DOI:10.3969/j.issn.1000-1298.2010.02.028

基于小波变换的大米直链淀粉波长选择方法*

张巧杰[1]　张　军[2]

(1. 北京信息科技大学自动化学院,北京 100192; 2. 北京联合大学信息学院,北京 100101)

【摘要】 提出了一种基于小波变换的近红外光谱波长选择方法,小波分解低频系数是原光谱的离散近似,将最佳小波分解低频系数与原光谱数据进行关联,求出小波分解低频系数与原光谱数据的列相关系数 R,取与原光谱数据相关系数较大的波长组合,作为最后参与建模的谱区。不仅考虑了浓度矩阵对波长选择的影响,且由于把小波分解结构中的高频系数全部滤除,避免了高频噪声的干扰,减小建模和预测运算时间,使最终建立的近红外光谱模型的预测精度提高。在大米直链淀粉含量的近红外光谱分析中进行了验证,并与其他常用波长选择方法进行了比较,结果表明,该方法波长点数最少,减小为原光谱数据点数的 20%,校正模型和预测效果都较理想。

关键词: 直链淀粉　近红外光谱　小波变换　波长选择

中图分类号: TS207.3; O657.33　　**文献标识码**: A　　**文章编号**: 1000-1298(2010)02-0138-05

Region Selecting Methods of Near Infrared Wavelength Based on Wavelet Transform

Zhang Qiaojie[1]　Zhang Jun[2]

(1. *College of Automation*, *Beijing Information Science & Technology University*, *Beijing* 100192, *China*
2. *College of Information*, *Beijing Union University*, *Beijing* 100101, *China*)

Abstract

An efficient method was presented to select wavelength regions of NIR(near infrared spectroscopy) based on WT(wavelet transform) for building a PLS(partial least squares) calibration model. Wavelet approximation coefficient(WAC) is similar matrix of original NIR. By linking the correlation coefficients of the optimal WAC with original NIR data, a research space with the all combinations of strong correlativity wavelength was selected as final wavelength regions to build a PLS calibration model of NIR. This selecting method considered the influence of density matrix on wavelength selection, filtered wavelet detail coefficient entirely, and avoided the interference of high frequency noise. The running time of building model was reduced enormously so that the final model is of higher accuracy. Compared with other traditional wavelength selecting method, This method is validated in the measurement of rice apparent amylose content by NIR. The apparent amylase content test results showed that the number of wavelengths for building the models can be reduced to 20% of the original method, the calibration model and the prediction precision are greatly improved by wavelet transform algorithm.

Key words　Amylose, NIR, Wavelet transform, Wavelength selecting

引言

对于农产品及食品成分检测,近红外光谱分析法是一种较理想方法。在直链淀粉含量检测方面,国内外进行了大量的研究[1~6],结果表明,依据全区光谱数据建立的 PLS 模型具有最佳的性能。用神经网络对不同粒度、不同类型的大米样品进行近红外光谱分析,建立了大米直链淀粉含量的预测模型,

收稿日期: 2009-01-09　修回日期: 2009-03-18

*北京市优秀人才培养资助项目(9110823301)

作者简介: 张巧杰,讲师,主要从事农产品品质检测方法研究,E-mail: qiaojiezhang@yahoo.cn

精米样品模型预测值与化学分析值的相关系数达0.95，预测标准差、平均相对误差分别为 0.56 和 3.1%[3~4]。还可以利用近红外反射光谱技术研究建立回归方程时样品群的界定与选择。

目前，波长选择的方法主要有相关系数法、方差分析法、逐步回归分析方法、无信息变量消除方法、遗传算法等[7]。小波变换（wavelet transform，简称 WT）具有很好的时频分离特征，信息处理能力强，已广泛用于分析化学领域，在近红外光谱分析的预处理、数据压缩、模式识别以及模型传递等方面也被应用。本文将小波变换用于近红外光谱分析的波长选择，提出一种基于小波变换的近红外光谱波长选择方法，并研究其在大米直链淀粉波长选择中的应用。

1　基本原理

1.1　小波变换的计算方法

小波变换的实质就是将信号 $f(t)$ 投影到小波，得到便于处理的小波系数 w，按照分析的需要对小波系数进行处理，然后对处理后的小波系数进行反变换得到处理后的信号。化学分析的信号一般是离散信号，常用离散小波变换。

