基于迭代初始化遗传算法的光谱波段选择及其在感冒液多组分测定中的应用

成 飙, 吴晓华, 陈德钊

浙江大学化学工程与生物工程学系, 浙江 杭州 310027

摘 要 光谱数据用于多元校正时, 组分间的交互作用会使部分波段与组分浓度呈非线性关系, 在用偏最小二乘法(PLSR)建模前, 宜作波长筛选。基于迭代初始化的遗传算法(IRGA)将运行多轮GA, 递归地以上轮结果作为先验知识支持下轮的初始化, 并对入选波长点的统计频率进行平滑处理, 由此可高效地从全谱中选出校正性能良好的波段, 筛选结果稳定。入选波段对全谱既作了适当简化, 又充分保留了有效信息。再采用PLSR建模, 模型更具稳健性。将该法用于感冒液的五组分测定, 与全谱建模法相比, 其预测性能和稳健性有显著提高。

主题词 遗传算法; 波长选择; 偏最小二乘; 多组分
中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2006)10-1923-05

引 言

现代光谱分析技术的发展极大地提高了仪器的分辨率和灵敏性, 可为分析者提供精确的光谱数据。光谱数据通常遵从朗伯-比尔定律, 可用偏最小二乘回归(PLSR)建模, 并作多元校正。近年研究发现, 在多组分体系中, 由于组分间的交互作用产生了新分子键等原因, 全谱中会出现不符合朗伯-比尔定律而与组分浓度呈非线性关系的波段。理论与试验均表明, 在PLSR建模前先筛选波长, 消除与待分析组分无关或呈非线性关系的波长点, 将可简化模型, 并提高其预测精度和稳健性[1, 2]。

波长选择方法有逐步回归法、无关信息变量消除法(uninformation variable elimination, UVS)[3]、迭代变量选择法(iterative variable selection, IVS)[4]、模拟退火法(simulated annealing, SA)[5]和遗传算法(genetic algorithm, GA)[6]等。遗传算法的研究和应用较为广泛, 并衍生出很多改进形式, 以迭代型遗传算法最为引人注目, 其选择效果良好。

速效感冒液包含扑热息痛、扑尔敏、咖啡因、愈创木酚甘油醚和对氨基酚五种成分, 组分间存在交互作用, 在紫外区其吸收峰相互重叠, 采用一般光度分析法难于定量每一组分[7]。本文拟用基于迭代种群初始化的遗传算法, 先筛选紫外区波长, 找出和组分浓度相关性最好的几个波长区间, 再进行PLSR建模, 以期提高感冒液多元校正精度。

1 遗传算法用于波长选择

1.1 基于遗传算法的波长点选择

遗传算法利用生物界物竞选择的进化机制, 是一种随机优化方法, 将其用于波长选择, 其主要步骤为[6, 8]。

(1) 染色体编码: 设被选的波长点共有 m 个, 则染色体长度为 m, 每个基因对应一个波长点, 以一个二进制位编码, 编码为1, 表示选中该波长; 为0, 则未选中。

(2) 种群初始化: 设初始群体包含 N 个个体, 随机产生 N 个染色体作为初始种群。可以限定每条染色体中基因编码为1的总数, 以免初始选入的波长点过多。

(3) 适应度函数: 常采用交互验证法评价模型的预测能力, 评价指标为交叉验证均方根偏差, 其值越小, 校正模型的预测性能越优。对各染色体选中的波长点数据, 用PLSR方法交叉地建模和预测, 计算值, 并经变换后作为该染色体的适应度函数。

(4) 遗传操作

选择: 以"轮盘赌"的方式进行正比选择。

交叉: 采用均匀交叉方式。交叉是产生新个体的主要方式, 故选用较大的交叉概率, 常为0.8左右。

变异: 采用简单翻转变异。为维持种群的多样性, 需引

收稿日期: 2005-05-28, 修订日期: 2005-08-28
基金项目: 浙江省重点科技项目(2004C21SA120002)资助
作者简介: 陈德钊, 1981年生, 浙江大学化学工程与生物工程学系博士研究生

