品集,用于模型适用性的研究。

近红外仪器有关参数设置:吸光度数据格式为 Absorbance,扫描次数为 32,分辨率为 4 cm^{-1} ,光纤透

射式探头光程 2 mm,以空气为参比。扫描光谱范围为 $4500 \sim 12000 \text{ cm}^{-1}$ 。

3.3 HPLC 定量测定^[21]

红花提取液以 1500 r/min 高速离心 10 min ,取上清液,用于分析 HSYA 浓度。

色谱条件: Agilent eclipse C_{18} 分析柱 ($250 \text{ mm} \times 4.6 \text{ mm}$, $5 \mu\text{m}$); 甲醇-乙腈- $0.7\% \text{ H}_3\text{PO}_4$ 溶液 ($26:2:72$, V/V) 为流动相; 流速 1 mL/min ; 检测波长 403 nm ; 柱温 40°C ; 进样量 $5 \mu\text{L}$ 。

标准曲线绘制: 精密称取 HSYA 对照品 5.8 mg 采用流动相溶解并定容至 10 mL 梯度稀释后, 分别进样分析。以浓度为横坐标, 峰面积为纵坐标, 绘制标准曲线。

3.4 固含量测定

称定烘干至恒重的扁形瓶(两次烘干后重量 X_0 的差值小于 5 mg) 量取 4 mL 提取液至扁形瓶, 称重 (X_1) 水浴蒸干, 105°C 烘 5 h , 取出置于干燥器内冷却 30 min , 迅速称重 (X_2)。按式 (3) 求得固含量 (Y)。

$$Y = 100\% \times (X_2 - X_1) / (X_1 - X_0) \quad (3)$$

3.5 数据处理方法与模型性能评价指标

选择合适的波段, 在适宜的光谱预处理基础上, 采用偏最小二乘回归, BP 神经网络 (BP-ANN) 和基于粒子群算法的最小二乘支持向量机 (PSO-LS-SVM) 建立近红外数据与固含量和 HSYA 浓度这两个质控指标之间的定量校正模型。以相关系数 (R)、校正集均方差 (Root mean square error in calibration, RMSEC)、预测均方差 (Root mean square error in prediction, RMSEP)、相对偏差 (Relative standard errors of prediction, RSEP) 为指标优化建模参数, 考察模型性能。RMSEC, RMSEP, RSEP 和 R 的计算方法见文献 [22-23]。此外, 使用相对分析误差 (Residual predictive deviation, RPD) 对模型进行深入评价。当 RPD 值大于 3 时认为模型具有较好的预测能力, 可以进行指标的定量控制^[24]。

4 结果与讨论

4.1 质控指标测定结果

建模样品中固含量和 HSYA 浓度的测定结果见图 1。提取过程中固含量和 HSYA 浓度均随着提取进程而不断增加, 并在一定时间后趋于稳定。但是, 在到达提取时间终点后, 不同批次提取液中的固含量和 HSYA 浓度均存在一定的差异, 主要原因在于不同批次间药材质量存在差异。

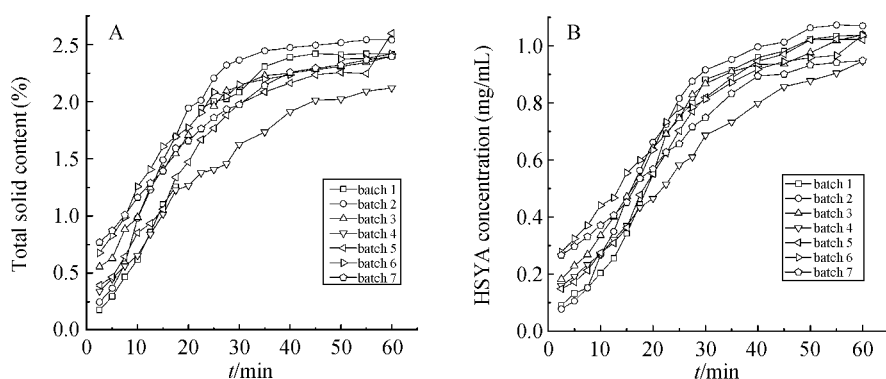


图 1 不同批次样品提取过程中固含量 (A) 和 HSYA 浓度 (B) 的变化曲线

Fig. 1 Time evolution of total solid content (A) and hydroxysafflor yellow A (HSYA) concentration (B) for different batch samples