已有几种不同的离散小波变换计算方法，其中应用最广泛的是 Mallat 提出的多分辨信号分解算法或塔式算法，又称 Mallat 算法。Mallat 算法小波分解[8~9]过程可以写成

$$C^j(k)=\sum_{n\in Z}h^*(n-2k)C^{j-1}(n)\quad(j=0,1,\cdots,J) \tag{1}$$

$$D^j(k)=\sum_{n\in Z}g^*(n-2k)C^{j-1}(n)\quad(j=0,1,\cdots,J) \tag{2}$$

式中　J——最高分解层数

$C^j(k)$——原始数据的低频近似

$D^j(k)$——原始数据的高频细节

$C^j(k)$ 包括信号的主要信息，表示原始信号中频率低于 2^{-j} 的低频分量，$D^j(k)$ 包括信号的细节部分，表示频率介于 2^{-j} 与 $2^{-(j-1)}$ 之间的高频分量。

1.2　基于小波变换的波长选择方法

由于小波分解低频系数是原光谱的近似，是对除去高频噪声系数后原光谱的替代，能够最大限度地表征原数据结构特征。基于小波变换的波长选择方法首先将原光谱用小波变换进行数据分解，根据小波分解低频系数与直链淀粉含量矩阵所建模型的优劣，找出最佳小波分解低频系数。然后将最佳小波分解低频系数与原光谱数据进行关联，求出最佳小波分解低频系数与原光谱数据的列相关系数 R，取与原光谱数据相关系数较大的波长组合来建模。最佳小波分解低频系数是根据所建模型的质量选择出来的，最后参与建模的波长是最佳低频系数与原光谱相关性大的波长组合，不仅考虑了浓度矩阵对波长选择的影响，而且由于将小波分解的高频系数全部滤除，避免了高频噪声的干扰。之所以选择与最佳小波分解低频系数相关系数大的波长组合来建模，是因为相关系数大的光谱区间携带原光谱矩阵的信息比较多，对样品的组成或性质影响也较大。

采用预置的相关系数阈值，自动选择那些相关系数大于阈值的波长点建立校正模型。在用偏最小二乘法[9]（partial least squares，简称 PLS）建立近红外光谱模型时，波长范围选择的目标是，选择最合适的一个或几个波长范围的吸光度光谱数据参与建模，要求所建模型在最佳主成分的交叉校验预测值与标准值的决定系数 R^2 最大且交叉验证标准差 S_{ECV} 最小（否则预测能力和精度不高）。阈值的选择根据所建模型 R^2 以及 S_{ECV} 值来确定。取不同的阈值对波长选择并建模，可得到相应的 R^2 和 S_{ECV}，R^2 最大且 S_{ECV} 最小的阈值对应的波长组合即为所求。

在最佳主成分 f_{opt} 时的交叉验证预测值与标准值的决定系数 R^2 最大可表示为

$$\max R^2=\frac{\left[\sum_{i=1}^{n}(c_i^P-\bar{c}^P)(c_i^O-\bar{c}^P)\right]^2}{\left[\sum_{i=1}^{n}(c_i^P-\bar{c}^P)^2\right]\left[\sum(c_i^O-\bar{c}^P)^2\right]} \tag{3}$$

交叉验证标准差 S_{ECV} 最小可表示为（1 + S_{ECV}）的倒数最大

$$\max\lambda=\frac{1}{1+S_{ECV}}=\frac{1}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i^O-c_i^P)^2}} \tag{4}$$

其中　$\bar{c}^O=\frac{1}{n}\sum_{i=1}^{n}c_i^O$　$\bar{c}^P=\frac{1}{n}\sum_{i=1}^{n}c_i^P$

