

波长选择算法在近红外光谱法中药有效成分测量中的应用

谷筱玉, 徐可欣, 汪^{*}

天津大学精密测试技术及仪器国家重点实验室, 天津 300072

摘 要 建立基于仪器分析方法的质量标准是中药进入国际市场的必要条件。近红外光谱技术以其能够反映样品的多种信息、易于在线应用的优势, 应用于中成药生产的在线质量监控, 可以提高中成药的质量控制标准, 加快中药现代化的进程。但在近红外光谱检测中存在着各成分谱图重叠严重, 光谱信息冗余, 特征吸收区域不明显的问题, 需要对采集到的波长进行优选, 以达到提高模型预测精度和简化模型的目的。从近红外光谱方法测量中药有效成分的基础研究入手, 以冰片含量的检测为例, 尝试采用遗传算法与模拟退火算法结合的模拟退火遗传算法及物理意义相对明确的多链逐步选择法对校正模型的波长进行优选。结果表明, 波长选择的方法可以使模型采用的波长数减少的同时提高预测精度, 波长选择最多可将波长数减少 84%, 预测精度提高 47.6%。

主题词 近红外光谱; 波长选择; 模拟退火遗传算法; 多链逐步选择法

中图分类号: O657.3, R927.2 **文献标识码:** A **文章编号:** 1000-0593(2006)09-1618-03

引 言

中药现代化的核心问题之一是进一步加强中成药的质量控制。建立基于仪器分析方法的质量标准是中药进入国际市场的必要条件。近红外光谱^[1]能够反映样品的多种信息, 易于在线应用。若将近红外光谱技术应用于中成药生产的在线质量监控, 可以提高中成药的质量控制标准, 加快中药现代化的进程。但在近红外光谱检测中存在着各成分谱图重叠严重, 光谱信息冗余, 特征吸收区域不明显的问题, 需要对采集到的波长进行优选, 以达到提高模型预测精度和简化模型的目的。本文从近红外光谱方法测量中药有效成分的基础研究入手, 以冰片含量的检测为例, 尝试采用不同的波长选择算法对校正模型的波长进行优化选择, 提高中药有效成分的测量精度, 并探讨不同方法间的优劣。

冰片作为一种化学合成类的中药原药材, 具有促进血脑屏障(BBB)通透性增加和引药上行的作用^[2]。在许多治疗心脑血管疾病的中成药中, 冰片常作为“药引”以增加其它药物的治疗效果^[3]。比如在速效救心丸中, 冰片作为其主要成分之一, 可以促进川芎嗪和方中其它成分的快速吸收, 提高其血药浓度和生物利用度, 还可抗心肌缺血, 使冠状窦血流回升, 减慢心率, 降低心肌耗氧量。

1 波长选择的物理意义

采用全谱建立多变量校正模型时, 不仅计算工作量大, 而且校正模型的预测精度未必能达到最优值。其中建模波长的不同将直接影响模型的测量精度。波长选择的物理意义在于: (1) 在有些光谱区域样品的有效光谱差异很小, 即光谱信息很弱, 不能提供样品结构的信息; (2) 由于近红外区域光谱重叠现象严重, 使得有些波长处的光谱信息的测量选择性和灵敏度较差, 直接影响模型的预测精度; (3) 由于近红外光谱信息存在多重相关性, 即不同波长下的光谱存在着线性相关的现象, 所以光谱信息中存在冗余信息; (4) 由于光谱仪器噪声的影响, 一些波段的样品光谱信噪比(SNR)较低, 光谱质量较差。

2 波长选择方法

2.1 模拟退火遗传算法

遗传算法^[4]是模拟达尔文的遗传选择和自然淘汰的生物进化过程的计算模型, 是一个以适应度函数为依据, 通过对群体中个体施加遗传操作来实现群体内个体结构重组的迭代优化过程。遗传算法在个体差异较大的运行早期, 容易使个别好的个体的后代充斥整个种群, 造成早熟; 在遗传算法后