入变异机制，但概率较小，常为 0.1 左右。

(5) 算法终止准则：算法重复执行各遗传操作，种群不断进化，且进化速度渐趋平缓，常设置进化代数的上限，至此算法终止。

(6) 波长选择：选用具有最大适应度值的染色体对应的波长点。

1.2 基于遗传算法的波长区段选择

1.1 节的 GA 为用于波长选择的早期形式，称为随机初始化 GA(randomly initialized GA，RIGA)，RIGA 和引言中提及的其它算法往往只选出散落于全波段的，各不相邻的若干个原始波长点，与它们相邻的很多波长点因显著相关而被摒弃。光谱测试中难免出现瞬时扰动，它们对个别点的影响往往更甚于区段。基于较少分离的波长点建模，将使模型预测的抗干扰能力严重受损。由此文献[9]等提出应筛选出一些波段，它们将有效地抵御偶然的瞬时扰动，并充分地保留有效的光谱信息，以利于发挥后续 PLSR 的优越性，提高模型的稳健性。用于波段选择的 GA 常用以下两种策略：(1) 将全波段等长地划分为若干区段，染色体的每个基因对应于一个区段，被选中用于建模的将是波段，而不是波长点。(2) 多次运行 GA，统计各波长点的入选频率，再对各点的入选频率作滑动平均，并以平滑后的入选频率作为选择波长的依据。显然，此策略的适应性更好，它为迭代型遗传算法的出发点。

2 迭代初始化的遗传算法

RIGA 的筛选受初始种群影响甚大，各次运行结果往往有显著差异，颇不稳定。为此可多次独立运行 RIGA，统计各波长点的入选频率并作平滑，然后按平滑频率以降序逐个引入波长点作为自变量，进行 PLSR 建模，计算评价函数(如 RMSECV)，并以其最优为目标，确定入选波长点的个数。常随 P 的增大先剧烈下降，再逐步上升。

初始种群的分布对波长点的选择影响甚大，若以先验知识支持初化，将有助于 GA 的筛选。由于难免存在背景和其他组分的干扰，不宜选用纯组分的光谱响应作为先验依据。其实每轮 GA 的运行结果中往往包含非常有用的信息，以此作为先验知识支持下轮 GA 的初化，将使 GA 可在信息相对丰富的区域中细化搜索，高效地实现波段筛选。这种多轮运行，递归地以上轮结果支持下轮初化的遗传算法称为迭代初始化遗传算法(iteratively reinitialized GA，IRGA)[9-11]，其流程如图 1 示，它区别于多次独立运行 RIGA 的方法，还在于采用了更恰当的模型性能评价函数。

2.1 IRGA 的初始化方法

RIGA 以相同的概率将各基因初化为 1，IRGA 对第个基因初始化为 1 的概率如(1)式示[9]。

$$p_{wi} = \frac{(R - r) \cdot p_0 + r \cdot p_{si}}{R} \qquad (1)$$

其中 R 为设定的运行 GA 的总轮数，r 为当前轮数。它是两种概率的加权平均，其一为原始概率 $p_0 = n/m$，其二为平滑概率 $p_{si}(p_{i-1} + p_i + p_{i+1})/3$，记 m 为基因总数，n 为预定初始基因为 1 的个数，p_i 为第 i 个基因的经验概率 $p_i = sel_i / \sum_{j=1}^{m} sel_j$，其中 sel_j 为第 j 个波长点在前几轮 GA 中的入选次数，p_i 反映了波长点 i 入选的先验知识，经平滑为 p_{si}，可使初化时入选的波长点更具连续性。可见 p_{wi} 折中了原始与经验概率，其加权系数开始时侧重于原始概率，随着 GA 轮次的增大，逐渐侧重于平滑概率，使算法渐趋稳定。