式中　n——参与建模样品数目

c_i^O——样品 i 的某一组分浓度标准值

c_i^P——交叉校验时样品 i 的某一组分浓度预测值

$\bar{c}^O$——参与建模 n 个样品的某一组分浓度标准值的平均值

$\bar{c}^P$——交叉校验时参与建模 n 个样品的某一组分浓度预测值的平均值

基于小波变换的波长选择算法如图 1 所示。

2　实验

2.1　仪器

采用德国 Bruke 公司 MATRIX－I 型傅里叶变

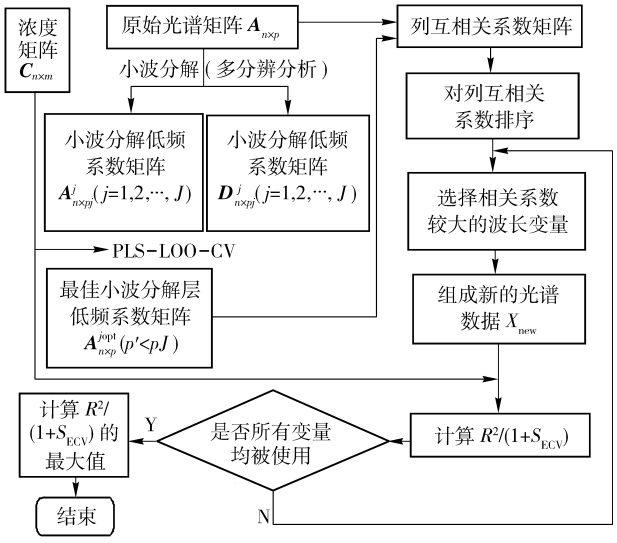

图 1 基于小波变换的波长选择算法示意图

Fig. 1 Scheme for select wavelength regions of spectrums based on wavelet approximation coefficient

换近红外光谱仪,高灵敏度 24 位数字化 PbS 监测器,光谱采集范围为 800 ~ 2 500 nm,分辨率为 16 cm^{-1},波长点数 1 102,石英样品池。

2.2 样品与基础数据来源

107 个大米样品由中国农业科学院作物品种资源所提供,选用的样品覆盖了全国主要产稻区的典型品种,直链淀粉含量的变幅范围为 1% ~26.6%,标准值由该所按照 NY/T 55 - 1987《水稻、玉米、谷子籽粒直链淀粉测定法》测定。将样品随机分为两集:校正集样品 90 个,预测集样品 17 个。

2.3 光谱采集

以空气为参比,扫描次数为 64 次,107 个糙米粉的近红外光谱图如图 2 所示。

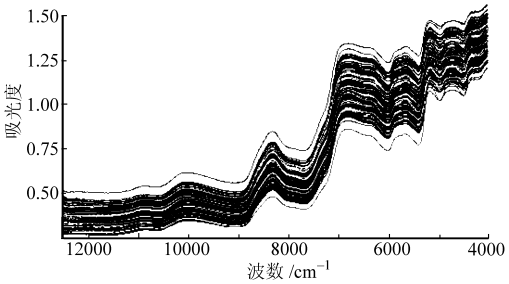

图 2 107 个糙米粉的近红外光谱图

Fig. 2 Near-infrared spectra of 107 samples

3 结果与分析

3.1 基于小波变换的近红外波长选择

分别对 90 个校正集样品的原光谱,选择 db2 基本小波进行 $J=6$ 的小波分解,每一个样品光谱都可以得到 6 个小波分解低频系数向量,取所有样品对应各层的小波分解低频系数组合成 6 个小波分解低频系数矩阵。以这 6 个低频系数矩阵代替原始光谱矩阵,对直链淀粉含量进行偏最小二乘留一法交叉验证(partial least squares leave one out cross validation,简称 PLS - LOO - CV),其验证结果如表 1 所示。从表中可以看出,在 6 个小波分解层上,当分解层数为 5 时 PLS - LOO - CV 效果最好。与使用原始光谱相比,使用小波低频系数进行 PLS - LOO - CV 的相关系数 R 从 0.921 2 提高到 0.950 7,决定系数 R^2 从 0.848 7 提高到 0.903 8,交叉验证标准差 S_{ECV} 从 3.366 7 减小到 2.721 3。当小波分解层数较小时,PLS - LOO - CV 效果改善不大,主要原因是在较低的分解层,小波低频系数中仍然包含原信号的高频成分。而在较高的分解层,PLS - LOO - CV 效果变差,主要是因为在较高的小波高频系数中已经包含了原信号的有用信号,即低频系数中已经失去了部分有用信息,直接以低频系数代替原光谱进行建模,效果当然会变差。

表 1 90 个糙米粉样品原光谱、6 个小波低频系数矩阵 PLS - LOO - CV 结果

Tab. 1 Results of PLS - LOO - CV for 90 samples near-infrared spectras and sixth WAC

类型	分解层数 j	数据点数	相关系数 R	决定系数 R^2	S_{ECV}
原始光谱	无	1 102	0.921 2	0.848 7	3.366 7
小波低频系数	1	552	0.931 1	0.867 0	3.192 7
	2	277	0.933 4	0.871 2	3.141 3
	3	140	0.938 3	0.880 4	3.061 6
	4	71	0.950 3	0.903 1	2.743 6
	5	37	0.950 7	0.903 8	2.721 3
	6	20	0.940 6	0.884 7	2.980 5

