文章编号：0253-2239(2007)07-1316-6

用遗传算法快速提取近红外光谱特征区域和特征波长*

邹小波[1,2] 赵杰文[1]

(1 江苏大学农产品加工研究所，镇江 212013
2 江苏恒顺集团有限公司，镇江 212004)

摘要： 提出了一种遗传区间偏最小二乘法(GA-iPLS)，并用该方法快速提取苹果糖度近红外光谱的特征区域，在此基础上采用遗传偏最小二乘法(GA-PLS)提取苹果糖度近红外光谱的特征波长，进行苹果糖度预测。结果表明，整个光谱等分为 40 个子区间，遗传区间偏最小二乘法能快速寻找出 5 个特征子区间(第 4，6，8，11，18 号)；在 5 个特征子区间的基础上用遗传偏最小二乘法继续优化，从中提取 44 个特征波长。建立在 5 个特征子区间和 44 个特征波长上的偏最小二乘法模型精度均优于全光谱偏最小二乘法模型，对预测集的预测相关系数提高了近 10%；且模型得到了很大的简化，用于建模的主因子数减少了 7 个。这些结果表明，用这两种方法不但可以建立简洁、数据运算量少的模型，还可以快速地提取近红外光谱的特征区域和特征波长。

关键词： 光谱学；近红外光谱；遗传算法；偏最小二乘法；糖度；苹果

中图分类号：O433；S123 文献标识码：A

Methods of Characteristic Wavelength Region and Wavelength Selection Based on Genetic Algorithm

Zou Xiaobo[1,2] Zhao Jiewen[1]

(1 *Agricultural Product Processing Research Institutes, Jiangsu University, Zhenjiang* 212013
2 *Jiangshu Hengshun Group, Limited Liability Company, Zhenjiang* 212004)

Abstract: Genetic algorithm interval partial least square (GA-iPLS) and genetic algorithm partial least square (GA-PLS) were proposed to select the characteristic wavelength region and characteristic wavelength of sugar content against apple near-infrared spectra for sugar content prediction. The apple near-infrared spectra data were divided into 40 intervals. Consequently, 5 subsets (No. 4, 6, 8, 11, 18) and 362 data points were selected quickly by GA-iPLS, and 44 characteristic wavelengths were selected by GA-PLS based on the 5 subsets. Compared with the whole spectra data model, the GA-iPLS and GA-PLS models could not only improve precision with the coefficients of determination for prediction set improved by 10%, but also simplify the model with 7 primary factors decreased in the model. With the proposed methods, a concise easily computed model can be built to select the characteristic reigon and wavelength of near-infrared spectra.

Key words: spectroscopy; near-infrared spectra; genetic algorithm; partial least square; sugar content; apple

1 引 言

随着近红外光谱技术和化学计量方法的发展，近红外光谱技术在农产品品质分析中的应用越来越广泛。借助先进的近红外光谱仪，研究者可以在短

* 国家自然基金(30370813，30671199)、江苏省创新人才启动基金(BK2006552)和江苏省自然科学基金(BK2006707-1)资助课题。

作者简介：邹小波(1975—)，男，湖南汨罗人，博士后，主要从事农产品品质检测方面的研究。
E-mail：zou_xiaobo@ujs.edu.cn

导师简介：赵杰文(1945—)，男，江苏苏州人，教授，博士生导师，主要从事农产品品质无损检测技术方面的研究。
E-mail：zhao @ujs.edu.cn

收稿日期：2006-08-22；收到修改稿日期：2006-11-16

时间内很方便地获得大量光谱数据。

在测定苹果中的糖度(可溶性固形物)含量时，由于其存在形式已经不是简单的单糖和多糖，特征谱区也就不是某种单糖或多糖的特征谱区，所以确定苹果中的固形物含量的特征谱区是比较困难的。另外，由于近红外区的谱带复杂、重叠多，通过苹果近红外光谱的分析可以看出，光谱的总体走势比较平缓，波峰和波谷没有剧烈的起伏，且不同区域的信噪比不同[1,2]。因此，以往的研究中，很少有将光谱仪所采集的所有数据用来建模，大都采用其中一段光谱来进行建模[2~14]。因此，通过一些现代的数据处理方法从这些烦杂的数据中提取有用的信息——提取特征光谱区域和特征波长是一个值得研究的问题。