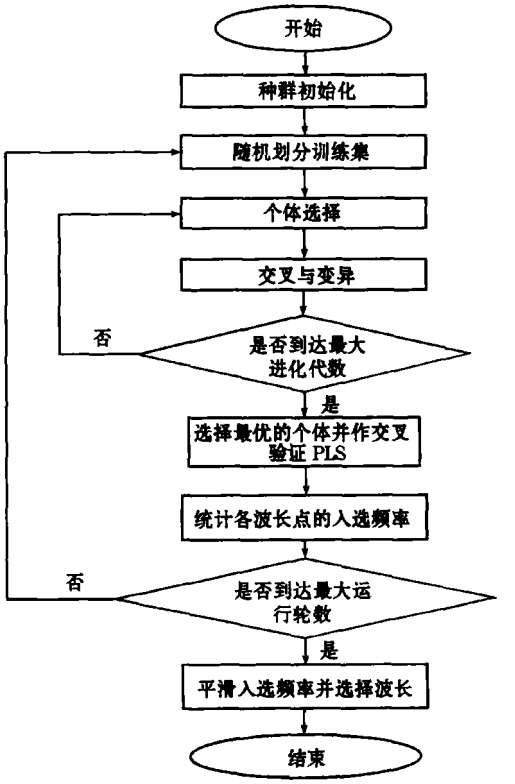

Fig 1 Flow sheet of IRGA

2.2 模型性能评价函数

GA 运行时以作为模型性能评价函数，需将训练样本集划分为校正集和监视集，以监视集的预测偏差作为适应度函数。若固定监视集，将导致对其过度训练，造成过拟合而引入噪声。为此，本文设计的 IRGA 在 GA 的每轮运行前，都将训练样本集重新随机地划分为若干组，并交替地取为校正集和监视集，使每轮的适应度函数独立，以求最大程度地避免过拟合。为了提高速率，在每轮 GA 中，将以固定的 PLS 成分数计算 RMSECV，通常在每轮运行结束时，对最佳个体作交叉验证，选定最优 PLS 成分数，作为下轮计算评价函数时所提取的 PLS 成分数。

3 试验与数据

3.1 仪器和试剂

仪器：岛津 260 型紫外分光光度计。

试剂：扑热息痛、咖啡因、扑尔敏、对氨基酚、愈创木酚甘油醚标准溶液(20 μg·mL^{-1})；盐酸溶液(0.1 mol·L^{-1})，作溶剂用。所用试剂均为分析纯。

3.2　混合标液的配制与测定

由于样品组分数较多，浓度变化范围较大，采用正交设计法设定配方，共配制 81 组混合标准溶液用于建模，称为训练集，另配制 9 组供独立检验，称为验证集。溶液测定时，在紫外分光光度计上以溶剂作参比，在 206~ 320 nm 范围内，每隔 2 nm 测一次吸光度值，共采集 58 个波长点的吸光度值，标样的吸收曲线如图 2 上半部分所示。

3.3　建模方法

基于训练集采用两种方法建模，其一采用全谱，其二用 IRGA 选择波长区段，然后都再用 PLSR 方法建模，分别记为 FULL- PLS 和 IRGA- PLS，所提取的 PLS 成分数标记为# PLSs。IRGA 的运行参数设定为：种群规模 30，最大入选波长点数 20，交叉概率 0.8，变异概率 0.1，运行轮数(Cycle) 100，每轮进化代数(Generation) 50。另外还将采用两种数据预处理方法：中心化(Centering) 和标准化(Autoscaling)；应用两种方式划分训练集，进行交互验证，计算 RMSECV，分析讨论它们对 IRGA 选择波长点和建模性能的影响。