计算小波分解 $J=5$ 时的低频系数与原光谱的列相关系数,设置不同的相关系数阈值,对选择出的波长组合进行建模,计算模型的 R^2 及 S_{ECV},相关系数阈值与对应模型的 $R^2/(1+S_{ECV})$ 关系图如图 3 所示。

从图 3 可以看出,当阈值为 0.999 83 时,PLS - LOO - CV 效果最好。用阈值为 0.999 83 时选择的波长组合建立 90 个样品的校正模型,波长选择前波长点数为 1102,决定系数 R^2 为 0.850 1,S_{ECV} 为 3.404 3。波长选择后波长点数变为 221,而决定系数 R^2 为 0.927 6,S_{ECV} 为 2.357 5。

用该模型预测 17 个预测集样品的测定结果如表 2 所示。从表 2 可以看出,波长选择前,样品预测值的最大偏差为 -8.72,波长选择后,预测值的最大偏差为 -4.66,SEP 由 4.45 减小到 2.65,预测结果得到较大幅度提高。

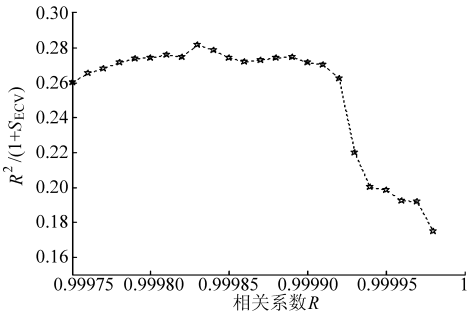

图 3　小波变换相关系数阈值与模型精度关系图

Fig. 3　Correlation coefficient threshold with $R^2/(1+S_{ECV})$ of calibration

表 2　小波变换波长选择前后校正模型对 17 个样品直链淀粉含量的预测结果

Tab. 2　Prediction results before and after selecting wavelength regions by WAC

样品号	标准值/%	波长选择前		波长选择后	
		预测值/%	偏差	预测值/%	偏差
1	13.30	12.32	0.98	10.67	2.63
2	17.20	22.59	−5.39	17.91	−0.71
3	16.20	22.00	−5.80	20.86	−4.66
4	14.90	21.64	−6.74	16.23	−1.33
5	13.70	18.95	−5.25	18.16	−4.46
6	1.00	9.72	−8.72	1.29	−0.29
7	16.30	22.76	−6.46	18.22	−1.92
8	1.70	−1.29	2.99	2.32	−0.62
9	1.30	4.26	−2.96	5.34	−4.04
10	1.40	3.24	−1.84	3.82	−2.42
11	15.20	15.13	0.07	17.19	−1.99
12	25.40	22.35	3.05	22.64	2.76
13	13.10	20.33	−7.23	16.92	−3.82
14	18.20	17.07	1.13	18.93	−0.73
15	17.20	17.35	−0.15	19.29	−2.09
16	19.30	22.30	−3.00	21.73	−2.43
17	15.30	17.77	−2.47	17.62	−2.32
SEP			4.45		2.65

3.2　常用波长选择方法比较

目前，波长选择的常用算法主要有相关系数法[9]、显变分析法[9]、无信息变量消除法[9]、遗传算法[9]等，这些算法的具体实现步骤参见文献[9]，其中遗传算法的研究和应用较为广泛。基于小波变换的波长选择方法与上述算法用于 90 个校正集样品和 17 个预测集样品的波长选择结果比较如表 3 所示。

表 3　不同波长选择方法对 90 个糙米粉样品的校正结果和 17 个糙米粉样品的预测能力

Tab. 3　Calibration and prediction results before and after selecting wavelength regions by different methods

方法	波长点数	校正集		预测集	
		R	S_{ECV}	R	RMSEP
原光谱	1 102	0.921 2	3.404 3	0.876 4	4.448 3
相关系数法	916	0.933 0	3.149 8	0.888 5	4.268 0
显变分析法	810	0.932 6	3.160 2	0.876 8	4.705 0
无信息变量消除法	726	0.964 8	2.298 5	0.934 7	3.772 3
遗传算法	332	0.958 7	2.488 1	0.958 2	2.641 7
小波变换法	221	0.963 1	2.357 5	0.956 2	2.651 8