在利用近红外光谱技术检测苹果糖度的研究中发现，目前常用的图示法和专家经验选取波峰、波谷和组分特征波长建立模型都具有主观性，且模型预测精度不高。1998 年 R. Leardi[15] 提出一种遗传偏最小二乘法(GA-PLS)来进行光谱特征波长的筛选，并在短波近红外光谱中得到成功应用，但参与该方法的光谱点个数不能太多，否则算法很难收敛。2000 年 Lars Nørgaard[14] 提出一种区间偏最小二乘法(iPLS)来进行光谱区间的筛选，但当所选的光谱区间比较多时，该方法运算时间会很长。本研究尝试将遗传算法和区间偏最小二乘法相结合以提高特征区间的选择速度，并将该方法称为遗传区间偏最小二乘法(GA-iPLS)。针对所研究的苹果近红外光谱，首先用遗传区间偏最小二乘法进行糖度特征光谱区域的选择，并在此基础上采用遗传偏最小二乘法进行糖度特征波长的筛选。

2　算法简介

本文在进行糖度特征光谱区域选择时所用遗传区间偏最小二乘法是对 Lars Nørgaard[14,16] 于 2000 年提出的区间偏最小二乘法的一种改进和发展。而用于苹果糖度特征波长筛选的遗传偏最小二乘法就是 R. Leardi[15,17] 提出的方法。遗传区间偏最小二乘法和遗传偏最小二乘法都利用遗传算法全局快速搜索的优点，将遗传算法和偏最小二乘法有机的结合起来，发挥各自的长处，建立更加稳定、简便、预测能力更强的模型。其基本思想是将偏最小二乘法交互验证中因变量的预测值和实际值的相关系数(r)作为遗传算法的适应度函数，用遗传算法进行近红外光谱快速分析中的波长筛选，再用偏最小二乘法方法对筛选后的波长变量建立分析校正模型。下面对作者所提出的遗传区间偏最小二乘法的算法进行介绍。

遗传区间偏最小二乘法包括遗传编码、适用度函数设计、遗传操作等。

1) 特征光谱区间入选编码

首先将整个苹果近红外光谱等分为 s 个区间，对这 s 个区间入选的问题，可用一含有 s 个 0/1 字符(基因)的字符串(染色体串)来表示每种区间组合。字符串 0 和 1 分别代表对应区间未被选中和选中，例如对 8 个区间的问题区间组合“00110101”表示第 3，4，6，8 个区间被选中，其余则未被选中。

2) 适应度函数的设计

采用偏最小二乘法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：对每个个体所选的区间进行数据重新组合，再用偏最小二乘法交互验证得到相关系数(r)。相关系数(r)的计算公式如下：

$$r=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}},\qquad(1)$$

式中 N 为样品个数，$\bar{x}$ 为交互验证预测值的均值，$\bar{y}$ 为实际测量值的均值。

3) 初始群体

本研究的初始种群由计算机随机产生的 m 个个体组成，而每个个体由 s 个字符组成。

4) 遗传操作设计

选择算子采用最常用的选择方法 —— 适应度比例方法，也称转轮法，即每个个体的选择概率与其适应度成比例。

交叉算子采用单点交叉方法(如图 1 所示)，参与交叉的个体概率为一个小于 1 的小数(如 0.8)。

变异算子采用基本变异算子，即在某个个体(字符串)中随机挑选一个或多个基因(字符)进行变异，参与变异的个体概率也为一个小于 1 的小数(如 0.1)，它通常比较小。