收稿日期: 2005-06-16, 修订日期: 2005-09-28

基金项目: 国家“十五”科技攻关(2004BA706B12)和天津市自然科学基金(023800411)资助

作者简介: 谷筱玉, 1977年生, 天津大学精仪学院博士研究生^{*} 通讯联系人

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

期, 适应度趋向一致, 优秀的个体在产生后代时, 优势不明显, 从而使整个种群进化停滞不前。因此引进模拟退火算法^[5, 6], 对温度低时(遗传算法后期)的适应度函数进行适当地拉伸。在温度高时(遗传算法前期), 适应度相近的个体产生的后代概率相近; 而当温度不断下降后拉伸作用加强, 使适应度相近的个体适应度差异放大, 从而使得优秀的个体优势更明显。评价个体优劣的适应度函数为

$$F = \frac{1}{1 + \text{RMSEP}}$$

(RMSEP: root mean square error of prediction), 是由对模型的评价指标进行变换得到的。该算法不仅能够发挥遗传算法的优势, 把握正确的搜索方向, 而且可以发挥模拟退火算法快速收敛的优势。同时, 模拟退火算法部分的设计还能够使搜索避免过早地收敛于局部最优状态。

2.2 多链逐步选择法

多链逐步选择算法^[7]中定义信噪比为

$$\text{SNR} = \left| \frac{\beta_j}{\sigma_j} \right|$$

其中 β_j 表示第 j 个波长处的回归系数计算值, σ_j 表示在波长 j 处的预测标准差。波长变量的选择依据迭代的 SNR 从高到低, 并在每一次迭代中采用交互验证的方法去寻找预测误差最小的波长点。首先计算每个波长下的 SNR 值, 并按照 SNR 的降序来排列波长点, 并将其称为一条链。在这条链下采用交互验证的方法得到一条预测光谱。继续使用 SNR 的降序来排列残差光谱的波长点, 由此得到第二条链。链的选择一般是根据光谱的特征吸收的多少来确定。N 条链可以形成一个 $n \times m$ 的矩阵(m 代表波长数)。按列对矩阵进行搜索, 并剔除相同的点, 将得到的新向量以逐步累加进入的方式进行交互验证, 验证的指标为 RMSECV (root mean square error of cross validation)

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n_c} (\hat{c}_i - c_i)^2}{n_c}}$$

如果加入的波长点不能使 RMSECV 值有显著的降低, 则剔除该波长点。

波长优选可以从已知的光谱中提取最有效的谱图信息, 使校正模型具有最佳的预测能力, 并简化数据运算。

3 实验部分

3.1 仪器设备与参数设置

仪器设备: 采用美国 Perkin-Elmer 公司 Spectrum GX 傅里叶变换红外光谱仪, 液态氮冷却的 InSb 检测器, 光程长为 3 mm 的石英样品池, 蠕动泵自动进样系统。

参数设置: 光谱采集范围 4 000~ 10 000 cm^{-1} , 扫描分辨率 4 cm^{-1} , 16 次扫描累加。

3.2 样品来源和制备

样品来源: 冰片由天津中新药业集团股份有限公司技术中心提供。

样品制备: 用无水乙醇分别将冰片配制为实验设计的不同浓度的溶液。实验所用的试剂均为分析纯。

4 结果与讨论

4.1 光谱测量

采用 3 mm 的石英样品池测量冰片无水乙醇溶液的近红外光谱。实验中有 31 个样品, 1.0~ 25 $\text{mg} \cdot \text{mL}^{-1}$, 浓度间隔为 0.8 $\text{mg} \cdot \text{mL}^{-1}$ 。