4.2 光谱分析

红花提取过程中采集的原始近红外光谱见图 2。由于红花采用水提法, 水含有 OH 基, 极性很强, 在近红外谱区的 1440 nm (6944 cm^{-1}) 和 1940 nm (5155 cm^{-1}) 附近有很强的合频与倍频吸收谱带, 形成两大“水峰”^[25], 即 $4500 \sim 5400 \text{ cm}^{-1}$ 和 $6500 \sim 7500 \text{ cm}^{-1}$ 光谱区间。在这两个波段内, 其它各种物质分子的倍频与合频吸收相对较弱, 因此对水溶液物质体系进行近红外吸收光谱分析时, 应尽可能减少来自

溶剂吸收的干扰。此外,吸收度大于 1.5 的波长区域属于饱和吸收,建模时不宜采用该区域内的信息,因此可用的波段为 $5400 \sim 6500 \text{ cm}^{-1}$ 和 $7500 \sim 12000 \text{ cm}^{-1}$ 。

为了提高模型的预测准确性,在对近红外光谱进行分析之前,一般要对光谱数据进行预处理。本研究采用导数法^[26]对光谱进行预处理。将近红外光谱进行一阶导数处理能有效消除基线偏移,减少峰与峰之间的重叠,获取更多的有效信息,但与此同时导数的计算会引进噪音,降低信噪比,而平滑可以降低高频随机噪声。因此,本研究将一阶导数法和 Norris 平滑滤波结合使用对光谱数据进行预处理。由于 $7500 \sim 12000 \text{ cm}^{-1}$ 区间的光谱存在较大的噪声,且无显著的特征吸收。因此,选择 $5400 \sim 6500 \text{ cm}^{-1}$ 波段用于固含量和 HSYA 浓度定量模型的建立。为了保证光谱选择区域的准确性,对光谱和质控指标的相关系数进行了考察。以 HSYA 为例(图 3)相关系数大于 0.6 的光谱主要集中在所选择的光谱区域内。此外,对于固含量指标也有类似的相关结果。

建立定量校正模型时,为减少计算量提高运算效率,采用主成分分析技术对光谱数据进行降维处理。当固含量和 HSYA 模型的主成分数分别为 4 和 8 时,其累计总贡献率分别达到 99.57% 和 99.75%,能够较好的反映原有光谱信息。因此,对于固含量和 HSYA 模型,分别选择贡献率较大的前 4 和 8 个主成分数作为新变量输入模型。