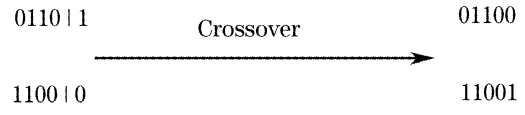

图 1　交叉算子

Fig. 1　Crossover

5) 运算终止条件

本文以遗传迭代次数达到设定的交互验证均方根误差(RMSECV)为收敛终止条件。交互验证均

方根误差值用 E_R 表示可按下式计算：

$$E_R=\sqrt{\sum_{i=1}^{n}(\hat{y}_l-y_l)^2/n},\qquad(2)$$

式中 y_l 和 $\hat{y}_l$ 分别为交互验证集中第 l 个样本的糖度实测值和预测值，n 为交互验证集样本数。

6）区间选取

本文采用的方法为：在遗传迭代后，具有最小均方根差交互验证 E_R 的区间组合中的所有入选区间为特征光谱区间。

3 实 验

实验选用山东水晶红富士苹果 85 个，购回后从中随机地选取 63 个作为校正集，余下的 22 个作为预测集，将它们分别编号后置于 4℃ 冰柜中贮藏。光谱检测实验在环境温度可控的实验室（本实验环境温度控制为 24℃）内进行。实验前，将冰柜中取出的苹果置于实验室中 3 h，以使苹果整体温度达到与环境温度的一致。实验时，由近红外光谱仪（Nexus670 FT-IR，美国 Nicolet 公司生产）在每个苹果的最大横径上进行光谱扫描，扫描波数范围为 4279～9843 cm^{-1}，扫描次数为 32 次，波数间隔为 1.924 cm^{-1}（共 2886 个波数点），分辨力为 4 cm^{-1}，动镜速度为 0.9494 cm/ s，光圈为 50，以 $BaSO_4$ 作为参比材料。扫描时光纤探头与苹果表面之间间隔保持 1～3 mm 的距离。图 2 为所采集的苹果近红外光谱经过去均值后的结果，去均值的目的是去除每次测量光谱整体能量的影响。采集完光谱后将该苹果削皮，取可食用部分榨汁，并用手持式糖度计（WYT0-32 型，泉州韦达计量仪器厂生产）测定其糖度值，表 1 列出了被测苹果糖度实测值的变化范围、平均值、标准偏差及变异系数。

表 1 糖度实测值的统计表

Table 1 Statistic of apple sugar content

	Number of samples	Mean (° brix)	Maximum (° brix)	Minimum (° brix)	Standard deviation	Coefficient of variance /%
Calibration set	63	12.855	16.6	9.4	1.50	11.67
Prediction set	22	12.795	15.8	9.0	1.473	11.51

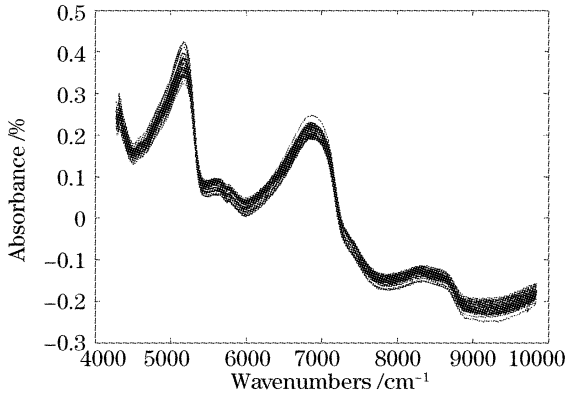

图 2 用于建模的富士苹果光谱

Fig. 2 "Fuji" apple spectra for model building

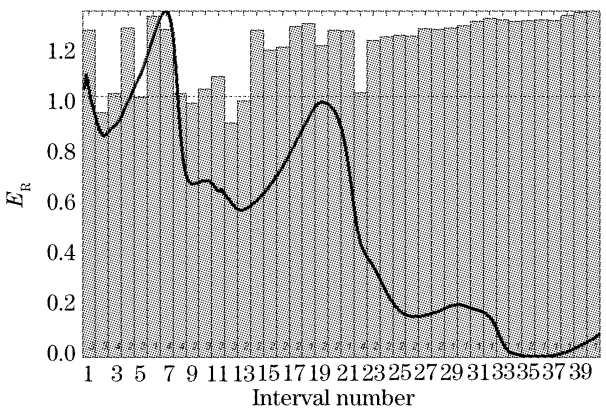

图 3 各区间模型的 E_R 值与全光谱模型的 E_R 值比较图（点线表示全光谱模型的 E_R 值，横轴上的斜体数字表示各局部模型的最佳主因子数，全光谱模型的主因子数为 3）

Fig. 3 E_R comparison of interval model and global model [dotted line is E_R (3Latent varibles) for global model/Italic numbers are optimal Latent varibles in interval model]

4 实验数据处理

将图 2 的光谱数据（光谱范围 4279～9843 cm^{-1}）等分为 40 个区间（其中第 1～6 号区间每个区间波数点为 73 个，余下的区间每个区间波数点为 72），用 Lars Nørgaard[16] 的区间偏最小二乘法进行处理，图 3 为处理后的情况。由图 3 可以看出，其中建立在第 2，9，12，13 各个区间上的偏最小二乘法模型的 E_R 比全光谱模型的 E_R 小，说明并不是光谱数据越多越好。因此，怎样才能得到最优的区间组合是本文的研究重点。