4.2 波长选择算法的应用

采用交互验证的方法对由偏最小二乘法得到的校正模型进行评价。评价指标为预测标准偏差 RMSEP, 其值越小, 表明模型的预测能力越好。波长优选的范围^[8]为 5 000~ 6 000 cm^{-1} 。由表 1 的 RMSEP 值可以看出, 采用从纯数学角度进行优化选择的模拟退火遗传算法提高了模型的预测能力, 但作为一种随机优化算法, 其中群体的选择、交叉、变异都带有很强的随机性, 所以波长的选择, 每一次不会完全相同, 但位置大体相同, 得到的 RMSEP 值也略有差异。根据各波长下的信噪比进行波长优选的多链逐步选择方法物理意义明确, 最后选择的 19 个波长基本上都是冰片特征吸收峰所在的波段^[8]。在简化模型的同时提高了模型的预测精度, 但模型的稳定性有待进一步的研究。

Table 1 Calibration and prediction results in different wave bands

算法	波长数	RMSEP/($\text{mg} \cdot \text{mL}^{-1}$)
无	126	0.191
Simulated annealing-genetic algorithm	37	0.106
Multi-chain stepwise	20	0.100

在实际的工业在线生产中, 要求模型在满足预测精度的前提下, 尽可能地简化。实验结果表明, 波长选择的方法可以使模型采用的波长数在减少的同时提高预测精度。在我们的实验中, 通过波长选择的方法, 最多可将波长数减少 84%, 预测精度提高 47.6%。因此, 波长选择的方法可以有效地应用于中成药有效成分的检测和中成药质量控制中。

参 考 文 献

- [1] CHEN Bin, ZHAO Long-lian, LI Jun-hui, et al(陈 斌, 赵龙莲, 李军会, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2002, 22(6): 976.
- [2] WU Shou-rong, CHENG Gang, FENG Yan(吴寿荣, 程 刚, 冯 岩). Chinese Traditional and Herbal Drugs(中草药), 2001, 32(12): 143.
- [3] ZHAO Bao-sheng, LIU Qi-de(赵保胜, 刘启德). Chinese Traditional New Drugs and Clinic Pharmacology(中药新药与临床药理), 2002, 13: 287.
- [4] XU Lu(许 禄). Chemometrics: Principles and Applications of the Important Methods(化学计量学: 一些重要方法的原理及应用). Beijing: Science Press(北京: 科学出版社), 2004.
- [5] SONG Xiang-zhen, ZHANG Zun-jian, XU Jian-ping(宋祥珍, 张尊建, 徐建平). Journal of China Pharmaceutical University(中国药科大学学报), 1997, 28(5): 294.
- [6] WANG Xue-mei, WANG Yi-he(王雪梅, 王义和). Chinese J. Computer(计算机学报), 1997, 20(4): 381.
- [7] Michael J McShane, Brent D Cameron, Gerard L Cote, et al. Analytica Chimica Acta, 1999, 388: 251.
- [8] GU Xiao-yu, WANG Yan, XU Ke-xin, et al(谷筱玉, 汪 彦, 徐可欣, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(2): 155.

Application of Wavelength Selection Algorithm to Measure the Effective Component of Chinese Medicine Based on Near-Infrared Spectroscopy

GU Xiao-yu, XU Ke-xin, WANG Yan*

State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China

Abstract Near infrared (NIR) spectroscopy has raised a lot of interest in the pharmaceutical industry because it is a rapid and cost-effective analytical type of spectroscopy with no need for extensive sample preparation, and with the easy-realizable ability of on-line application. The NIR technology can increase the quality control standard of the Chinese medicine and accelerate the entry into the international market. In the present paper, two methods for wavelength selection are applied to the measurement of borneol, one of which is the multiple-chain stepwise, which tends to select many variables in the same area containing valuable information, and the other is the mixture genetic algorithm, which incorporates simulated annealing so as to improve the local searching ability while maintaining the global searching ability. The results present that the number of wavelength is reduced to 16% compared with the original number of wavelength, and the prediction accuracy has increased 47.6%. Therefore, the method of wavelength selection is a good way to enhance the prediction accuracy and simplify the model in NIR region.

Keywords Near infrared spectra; Wavelength selection; Simulated annealing-genetic algorithm; Multiple chain stepwise

(Received Jun. 16, 2005; accepted Sep. 28, 2005)

* Corresponding author