遗传算法用于偏最小二乘方法建模中的变量筛选

褚小立* 袁洪福 王艳斌 陆婉珍
(石油化工科学研究院,北京 100083)

摘 要 利用全局搜索方法——遗传算法(genetic algorithms, GA)对近红外光谱快速分析中的波长变量进行筛选,再用偏最小二乘方法(partial least squares, PLS)建立分析校正模型。对两类样品的近红外光谱分析应用实例表明,这种选取变量进行校正的方法,不仅简化、优化了模型,而且增强了所建模型的预测能力,尤其适用于单纯 PLS 较难校正关联的体系。

关键词 遗传算法,偏最小二乘方法,变量筛选,汽油,芳烃,润滑油,饱和烃

1 引 言

偏最小二乘方法(partial least squares, PLS)是目前多元校正中最常用的方法之一,在光谱结合 PLS 方法建模中,传统观点认为 PLS 具有较强的抗干扰能力,可全波长或根据相关性选取某波段参与建模[1]。随着对 PLS 方法的深入研究和应用,通过特定方法筛选特征变量有可能得到更好的定量校正模型[2]。筛选特征变量一方面可以简化模型,更主要的是由于不相关或非线性变量的剔除,可以得到预测能力更强的校正模型。

目前,变量筛选的方法主要有逐步回归方法[3]、模拟退火算法(simulated annealing algorithm, SAA)[4]、遗传算法(genetic algorithms, GA)[5]和多链方法(mutiple-chain method, MCM)[6]等,其中遗传算法的研究和应用较为广泛。遗传算法是借鉴生物界自然选择和遗传机制,利用选择、交换和突变等算子的操作,随着不断的遗传迭代,使目标函数值较优的变量被保留,较差的变量被淘汰,最终达到最优结果[7]。遗传算法自 70 年代提出以来,在国内已有许多领域得到应用[8~11],在特征变量筛选方面也获得了较好的结果[12],但尚未见到遗传算法用于光谱多元校正中波长变量的选取。本文将遗传算法用于近红外光谱快速分析中的波长筛选,筛选后的波长变量再由 PLS 方法建立分析校正模型。应用实例表明,这不仅优化了模型,而且增强了所建模型的预测能力。

2 遗传算法实现过程

遗传算法的实现主要包括如下几个基本要素,具体的遗传算法实现流程框图参见图 1。

(1)参数编码

由于遗传算法不便直接处理空间数据,需通过编码将它们表示成遗传空间的基因型串结构数据,一般采用基于 0/1 字符的二进制串形式。对于包含 n 个参数(如波长)的问题,可用一串含有 $n \times m$ 个字符(对应于基因)的向量(对应于染色体)表示,m 表示每个参数需要的基因位数。本文 m 选取 1,即一条染色体中的每个基因对应一个实际参数,若基因为 1 表示其代表的参数被选中,基因为 0 则未被选中。

(2)群体的初始化

随机或根据一定的限制条件产生一个给定大小的初始群体,群体的大小即个体(染色体)的数目根据参数(基因)的多少选定,一般选 30 ~ 100,本文选取 70,限制条件是个体选定的最大变量数目。

(3)适应度函数的设计

遗传算法根据适应度函数来评价个体的优劣,作为以后遗传操作的依据。由于在整个搜索进化过程中,只有适应度函数与所解决的具体问题相联系,因此,适应度函数的确定至关重要。本文采用 PLS 交互验证中因变量的预测值和实际值的相关系数(r)为适应度函数。具体实施方法为:对每个个体所选

2000-05-29 收稿;2000-11-24 接受。

的变量进行数据重新组合,再用 PLS 交互验证得到相关系数(r)。

(4)遗传操作设计

选择:选择的目的是把优胜的个体直接遗传到下一代或通过交叉或变异产生新的个体再遗传到下一代。选择操作是建立在群体中个体的适应度评估基础上的。本文采用最常用的选择方法——适应度比例方法,也称转轮法,每个个体的选择概率与其适应度成比例。

交叉:交叉是遗传算法中最主要的算子,寻优的搜索过程主要是通过它来实现的。本文采用随机一点交叉方法,交叉概率为 0.8。

变异:引入变异算子的目的是维持群体的多样性,防止出现未成熟收敛现象,此外是使遗传算法具有局部的随机搜索能力。本文采用基本变异算子,即在个体中随机挑选一个或多个基因以变异概率做变动,变异概率为 0.1。