4.1 遗传区间偏最小二乘法选取特征光谱区域

下面就用遗传区间偏最小二乘法从这 40 个区间中选取特征光谱区域。设定优化参量：区间数 40，初始群体 60，最大选取变量数 40，交叉概率0.8，变异概率 0.1，遗传迭代次数 200，均方根差交互验

证 E_R 为 0.4 布里克斯糖度(E_R = 0.4 ° brix)。图 4 为每代中最小 E_R 随遗传算法进化 73 代的情况。对应偏最小二乘法模型最小 E_R 的光谱区间为第 4,6,8,11,18,如图 5 所示。因此遗传区间偏最小二乘法所选取的特征光谱区域为第 4,6,8,11,18 五个区间所对应的区域,此时对应的波长变量数为 362。

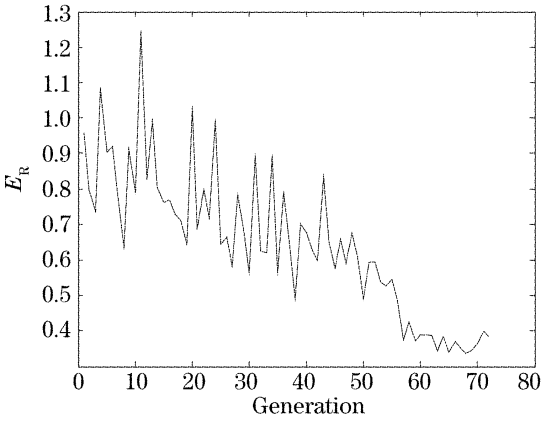

图 4　每代中最小 E_R 随遗传代进化的情况

Fig. 4　Minimum E_R values of partial least square regression models variation with generation

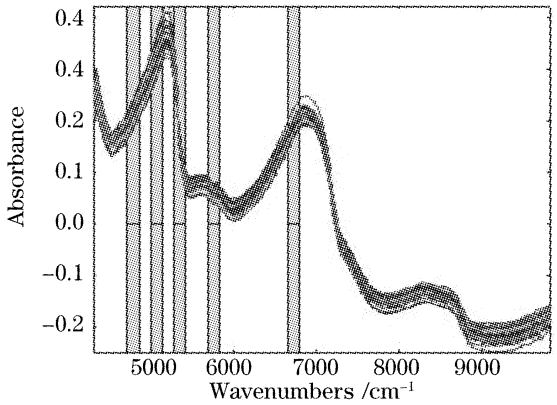

图 5　遗传偏最小二乘法所选取的特征光谱区域

Fig. 5　Characteristic spectral region selected by genetic algorithm interval partial least square

4.2　遗传偏最小二乘法选取特征波长

用 R. Leardi 所提出的遗传偏最小二乘法方法对前面遗传区间偏最小二乘法所选择的 5 个区间内的变量进行进一步筛选。遗传迭代的参量设置为:初始群体 70,最大选取变量数 362,交叉概率 0.8,变异概率 0.1,遗传迭代次数 400,在遗传迭代后,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,便得到最佳模型。图 6 为经过 400 次迭代后变量选取的频率图。图 7 为所有变量按选取频率重新排列后,相关系数(r%)随选取变量数的逐步增加而变化的趋势图,由图可以看出最佳的变量数为 44,此时建立的糖度偏最小二乘法模型对校正集的测试结果相关系数达到 0.9355。因此这 44 个变量即为遗传偏最小二乘法选取的苹果糖度近红外特征波长。