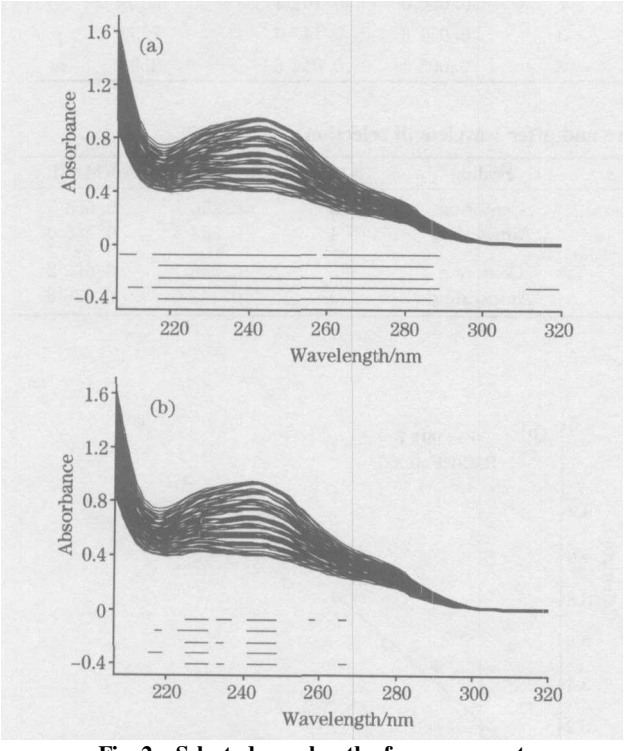

Fig 2　Selected wavelengths for component 2 of 5 IRGA runs

4　结果与讨论

4.1　模型性能评价标准

本文对模型性能采用两个评价标准，其一为 RMSECV，它针对训练集，用为 GA 的适应度值；其二为 RMSEP，它针对未参与建模的独立验证集，如(2) 式所示，它更确切地反映模型的预测性能。

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y_{\mathrm{pred}} - y_{\mathrm{ref}})}{f}} \qquad (2)$$

y_{pred} 为预测值，y_{ref} 为实测值，f 为预测样本数。两种方法所建模型为 5 种组分的检测结果列于表 1。用两种预处理方法，为组分 2(Guaifenesin) 建模的结果列于表 2，其中还有 IRGA 所选择的波长点。

4.2　预处理方法的影响

由表 2 的数据可见，预处理方法将显著影响 IRGA 对波长点的选择，并影响建模性能。标准化预处理，将使 IRGA 选出更多连续的波长点，所建模型的性能也都明显优于中心化的结果。表 1 所列的数据，为采用标准化预处理后的结果。

4.3　训练集划分方式的影响

在用 IRGA 选择波长点时，以两种方式划分训练集。一是每轮对训练集的划分不变，记为Ⅰ方式，另一方式则在每轮运行前，均重新随机地划分训练集，记为Ⅱ方式。以两种方式，针对组分 2 分别执行 IRGA 各 5 次，并在图 2 的下部，以水平线标识各次选出的波长点位置。显然，IRGA 选出的波长点都连续相邻，组成为波段，各次选出的主要波段基本相同，较为稳定。而Ⅰ方式选出的波长点总数远大于Ⅱ方式的选出量，此为Ⅰ方式在同一交互验证集上训练，导致过拟合，入选了过多的波长点。图 3 为两种方式下随# PLSs 的变化曲线，可见Ⅱ方式以较少的# PLSs，达到较低的值，在下文还可见Ⅱ方式下的值也较低，由此所建的模型性能更优。表 1 所列的数据，采用了Ⅱ方式。