(5)收敛判据

常规的数学规划方法在数学上都有比较严格的收敛判据,但遗传算法的收敛判据基本是启发式的。因此,遗传算法的判据较多,如计算时间、计算机变量或从解的质量方面等确定判据。本文以遗传迭代次数为收敛终止条件。

(6)变量选取

本文采用的方法为:在遗传迭代终止后,所有变量按选取频率重新排列,再由选取变量数与相关系数(r)作图选定最佳变量数,便得到所选的变量。

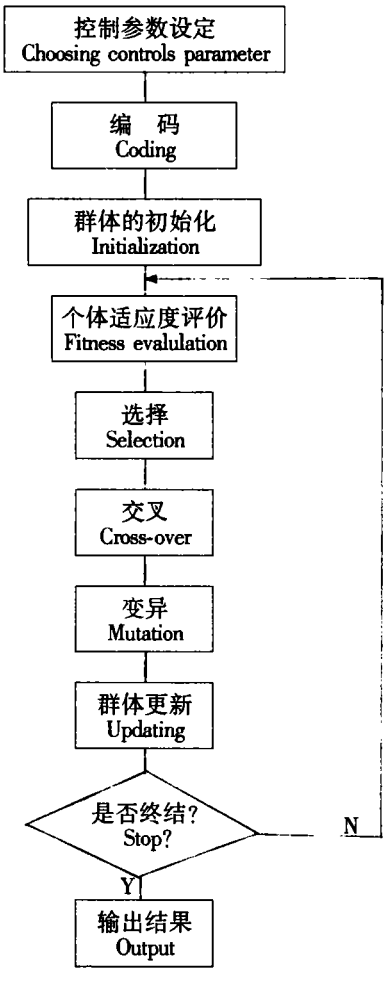

图 1 遗传算法实现流程框图

Fig.1 Flow diagram of the genetic algorithm

3 实验部分

3.1 仪器

NIR-2000近红外光谱仪(石油化工科学研究院研制、英贤仪器实业有限公司生产),5 cm玻璃样品池,CCD检测器,光谱范围700~1100 nm,取点间隔0.2 nm。

3.2 样品与基础数据来源

54个重整工艺汽油由石油化工科学研究院催化重整中型装置提供,重整汽油中芳烃组成包括苯、甲苯、二甲苯和重芳烃(C8以上芳烃),芳烃组成各含量均由气相色谱测得。

42个润滑油基础油来自国内不同单位和工艺装置,饱和烃含量由柱色谱法测得。

3.3 光谱采集

以空气为参比,样品放入3 min后开始扫描。CCD扫描累加次数为50。环境温度:22±5℃。

3.4 数据处理

偏最小二乘方法和遗传算法程序及光谱预处理程序(如微分)均采用MATLAB语言编制。

4 结果与讨论

4.1 光谱数据预处理

为消除颜色等因素对校正结果的影响,通常需对光谱数据进行波段选择和微分等基线预处理。表1列出了不同波段和微分处理对重整汽油重芳烃校正结果的影响,其中830~980 nm波段经11点一阶微分的校正结果最好,此模型命名为Model-all。综合以上结果,在用遗传算法筛选变量前,首先选取830~980 nm波段的光谱数据,再进行11点一阶微分处理。用于PLS建模的校正集由43个重整汽油样品组成,剩余的11个样品组成验证集,最佳主因子数采用交互验证法所得的预测残差平方和(PRESS)确定。

表1　不同波段和微分处理对重芳烃的校正结果

Table 1　Calibration results of different spectral region and derivative pretreatment