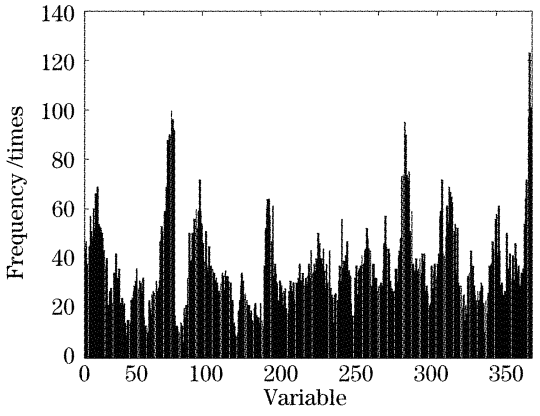

图 6　糖度变量选取频率图

Fig. 6　Selected frequency versus sugar content variable

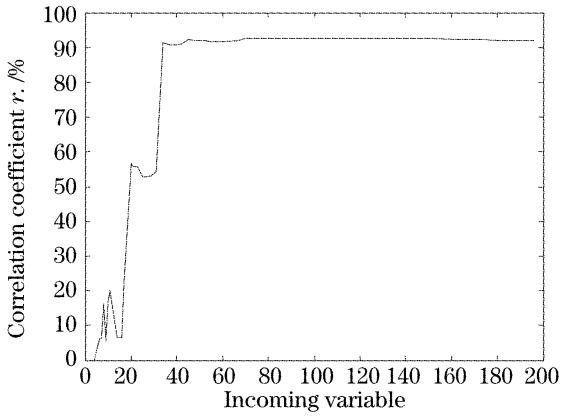

图 7　相关系数随选取变量数的变化趋势

Fig. 7　Correlation coefficient versus independent variables

4.3　与全光谱比较

为了比较遗传区间偏最小二乘法和遗传偏最小二乘法处理的效果,将所建模型分别与全光谱模型进行比较(全光谱模型的光谱范围为 4279～9843 cm^{-1}),结果如表 2 所示。所有模型建模过程中最佳主因子数由交互验证法(Cross-validation)确定,即由最小的预测残差平方和确定。

从表 2 可以看出,全光谱的偏最小二乘模型预测苹果糖度的精度不高,且该模型采纳的最佳因子数为 13,这使得模型显得过于复杂。遗传区间偏最小二乘法处理所得的最佳苹果近红外光谱模型是建立在 5 个光谱区间(共 362 个波数点),其不论对校正集还是预测集模型的预测能力都优于全光谱模型,且该模型得到了很大的简化:其实际采用的波数点个数比全光谱模型采用的波数点 2886 个大大减少;且采纳的最佳主因子数也减少了许多,这样运算量也减少了许多。遗传偏最小二乘是对遗传区间偏最小二乘选取的光谱区间内的变量进一步的筛选和

优化，选取的特征波长为 44 个，所建立的偏最小二乘法模型虽然其 E_R 有一些增加，但它对预测集的预测能力没有改变，且主因子数为 5，这说明遗传偏最小二乘法所建立的苹果糖度近红外光谱模型比全光谱模型更简洁、更稳健。

表 2 不同处理方法处理后的偏最小二乘校正结果

Table 2 Summary of partial least square results after apple spectral being treated by different methods

Election methods	Selected wavenumber range /cm^{-1}	Variables in model	Number of factors	Calibration set		Prediction set	
				E_R(° brix)	r_c	E_R(° brix)	r_p
Full-spectrum PLS	4279.34～9843.06	2886	13	0.5542	0.8808	0.6334	0.8362
GA-iPLS (5 intervals selected)	4701.6～4840.5; 4983.5～5122.0 5263.0～5399.8; 5679.5～5816.4 5818.3～5955.2	362	7	0.3346	0.962	0.3846	0.932
GA-PLS	5252.19, 5254.12, 5250.26, 5016.92, 5256.05, 5018.84, 5527.97, 5529.90, 5059.34, 5760.75, 5814.99, 5315.55, 5415.83, 5795.18, 5797.11, 5767.01, 5811.42, 5816.40, 5348.34, 5317.76, 5319.69, 5899.04, 5814.47, 5883.89, 5708.16, 5787.23, 5028.29, 5368.94, 5403.65, 5453.79, 5839.49, 5901.21, 6047.77, 6188.55, 6213.62, 6298.48, 6300.40, 6302.33, 5812.54, 5826.04, 5881.97, 5026.37, 4855.96, 5027.60.	44	5	0.3526	0.934	0.3757	0.936

PLS: Partial least square, GA-iPLS: Genetic algorithm interval partial least square; GA-PLS: genetic algorithm partial least square

5 结 论

用遗传区间偏最小二乘法和遗传偏最小二乘法对苹果近红外光谱进行特征光谱区域和特征波长的选取。结果发现，与全光谱模型相比，遗传区间偏最小二乘法和遗传偏最小二乘法不仅能有效地减少建模所用的变量数，而且能有效地提高苹果糖度模型的测量精度。通过这两种方法选取合适的光谱区间和波长进行建模，可以减小建模运算时间，剔除噪声过大的谱区，使最终建立的农产品品质检测近红外光谱模型的预测能力和精度更高。该结果可为设计滤光片式或激光式近红外快速检测仪提供一种客观选择特征波长的方法。