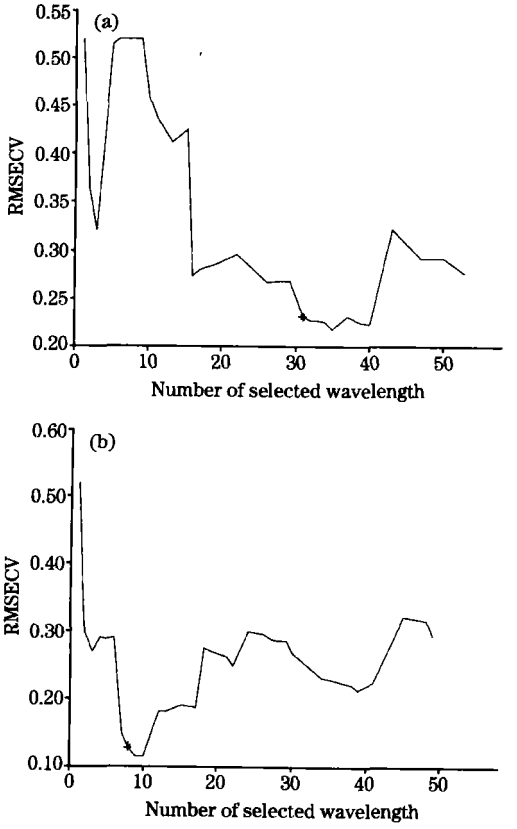

Fig 3　RMSECV versus number of selected variables for component 2

4.4 波长选择及模型性能比较

对各组分各独立运行 5 次 IRGA，所选出的波长点重现性好。图 4 显示了对 5 种组分所选出的复现性最好的波长点，均以水平线表示，数字对应于组分号。

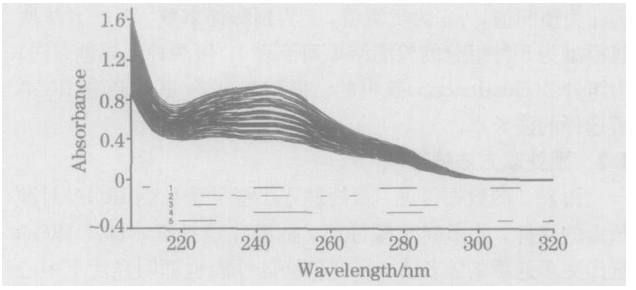

Fig 4 Selected wavelengths for 5 components respectively

由表 1 可见，对五种组分所建的 IRGA-PLS 模型，与 FULL-PLS 模型相比，其和值均有下降，尤其前 3 个组分的值下降明显，分别降低 44.94%，61.69% 和 57.74%，表示其预测性能改善明显。对于后 2 种组分，仅降低 7.52% 和 3.95%，是因为 2 种组分浓度较低，所含的光谱信息较弱，在全谱基础上的改进不够明显。

图 5 显示了验证集的 9 个溶液，其组分 2 的实际配制值与 2 种模型预测值之间的偏离情况，可直观地看出，经波长选择后，IRGA-PLS 模型的预测值对实际值的偏离较小。

Table 1 Calibration and prediction results by two models

Components	FULL-PLS			IRGA-PLS			RMSEP Improvement/%
	# PLSs	RMSECV	RMSEP	# PLSs	RMSECV	RMSEP	
Acetaminophen	5	0.125 4	0.299 3	3	0.111 0	0.164 8	44.94
Guaifenesin	4	0.283 3	0.586 0	3	0.135 7	0.227 4	61.19
Caffeine	4	0.1 504	0.242 3	7	0.056 0	0.102 4	57.74
Chlorphenaine meteate	6	0.051 7	0.1 596	3	0.051 3	0.147 6	7.52
p-aninophenol	3	0.005 2	0.015 2	3	0.005 1	0.014 6	3.95

Table 2 Results for Guaifenesin before and after wavelength selection

Model	Wavelength points	Scaling	# PLSs	EMSECV	RMSEP
FULL-PLS	Full(206~320)	Centering	3	0.296 1	0.658 5
		Autoscaling	4	0.283 3	0.586 0
IRGA-PLS	226, 228, 230, 234, 242, 244, 246, 248	Centering	2	0.288 1	0.690 2
	226, 228, 230, 234, 236, 242, 244, 246, 248, 258, 266	Autoscaling	3	0.132 6	0.225 8