波段范围(nm) Spectral region	基线处理方法 Baseline pretreatment method	主因子 Factor	PRESS	R
700~1100	无 None	5	40.93	0.7636
	9点一阶微分 9 point 1st derivative	4	43.04	0.7320
	11点一阶微分 11 point 1st derivative	4	40.67	0.7642
	13点一阶微分 13 point 1st derivative	4	42.59	0.7435
830~980	无 None	4	41.08	0.7617
	9点一阶微分 9 point 1st derivative	3	24.11	0.8245
	11点一阶微分 11 point 1st derivative	3	19.54	0.8950
	13点一阶微分 13 point 1st derivative	3	21.48	0.8743
880~930	无 None	4	42.11	0.7385
	9点一阶微分 9 point 1st derivative	3	24.36	0.8186
	11点一阶微分 11 point 1st derivative	3	21.01	0.8840
	13点一阶微分 13 point 1st derivative	3	22.72	0.8625

PRESS: prediction residual sum of squares

4.2 变量筛选结果

遗传算法控制参数设定:初始群体70,最大选取变量数200,交叉概率0.8,变异概率0.1,遗传迭代次数100。

将预处理后的光谱数据用遗传算法进行变量筛选,图2为经过100次迭代后变量选取的频率图,在875 nm、913 nm和934 nm区域变量选取频率最大,这正对应于芳烃、甲基和亚甲基中C-H键三倍频吸收,说明遗传算法所选变量的合理性。图3为所有变量按选取频率重新排列后,相关系数(r)随选取变量数的变化趋势图,由图可以看出最佳的变量为26,从而得到所选变量。将所选变量组成新的数据矩阵,再用PLS重新建立测定重芳烃含量的校正模型。

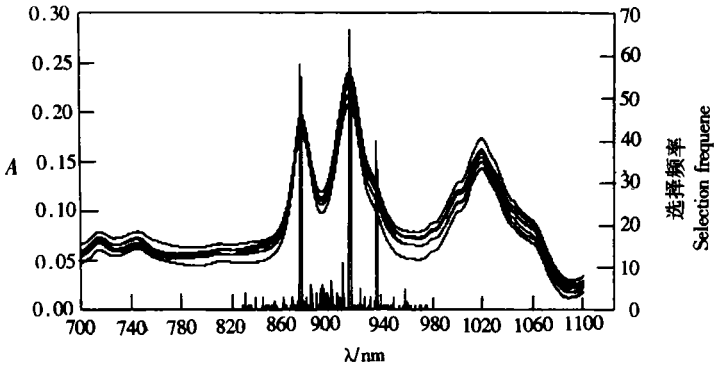

图2 近红外光谱与GA选取变量频率对照图

Fig.2 The contrast of near infrared spectra and variable selection frequency

由于遗传算法的初始群体是随机选取的，选择、交叉和变异也带有较强的随机性，为验证遗传算法的随机性对PLS建模结果的影响，本文连续进行了4次重复的遗传迭代过程，表2和表3分别列出了PLS建模结果及用所建模型对验证集的测定结果。结果表明，遗传算法的随机性由于经过多次遗传迭代，对PLS建模和预测结果影响不大，但其结果均好于模型Model-all的结果。

同时用这种方法建立了测定重整汽油的苯、甲苯、二甲苯含量的校正模型，其中遗传算法对苯、甲苯、二甲苯所选取的变量数分别为17、26和39，遗传算法变量选取前后的校正和预测对比结

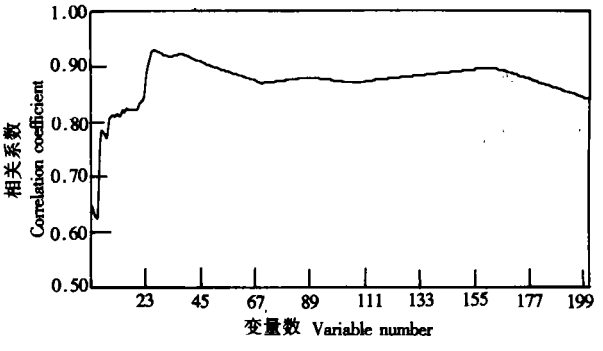

图3 相关系数 r 随选取变量数的变化趋势图

Fig.3 The trend of correlation coefficient with different variable number

表2 遗传算法变量选取后PLS校正结果

Table 2 Calibration results after variable selection by genetic algorithms (GA)

模型名称 Model name	次数 Time	选取变量数 Variable number selected	主因子 Factor	PRESS	R
Model-26	No.1	26	3	13.56	0.9280
Model-29	No.2	29	3	12.79	0.9323
Model-38	No.3	38	3	12.96	0.9313
Model-33	No.4	33	3	12.17	0.9357