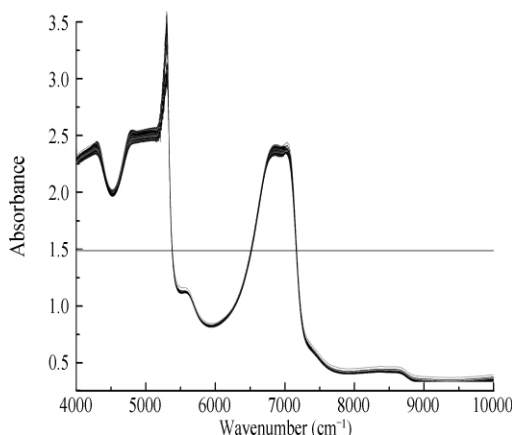


图 2 红花提取液的原始近红外光谱

Fig. 2 Raw spectra of extraction solution of safflower

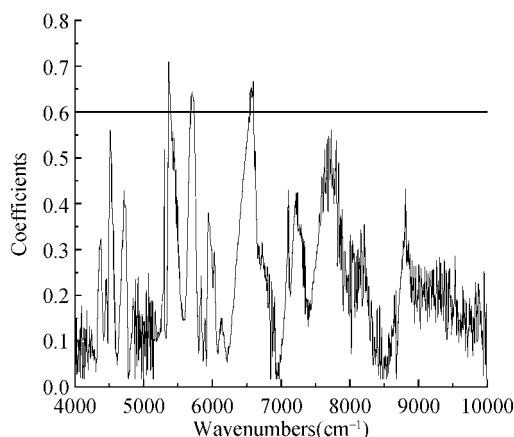


图 3 近红外光谱和 HSYA 浓度的线性关系图

Fig. 3 Correlogram of NIR spectra and HSYA concentrations

4.3 定量模型的建立

在 $5400 \sim 6500 \text{ cm}^{-1}$ 波段范围内,采用一阶导数和 Norris 平滑滤波预处理方法,将随机抽取的 6 批提取实验的数据用于建立 PSO-LS-SVM, BP-ANN 和 PLSR 模型。构建两层结构的 BP 网络,对于固含量和 HSYA 模型,输入层的神经元个数分别为 4 和 8 个,即选定的主成分数(新变量)。输出层神经元个数为 1,即固含量(HSYA 浓度)。隐含层的神经元个数经过网络测试,分别确定为 20 和 15 个。输入层到隐含层的传递函数为正切 S 形传递函数(TANSIG),隐含层到输出层的传递函数为线性传递函数(PURELIN)。为加快网络训练,本研究采用 Levenberg-Marquardt 规则训练网络,学习速率取为 0.05,训练目标为 0.001,网络分别经过 36 和 6 次训练达到预定目标。对于 PSO-LS-SVM 模型,初始化粒子数设置为 100,PSO 最大循环迭代数为 100。本研究选择 RBF 作为核函数,它只有一个待定参数 σ 。因此,对于采用径向基核的最小二乘支持向量机,PSO 需要优化的参数是正则化参数 C 和核函数宽度 σ ($D=2$),这两个参数在很大程度上决定了最小二乘支持向量机的学习和泛化能力。其它参数设置: $c_1=c_2=2$, $\omega=0.5$ 。利用粒子群算法搜索到的最优的 LS-SVM 参数: $C=2636861.4$, $\sigma=348.7$ (固含量模型); $C=7127948.6$, $\sigma=118.2$ (HSYA 浓度模型)。根据以上参数优化结果,分别建立固含量和 HSYA 浓度的 PSO-LS-SVM, PLSR 和 BP-ANN 定量校正模型。优化后各模型的校正和验证结果比较(见表 1), PLSR 和 PSO-LS-SVM 模型的校正集和验证集相关系数均大于 0.987, RMSEC 和 RMSEP 值相近且小于 0.074, RPD 值均大于 6.26, RSEP 值均小于 5.7%。虽然 BP-ANN 校正结果优于 PLSR 和 PSO-LS-SVM,但是对验证集的预测效果较差,特别是固含量模型, RMSEP 值远大于 2 倍 RMSEC 值,且 RPD 小于 3,

RSEP 值超过 10% ,HSYA 模型也有类似现象。与 BP-ANN 模型相比 ,PLSR 和 PSO-LS-SVM 模型的校正和验证结果相近 ,均有较好的预测能力和模型稳定性。研究表明 ,采用近红外光谱技术快速测定红花提取液中固含量和 HSYA 浓度时 ,使用 PLSR 和 PSO-LS-SVM 算法建立的模型性能优于经典的 BP-ANN 算法 ,这两种算法的预测精度高 ,泛化能力强 ,不容易出现过拟合倾向。

表 1 优化后各模型的校正和验证结果比较
Table 1 Statistics of established models in calibration and validation set

模型 Models	提取物 Extraction	校正集 Calibration set				验证集 Validation set			
		R^2	RMSEC	RPD	RSEP (%)	R^2	RMSEC	RPD	RSEP (%)
BP-ANN	固含量 Solid content	0.9976	0.032	20.636	1.816	0.9926	0.229	2.350	12.695
	HSYA	0.9941	0.022	12.956	3.191	0.9805	0.062	3.905	9.126
PLSR	固含量 Solid content	0.9896	0.066	9.855	3.804	0.9961	0.062	8.621	3.460
	HSYA	0.9971	0.018	18.627	2.220	0.9924	0.035	6.970	5.112
PSO-LS-SVM	固含量 Solid content	0.9871	0.074	8.830	4.245	0.9942	0.050	10.677	2.794
	HSYA	0.9946	0.021	13.675	3.024	0.9953	0.039	6.260	5.692

PSO-LS-SVM: Particle swarm optimization-least squares support vector machine.