参 考 文 献

1 Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. *Progress in Chemistry*, 2004, **16**(4): 528～542 (in Chinese)
褚小立，袁洪福，陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展, 2004, **16**(4): 528～542

2 Li Hui, Xie Shusen, Lu Zukang *et al*.. A new model of the light scattering in biological tissue for visible and near infrared region [J]. *Acta Optica Sinica*, 1999, **19**(12): 1661～1666 (in Chinese)
李 晖，谢树森，陆祖康 等. 生物组织的可见光与近红外光散射模型[J]. 光学学报, 1999, **19**(12): 1661～1666

3 Zhao Jiewen, Zhang Haidong, Liu Muhua. Preprocessing methods of near infrared spectra fro simplifying prediction model of surgar content of apples[J]. *Acta Optica Sinica*, 2006, **26**(1): 1136～1139 (in Chinese)
赵杰文，张海东，刘木华. 简化苹果糖度预测模型的近红外光谱预处理方法[J]. 光学学报, 2006, **26**(1): 1136～1139

4 Zhao Jiewen, Zhang Haidong, Liu Muhua. Non-destructive determinat ion of sugar contents of apples using near infrared diffuse reflectance[J]. *Transactions of the CSAE*, 2005, **21**(3): 162～165 (in Chinese)
赵杰文，张海东，刘木华. 利用近红外漫反射光谱技术进行苹果糖度无损检测的研究[J]. 农业工程学报, 2005, **21**(3): 163～165

5 Ann Peirs, Jeroen Tirry , Bert Verlinden *et al*.. Effect of biological variability on the robustness of NIR models for soluble solids content of apples [J]. *Postharvest Biology and Technology*, 2003, **28**(3): 269～280

6 O. Kleynen, V. Leemans, M.-F. Selection of the most efficient wavelength bands for "Jonagold" apple sorting[J]. *Postharvest Biology and Technology*, 2003, **30**(1): 221～232

7 I. Wayan Budiastra, Yoshio Ikeda, Takahisa Nishizu. Optical methods for quality evaluation of fruits (part 2)-prediction of individual sugars and malic acid concentrations of apples and mangoes by the developed NIR reflectance system[J]. *J. JSAM*, 1998, **60**(3): 117～128

8 V. Steinmetz, J. M. Roger, E. Molto *et al*.. On-line fusion of color camera and spectrophotometer for sugar content prediction of apples[J]. *J. Agric. Engng. Res.*, 1999, **73**(4): 207～216

9 K. H. S. Peiris, G. G. Dull, R. G. Leffler *et al*.. Spatial

variability of soluble solids or dry-matter content within individual fruits, bulbs, or tubers: implications for the development and use of NIR spectrometric techniques[J]. *Hori Science*, 1999, **34**(1): 114～118

10 Ann Peirs, J. Lammertyn, K. Ooms *et al*.. Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR-spectroscopy [J]. *Postharvest Biology and Technology*, 2000, **21**(3): 189～199

11 J. Lammertyn, Ann Peirs, Josse De Baerdemaeker *et al*.. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment[J]. *Postharvest Biology and Technology*, 2000, **18**(1): 121～132

12 Renfu Lu, Daniel E. Guyer, Randolph M. Beaudry. Determination of firmness and sugar content of apples using near-infrared diffuse reflectance [J]. *J. Texture Studies*, 2000, **31**(6): 615～630

13 B. Park, J. A. Abbott, K. J. Lee *et al*.. Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and Gala apples[J]. *Transactions of the ASAE*, 2003, **46**(6): 1721～1731

14 L. Nørgaard, A. Saudland, J. Wagner *et al*.. Interval partial least squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy [J]. *Applied Spectroscopy*, 2000, **54**(3): 413～419

15 R. Leardi, A. Lupiáez, González. Genetic algorithms applied to feature selection in PLS regression: how and when to use them [J]. *Chemometrics and Intelligent Laboratory Systems*, 1998, **41**(2): 195～207

16 R. Leardi, L. Nørgaard. Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions[J]. *J. Chemomelrics*, 2004, **18**(11): 486～497

17 R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets [J]. *J. Chemometrics and Intelligent Laboratory Systems*, 2000, **14**(5): 643～655