表3 不同模型对验证集样品的预测结果

Table 3 Predictive results in test sets by different models

样品号 Sample	实测值 Actual	Model-all		Model-26		Model-29		Model-38		Model-33	
		PRE	DEV	PRE	DEV	PRE	DEV	PRE	DEV	PRE	DEV
CZ01	13.96	12.35	-1.61	12.64	-1.32	12.78	-1.18	12.82	-1.14	13.20	-0.76
CZ02	8.99	9.23	0.24	9.11	0.12	9.50	0.51	9.52	0.53	8.89	-0.10
CZ03	8.82	9.19	0.37	8.42	-0.40	8.84	0.02	8.59	-0.23	8.51	-0.32
CZ04	11.14	11.74	0.60	11.34	0.20	11.39	0.25	11.29	0.15	11.98	0.84
CZ05	11.18	10.87	-0.31	11.84	0.66	11.63	0.45	11.64	0.46	11.73	0.55
CZ06	12.98	12.50	-0.48	12.67	-0.31	12.68	-0.30	12.48	-0.50	12.80	-0.18
CZ07	12.35	12.04	-0.31	12.47	0.12	13.25	0.90	13.06	0.71	12.33	-0.02
CZ08	12.62	12.23	-0.39	12.83	0.21	12.54	-0.08	12.71	0.09	12.49	-0.13
CZ09	9.20	8.32	-0.88	8.78	-0.42	9.17	-0.03	9.35	0.15	8.70	-0.50
CZ10	9.85	9.46	-0.39	9.10	-0.75	10.42	0.57	10.36	0.51	9.73	-0.12
CZ11	13.09	11.77	-1.32	12.40	-0.69	12.92	-0.17	12.89	-0.20	13.04	-0.05
RMSEP			0.68		0.56		0.56		0.54		0.44

PRE.预测结果(predictive results)；DEV.偏差(deviation)；RMSEP. 验证集均方根偏差(the root mean square error in predication)

表 4　遗传算法选取变量前后校正和预测结果
Table 4　Calibration and predictive results before and after selecting variable by GA

	变量选取前 Before selecting variable		变量选取后 After selecting variable	
	RMSEC	RMSEP	RMSEC	RMSEP
苯 Benzene	0.44	0.35	0.37	0.28
甲苯 Toluene	0.52	0.48	0.42	0.36
二甲苯 Dimethyl benzene	0.74	0.62	0.58	0.49

RMSEC. 校正集均方根偏差(the root mean square error in calibration);RMSEP. 验证集均方根偏差(the root mean square error in predication)

果见表 4。结果表明,通过遗传算法进行变量选取,可优化校正模型,使其具有较强的预测能力。润滑油基础油相对于汽油较重,氢碳比小,因此在短波近红外光谱区间的信息相对较弱,尽管单纯用 PLS 方法建立饱和烃的校正模型可以满足常规分析方法的要求,但偏差较大。用于 PLS 建模的校正集由 35 个润滑油基础油样品组成,剩余的 7 个样品组成验证集,遗传算法选取波长变量前后的校正结果见表 5,验证集的测定结果见表 6。结果可以看出,对于信息较弱,PLS 较难关联的体系,通过遗传算法选取波长变量可较大幅度地提高校正和预测结果。

表 5　遗传算法选取波长变量前后的校正结果
Table 5　Calibration results before and after selecting variable by GA

变量选取前 Before selecting variable			变量选取后 After selecting variable			
主因子 Factor	PRESS	*R*	选取变量数 Variable number selected	主因子 Factor	PRESS	*R*
8	153.1	0.8537	51	6	117.2	0.9280

波段范围(spectral region):850 ~ 980 nm;基线处理方法(baseline pretreatment method):13 点一阶微分(13 point 1st derivative)

表 6　遗传算法选取波长变量前后的预测结果
Table 6　Prediction results before and after selecting variable by GA