4.4 未知样品预测

将所建模型用于预测红花提取过程中的未知样品 ,预测结果如表 2 所示。通过对比 RMSEP , RPD 和 RSEP 值可以看出 ,PSO-LS-SVM 模型的预测准确度高于 BP-ANN 和 PLSR 模型。采用 PSO-LS-SVM 算法得到的固含量和 HSYA 模型的 RMSEP 值和 RPD 值均与校正集和验证集结果相似 ,并且 RMSEP 值小于 2 倍 RMSEC ,RPD 值大

表 2 未知样品集各模型的评价指标
Table 2 Statistics of the established models in unknown sample set

模型 Models	提取物 Extraction	未知样品集 Unknown sample set			
		R^2	RMSEP	RPD	RSEP (%)
BP-ANN	固含量 Solid content	0.9859	0.168	4.753	8.563
	HSYA	0.9903	0.046	7.828	6.008
PLSR	固含量 Solid content	0.9986	0.091	8.769	4.642
	HSYA	0.9821	0.051	7.039	6.681
PSO-LS-SVM	固含量 Solid content	0.9975	0.069	11.566	3.519
	HSYA	0.9913	0.044	8.064	5.832

于 8.06。此外 ,RSEP 值均小于 5.84% ,较 BP-ANN 和 PLSR 模型更低。使用 PSO-LS-SVM 模型预测未知样品集中固含量和 HSYA 浓度。如图 4 所示 ,近红外预测趋势与实际测定值的变化趋势基本一致。

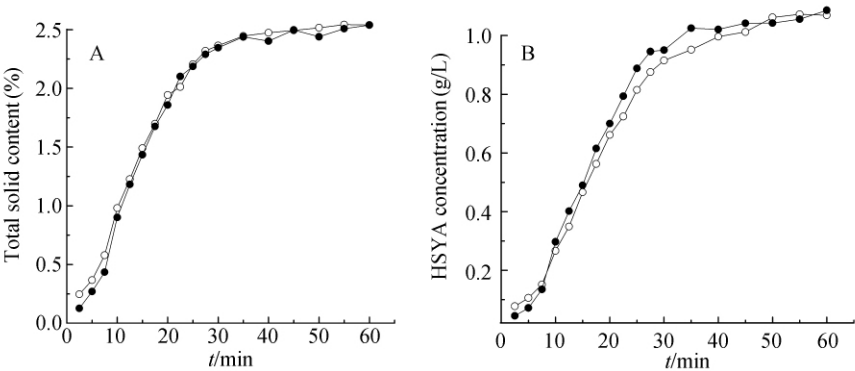


图 4 红花提取过程模型预测与实际测定趋势对照图(A. 固含量; B. HSYA)
Fig. 4 Evolution of total solid content (A) and HSYA concentration (B) during the extraction process of safflower
●: Predicted by NIR; ○: Measured by the reference assays.

5 结 论

本研究将 LS-SVM 理论与近红外光谱技术相结合,利用粒子群优化算法对 LS-SVM 进行参数寻优,建立红花提取过程中固含量和 HSYA 的定量分析模型,并与 PLSR 及 BP-ANN 建模结果作比较。结果表明,PSO-LS-SVM 模型具有稳定性好、预测准确度高、外推能力强及不存在过学习现象等优点。与传统的分析方法相比,所建立的方法快速、有效,能用于红花提取过程中固含量和 HSYA 浓度的快速测定。