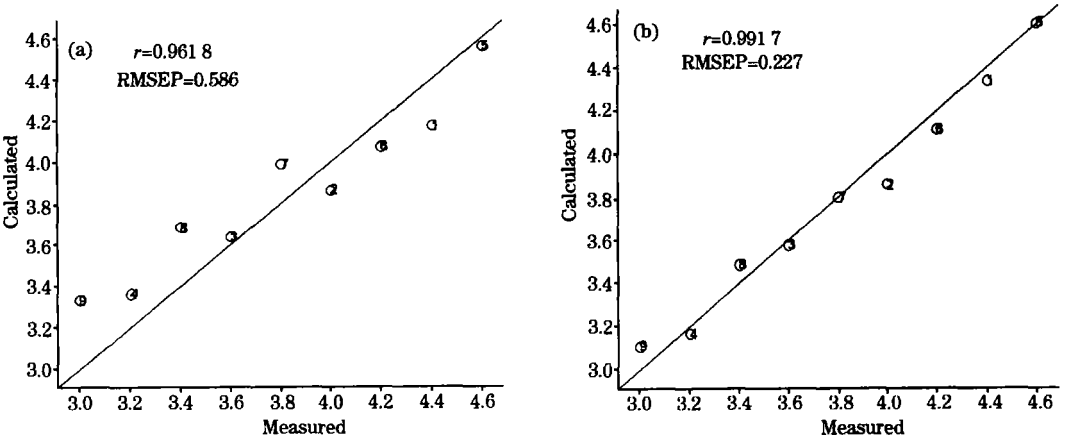

Fig 5 Correlation plots for guaifenesin in test set

(a): Before selection; (b): After selection

5 结 论

迭代初始化遗传算法 IRGA 是一种波段选择的有效方法，它多轮运行 GA，递归地以上轮结果作为先验知识支持下轮的初始化，并对入选波长点的统计频率进行平滑处理，由此可高效地从全谱数据中选出连续相邻的校正能力强的波长点，筛选结果稳定。入选波段对全谱既作了适当简化，又充分保留了有效信息。再采用 PLSR 建模，模型更具稳健性。将该法用于感冒液的紫外-可见光谱的多元校正，其预测性能和稳健性明显优于全谱建模法。

参 考 文 献

[1] Leardi R, Lupianez Gonzalez. Chemometrics Intell. Lab. Syst., 1998, 41: 195.
[2] Spiegelman C H, Mcshane M J, et al. Anal. Chem., 1998, 70: 35.
[3] Centner V, Massart D L, de Noord O E, et al. Anal. Chem., 1996, 68: 3851.
[4] Lindgren F, Geladi P, Rannar S, et al. J. Chemometrics, 1994, &: 349.
[5] John H Calivas, Nancy Roberts, Jon M Sutter. Anal. Chem., 1989, 61: 2024.
[6] Julia Acros M, Cruz Ortiz M. Analytica Chimica Acta, 1997, 339: 63.
[7] ZHANG Li-qing, WU Xiao-hua, TANG Xi, et al(张立庆, 吴晓华, 唐 曦, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2002, 22(3): 427.
[8] Christoffer Abrahamsson, Jonas Johansson. Chemometrics Intell. Lab. Syst., 2003, 69: 3.
[9] Riccardo Leardi. J. Chemometrics., 2000, 14: 643.
[10] Hector C Goicoechea. J. Chem. Inf. Comput. Sci., 2002, 42: 1146.
[11] Hector C Goicoechea. J. Chemometrics, 2003, 17: 338.

Wavelength Interval Selection by Iteratively Reinitialized GA and Its Application to Spectrophotometric Determination of Components in Cough Syrup

CHENG Biao, WU Xiao-hua, CHEN De-zhao

Department of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China

Abstract Wavelength selection in PLS calibration can be used to reach two goals: improve the predictive ability and simplify the model. Iteratively reinitialized GA is a modified genetic algorithm, and it gives an initializing procedure of selecting the first candidates for every run of GA, which uses the results of previous runs as the guiding information. This algorithm can select wavelength regions instead of scattering points, which is very helpful in understanding the relevant parts of spectra. Furthermore, the continuous wavelength points make the PLS model more robust. Appling IRGA based wavelength selection to the UV-Vis spectrum of cough syrup, the result illustrates that PLS regression can greatly benefit from variable selection when used for multicomponent spectrophotometric determination.

Keywords Genetic Algorithm; Wavelength selection; Partial least square; Multicomponent

(Received May 28, 2005; accepted Aug. 28, 2005)