样品号 Sample	实测值 Actual	变量选取前 Before selecting variable		变量选取后 After selecting variable	
		PRE	DEV	PRE	DEV
RH01	90.8	91.6	0.8	92.4	1.6
RH02	80.6	85.3	4.7	81.1	0.5
RH03	89.6	91.7	2.1	90.9	1.3
RH04	88.6	87.7	-0.9	88.4	-0.2
RH05	94.4	95.1	0.7	95.9	1.5
RH06	84.9	82.3	-2.6	83.8	-1.1
RH07	86.6	85.4	-1.2	86.7	0.1
RMSEP			2.3		1.1

PRE. 预测结果(predictive results);DEV. 偏差(deviation);RMSEP. 验证集均方根偏差(the root mean square error in predication)

5　结　论

(1)利用遗传算法对近红外光谱的波长变量进行筛选,再用 PLS 方法建立校正模型,不仅简化、优化了模型,而且增强了所建模型的预测能力。(2)遗传算法用于近红外光谱结合偏最小二乘方法建模中的波长变量筛选尤其适用于信息弱、单纯 PLS 较难关联校正的体系。(3)本文设计的遗传算法用于变量筛选是有效的,这种方法还可用于除近红外光谱以外的其他波谱数据的筛选。

致 谢　感谢意大利热那亚大学 Riccardo Leardi 教授在遗传算法上给予的帮助。

References

1　Thomas E V, Haaland M D. *Anal. Chem.*, **1990**, 62: 1091 ~ 1099
2　Thomas E V. *Anal. Chem.*, **1994**, 66: 759 ~ 803a
3　Zhu Eryi(朱尔一), Deng Zhiwei(邓志威), Huang Benli(黄本立). *Chem. J. Chinese Universities*(高等学校化学学报), **1993**,

14(11):1518 ~ 1521
4 Kalivas J H, Roberts N, Sutter M. *Anal. Chem.*, **1989**, 61:2024 ~ 2029
5 Riccardo Leardi. *Chemometr. Intell. Lab. Syst.*, **1998**, 41:195 ~ 207
6 Chen Guoliang(陈国良), Wang Xufa(王煦法), Zhuang Zhenquan(庄镇泉). *Genetic Algorithms and Applications*(遗传算法及其应用). Beijing(北京): People's Post Press(人民邮电出版社), **1996**: 1 ~ 14
7 Michael J M, Brent D C, Gerard L C. *Analytica Chimica Acta.*, **1999**, 388:251 ~ 264
8 Guo Ming(郭 明), Chen Qiande(陈前德), Liu Wenjie(刘文杰), Xu Lu(许 禄). *Chinese J. Anal. Chem.*(分析化学), **2000**, 28(1):6 ~ 11
9 Li Tonghua(李通化), Zhang Zhongjie(张众杰), Zhu Zhongliang(朱仲良). *Chem. J. Chinese Universities*(高等学校化学学报), **1995**, 16(3):354 ~ 359
10 Cai Wensheng(蔡文生), Shao Xueguang(邵学广), Zhao Guiwen(赵贵文), Zhang Maosen(张懋森). *Chinese J. Anal. Chem.*(分析化学), **1997**, 25(2):231 ~ 237
11 Liu Jia(刘 嘉), Deng Bo(邓 勃). *Chinese J. Anal. Chem.*(分析化学), **1997**, 25(7):784 ~ 788
12 Zhang Yuan(章 元), Zhu Eryi(朱尔一), Zhuang Shixia(庄峙厦), Wang Xiaoru(王小如). *Chem. J. Chinese Universities*(高等学校化学学报), **1999**, 20(9):1371 ~ 1375

Variable Selection for Partial Least Squares Modeling by Genetic Algorithms

Chu Xiaoli*, Yuan Hongfu, Wang Yanbin, Lu Wanzhen
(*Research Institute of Petroleum Processing, Beijing* 100083)

Abstract Genetic algorithms (GA), a global searching method, is applied to select wavelength variables of near infrared spectroscpy (NIR) for multivariate calibration made by partial least squares (PLS) method. Two application examples of NIR analysis show that this wavelength selection method for PLS modeling not only simplifies and optimizes calibration model but also increases the prediction ability of calibration model. The method is especially adequate for the system where only PLS is difficult to correlate.

Keywords Genetic algorithms, partial least squares, variable selection, gasoline, aromatics, lubricant, saturated hydrocarbon

(Received 29 May 2000; accepted 24 November 2000)