References

- 1 LI Ying, ZHANG Ge, GUO Mei-Li, ZHANG Wei, ZHONG Xiang-Ping. *Academic Journal of Second Military Medical University*, **2005**, 26(5): 587 ~ 588
李颖,张戈,郭美丽,章伟,仲向平. 第二军医大学学报, **2005**, 26(5): 587 ~ 588
- 2 SHI Ming. *Modern Journal of Integrated Traditional Chinese and Western Medicine*, **2004**, 13(15): 2077 ~ 2079
时明. 现代中西医结合杂志, **2004**, 13(15): 2077 ~ 2079
- 3 Zhu H B, Wang Z H, Tian J W, Fu F H, Liu K, Li C L. *Acta Pharmaceutica Sinica*, **2005**, 40(12): 1144 ~ 1146
- 4 ZHANG Ju-Hua, ZHU Xiang-Yang, LI Gao-Yang, SHAN Yang, SHANG Yue-Bo, HUANG Lu-Hong, SHUAI Ming. *Chinese J. Anal. Chem.*, **2011**, 39(5): 748 ~ 752
张菊华,朱向英,李高阳,单扬,尚雪波,黄绿红,帅鸣. 分析化学, **2011**, 39(5): 748 ~ 752
- 5 Chan C O, Chu C C, Mok D K W, Chau F T. *Anal. Chim. Acta*, **2007**, 592(2): 121 ~ 131
- 6 QU Hai-Bin, LI Bin, LIU Xue-Song, CHENG Yi-Yu. *Chinese pharmaceutical Journal*, **2005**, 40(24): 1897 ~ 1903
瞿海斌,李斌,刘雪松,程翼宇. 中国药学杂志, **2005**, 40(24): 1897 ~ 1903
- 7 HUANG Ya-Wei, WANG Jia-Hua, LI Xiao-Yun, Jacqueline J Shan, Lei Ling, HAN Dong-Hai. *Spectroscopy and Spectral Analysis*, **2010**, 30(1): 2954 ~ 2957
黄亚伟,王加华,李晓云, Jacqueline J Shan, Lei Ling, 韩东海. 光谱学与光谱分析, **2010**, 30(1): 2954 ~ 2957
- 8 Thissen U, Pepers M, Ustun B, Melssen W J, Buydens L M C. *Chemom. Intell. Lab. Syst.*, **2004**, 73(2): 169 ~ 179
- 9 Chauchard F, Cogdill R, Roussel S, Roger J M, Bellon-Maurel V. *Chemom. Intell. Lab. Syst.*, **2004**, 71(2): 141 ~ 150
- 10 Suykens J A K, Vandewalle J. *Neural Process. Lett.*, **1999**, 9(3): 293 ~ 300
- 11 Clerc M, Kennedy J. *IEEE Trans. Evol. Comput.*, **2002**, 6(1): 58 ~ 73
- 12 Kennedy J, Eberhart R. *Proceedings of the IEEE International Conference on Neural Networks*, Perth, Australia, vol. 4, **1995**: 1942 ~ 1948
- 13 Guo X C, Yang J H, Wu C G, Wang C Y, Liang Y C. *Neurocomputing*, **2008**, 71(16 ~ 18): 3211 ~ 3215
- 14 Balabin R M, Lomakina E I. *Analyst*, **2011**, 136(8): 1703 ~ 1712
- 15 Shahlaei M, Fassihi A, Saghaie L. *Eur. J. Med. Chem.*, **2010**, 45(4): 1572 ~ 1582
- 16 Lai Y H, Ni Y N, Kokot S. *Vib. Spectrosc.*, **2011**, 56(2): 154 ~ 160
- 17 TU Zhen-Hua, ZHU Da-Zhou, JI Bao-Ping, MENG Chao-Ying, WANG Lin-Ge, QING Zhao-Shen. *Chinese J. Anal. Chem.*, **2010**, 38(1): 45 ~ 50
屠振华,朱大洲,籍保平,孟超英,王林舸,庆兆坤. 分析化学, **2010**, 38(1): 45 ~ 50
- 18 YANG Chun-Hua, QIAN Xiao-Shan. *Chinese J. Anal. Chem.*, **2010**, 38(6): 838 ~ 842
阳春华,钱晓山. 分析化学, **2010**, 38(6): 838 ~ 842
- 19 JIN Hang-Feng, HUANG Ling-Xia, Wu Di, JIN Pei-Hua, LOU Cheng-Fu. *J. Infrared Millim. Waves*, **2010**, 29(3): 216 ~ 219
金航峰,黄凌霄,吴迪,金佩华,楼程富. 红外与毫米波学报, **2010**, 29(3): 216 ~ 219
- 20 Yang Y, Chen R S, Ye Z B. *Microw. Opt. Techn. Lett.*, **2006**, 48(1): 141 ~ 144
- 21 National Commission of Chinese Pharmacopoeia. *Pharmacopoeia of People's Republic of China*. Beijing: Chemical Industry Press, **2005**, Vol I: 141
国家药典委员会. 中国药典. 北京: 化学工业出版社, **2005**, 一部: 141
- 22 Blanco M, Bano R G, Bertran E. *Talanta*, **2002**, 56(1): 203 ~ 212
- 23 Xiang D, Berry J, Buntz S, Gargiulo P, Cheney J, Joshi Y, Wabuyele B, Wu H Q, Hamed M, Hussain A S, Khan M A.