结果表明，5 种波长选择方法进行波长选择后，和原光谱进行比较，模型得到不同程度的优化，预测能力也得到不同程度的提高。其中，显变分析方法提高得最少，无信息变量消除法所建模型最好，但其预测能力较遗传算法和小波变换法差。遗传算法的预测能力最好。小波变换法波长点数最少，减少到原光谱数据点数的 20%，校正模型比遗传算法好，预测效果和遗传算法的预测效果相当，但其计算量较遗传算法大大减小。

4　结束语

基于小波变换的波长选择方法，不需要先验知识，可以选择出待测组分浓度预测效果较好的谱区。最佳小波低频系数是根据小波低频系数与浓度矩阵所建模型质量选择出来的，而最后参与建模的波长是最佳小波低频系数与原光谱相关性较好的波长组合，体现出浓度对波长选择的影响。将小波分解结构中的高频系数全部滤除，避免了高频噪声的干扰，减小建模和预测运算时间，使最终建立的近红外光谱模型的预测精度得到提高。与其他常用近红外光谱波长选择方法比较结果表明，基于小波变换的波长选择方法不仅大大简化、优化了校正模型，而且提高了所建模型的预测能力。其校正效果和预测能力和遗传算法用于波长选择的校正效果和预测能力相当，但计算量较遗传算法大大减少。

参　考　文　献

1　Delwiche S. R, Bean M M, Miller R E, et al. Apparent amylase content of milled rice by near-infrared reflectance spectrophotometry[J]. Cereal Chemistry, 1995, 72(2): 182 ~ 187.

2　Villareal C P, Dela Cruz N M, Juliano B O. Rice amylase analysis by near-infrared transmittance spectroscopy[J]. Cereal Chem., 1994, 71(3): 292 ~ 296.

3 刘建学,吴守一,方如明. 大米直链淀粉含量的近红外光谱分析[J]. 农业工程学报. 2000,16(3):94~96.
Liu Jianxue, Wu Shouyi, Fang Ruming, Measurement of rice apparent amylose content by near infrared spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering, 2000,16(3): 94~96. (in Chinese)

4 刘建学,吴守一,方如明. 基于近红外光谱的神经网络预测大米直链淀粉含量[J]. 农业机械学报. 2001,32(2):55~57.
Liu Jianxue, Wu Shouyi, Fang Ruming. Determination of apparent amylose content in rice by neural networks based on near infrared spectroscopy[J]. Transactions of the chinese society for agricultural machinery, 2001, 32(2): 55~57. (in Chinese)

5 谢新华,肖昕,李晓方,等. 用近红外透射光谱技术测定精米直链淀粉含量研究[J]. 食品科学,2004,25(1):118~121.
Xie Xinhua, Xiao Xin, Li Xiaofang, et al. Study on calibration of polished rice amylose by NITS[J]. Food Science, 2004, 25(1): 118~121. (in Chinese)

6 舒晓尧,吴殿星,夏英武. 用近红外反射光谱技术测定精米样品表观直链淀粉含量的研究[J]. 中国水稻科学,1999,13(3):189~192.
Shu Xiaoyao, Wu Dianxing, Xia Yingwu. Apparent amylose content of rice by near infrared reflectance analysis of ground milled samples[J]. Chinese Journal of Rice Science, 1999, 13(3):189~192. (in Chinese)

7 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.
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)

8 祝诗平. 近红外光谱品质检测方法研究[D]. 北京:中国农业大学,2003.
Zhu Shiping. Study on quality detection methods of near infrared spectroscopy analysis[D]. Beijing: China Agricultural University, 2003. (in Chinese)

9 田高友,袁洪福,刘慧颖,等. 小波变换在近红外光谱分析中的应用进展[J]. 光谱学与光谱分析,2003,23(12):1 111~1 114.
Tian Gaoyou, Yuan Hongfu, Liu Huiying, et al. The application of wavelet transform in near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2003, 23(12): 1 111~1 114. (in Chinese)

10 郭婷婷,邬文锦,苏谦,等. 近红外玉米品种鉴别系统预处理和波长选择方法[J]. 农业机械学报,2009,40(增刊):87~92.
Guo Tingting, Wu Wenjin, Su Qian, et al. Effects of spectral pretreatment and wavelength selection on discrimination of maize seed varieties by NIR spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(Sup.): 87~92. (in Chinese)