- J. Pharm. Sci.*, **2009**, 98(3): 1155 ~ 1166
- 24 Williams P C, Sobering D C. *How Do We Do It: a Brief Summary of the Methods We use in Developing Near Infrared Calibrations*. In: Davies A M C, Williams P C(Eds.), *Near infrared Spectroscopy: The Future Waves*. Chichester: NIR Publications, **1996**: 185 ~ 188
- 25 Streefland M, Van Herpen P F G., Van de Waterbeemd B, Van der Pol L A, Beuvery E C, Tramper J, Martens D E, Toft M. *Biotechnol. Bioeng.*, **2009**, 104(3): 492 ~ 504
- 26 ZHEN Yong-Mei, ZHANG Tie-Qiang, ZHANG Jun, CHEN Xing-Dan, SHEN Xuan-Guo. *Spectrosc. Spect. Anal.*, **2004**, 24(12): 1546 ~ 1548
- 郑咏梅, 张铁强, 张军, 陈星旦, 申铨国. *光谱学与光谱分析*, **2004**, 24(12): 1546 ~ 1548

Application of Particle Swarm Optimization Based Least Square Support Vector Machine in Quantitative Analysis of Extraction Solution of Safflower Using Near-infrared Spectroscopy

JIN Ye, YANG Kai, WU Yong-Jiang, LIU Xue-Song, CHEN Yong*
(College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China)

Abstract A novel particle swarm optimization(PSO) based least squares support vector machine(LS-SVM) method was investigated for quantitative analysis of extraction solution of safflower using near-infrared(NIR) spectroscopy. The usable spectral region(5400–6500 cm^{-1}) was identified, spectral preprocessing of Norris derivative smoothing was employed, and spectral dimension was also reduced through principal component analysis(PCA). In this paper, the PSO algorithm was applied to select the LS-SVM hyper-parameters(including the regularization and kernel parameters). The calibration models of total solid content and hydroxysafflor yellow A(HSYA) were established using the optimum hyper-parameters of LS-SVM. The performance of LS-SVM models was compared with partial least squares regression(PLSR) and back-propagation artificial neural networks(BP-ANN). The feasibility of these three methods was examined on the unknown sample set. Experimental results showed that the calibration results of BP-ANN were superior to PSO-LS-SVM and PLSR, however, the prediction accuracy of validation and unknown sample set was inferior. For PSO-LS-SVM and PLSR models, the correlation coefficients of the calibration and validation set were above 0.987, the RMSEC and RMSEP values were close to each other and less than 0.074, residual predictive deviation(RPD) values were all above 6.26, and the RSEP values were controlled within 5.70%. For the unknown sample set, the RPD values of PSO-LS-SVM models were above 8.06, the RMSEP and relative standard errors of prediction (RSEP) values were less than 0.07 and 5.84% respectively, which were much lower than BP-ANN and PLSR models. The PSO-LS-SVM algorithm employed in this paper exhibited excellent model robustness and prediction accuracy, which has definite practice significance and application value.

Keywords Near-infrared Spectroscopy; Particle swarm optimization; Least squares support vector machine; Extraction solution of safflower

(Received 16 August 2011; accepted 9 December 2011)