DOI:10.6041/j.issn.1000-1298.2012.03.021

鲜枣可溶性固形物可见/近红外光谱检测建模方法比较*

张淑娟　张海红　赵艳茹　赵华民
(山西农业大学工学院, 太谷 030801)

【摘要】 对采摘于一枣园的 180 个壶瓶枣样本,随机分成 150 个样本校正集和 30 个样本预测集。用 Field Spec3 光谱仪采集光谱,并进行多元散射校正(MSC)预处理,之后分别利用连续投影算法(SPA)和逐步回归法(SRA)提取特征波长,并结合光谱理论分析确定,再分别基于偏最小二乘法(PLS)和最小二乘-支持向量机(LS-SVM)建立壶瓶枣可溶性固形物含量预测的简化模型和全波段模型。结果表明,全波段 PLS 模型预测的相关系数和预测均方根误差分别为 0.887 4 和 1.088 9,预测效果最好;建立的 MSC-SPA-PLS 模型预测的相关系数和均方根误差分别为 0.799 0 和 1.407 8,建立的 MSC-SRA-PLS 模型预测的相关系数和均方根误差分别为 0.822 4 和 1.385 1,与全波段的 MSC-PLS 相比,精度均降低;建立的 MSC-SPA-LS-SVM 模型预测的相关系数和均方根误差分别为 0.796 3 和 1.145 8,与全波段的 MSC-LS-SVM 相比,精度提高;建立的 MSC-SRA-LS-SVM 模型预测精度很低,不适用。

关键词: 鲜枣　可溶性固形物　可见/近红外光谱　无损检测
中图分类号: S123　　**文献标识码**: A　　**文章编号**: 1000-1298(2012)03-0108-05

Comparison of Modeling Methods of Fresh Jujube Soluble Solids Measurement by NIR Spectroscopy

Zhang Shujuan　Zhang Haihong　Zhao Yanru　Zhao Huamin
(*College of Engineering*, *Shanxi Agricultural University*, *Taigu* 030801, *China*)

Abstract

Totally 180 samples coming from one orchard were divided into calibration set with 150 samples and prediction set with 30 samples. Field Spec3 spectrometer was used for collecting spectra data of 180 fresh jujube samples separately. Then successive projection algorithm and stepwise regression analysis combined with spectral theory were used to process the spectral data after MSC pretreatment. Characteristic wavelengths of 150 samples in calibration set were selected by using SPA and SRA, and the partial least square (PLS) and LS-SVM methods were used to establish models of the fresh jujube soluble solids with the whole spectrum and characteristic wavelengths selected by using SPA and SRA. At last, the models of MSC-PLS model, MSC-LS-SVM model, the MSC-SPA-PLS model, the MSC-SPA-LS-SVM model, the MSC-SRA-PLS model and the MSC-SRA-LS-SVM model were used to predict the soluble solids of 30 samples in the prediction set. The results showed that the correlation coefficient and the root mean square error of prediction of MSC-PLS model for full band are 0.887 4 and 1.088 9 and that is the best. The correlation coefficient and the root mean square error of prediction of MSC-SPA-PLS model and MSC-SRA-PLS model are 0.799 0, 1.407 8 and 0.822 4, 1.385 1, and they are less precise than the MSC-PLS model. The correlation coefficient and the root mean square error of prediction of MSC-SPA-LS-SVM model are 0.796 3 and 1.145 8, and that is more precise than the MSC-LS-SVM model. The precision of MSC-SRA-LS-SVM model is very low and is not suitable.

Key words　Fresh jujube, Soluble solids, NIR spectroscopy, Nondestructive detection

收稿日期: 2011-08-05　修回日期: 2011-08-24
*高等学校博士学科点专项科研基金资助项目(20101403110003)
作者简介: 张淑娟,教授,主要从事数字农业关键技术和装备研究, E-mail: zsujuan@263.net

引言

可见/近红外光谱分析技术能够充分利用全谱段或多波长的光谱数据，对被测物体进行定性或定量分析，已被广泛应用于农业领域[1~13]。

在农产品品质近红外光谱无损检测建模过程中，较多的研究采用偏最小二乘（PLS）和最小二乘-支持向量机（LS－SVM）建模[1~6]。为了消除农产品品质近红外光谱无损检测建模过程中无用信息的干扰，提高建模的速度和效率，需要提取特征波长。目前常用的方法有连续投影算法（SPA）、逐步回归法（SRA）等。齐小明等[7]建立小麦样品蛋白质含量近红外光谱定量分析的 BP 模型时，采用逐步回归法压缩数据，预测结果满足要求。吴迪等[8]用连续投影算法选择由主成分分析（PCA）得到主成分的最佳组合，通过最小二乘-支持向量机对奶粉中脂肪和蛋白质含量进行预测，表明 SPA 能够用于快速有效选取最佳的主成分数，寻优过程简单快速。

可溶性固形物含量是评价鲜枣品质的重要指标，前期的研究中已经利用可见/近红外光谱分析技术对鲜枣中可溶性固形物含量进行快速、无损检测。在主成分分析的基础上，采用 BP 神经网络建模，对鲜枣可溶性固形物含量进行了预测。从预测的精度来看（RMSEP 为 1.990 3），还有上升的空间[11]。

本文采用连续投影算法和逐步回归法，结合近红外光谱理论，对壶瓶枣的全波段可见/近红外漫反射光谱数据进行有效信息的提取，建立偏最小二乘（PLS）和最小二乘-支持向量机（LS－SVM）的简化模型，并与相应的全光谱模型进行精度比较，探讨快速无损检测壶瓶枣可溶性固形物有效建模的新方法。

1　实验设备与方法

1.1　仪器与软件

实验所用光谱仪为美国 ASD（analytical spectral device）公司生产的 Field Spec3，每隔 1 nm 采集一个光谱数据样本，测定范围为 350 ~ 2 500 nm，探头视场角为 10°，光源采用 14.5 V 卤素灯。实验采用漫反射方式。分析软件为 ASD View Spec Pro V5.0，Unscrambler V9.8 和 Matlab V2009。

1.2　样本采集与可溶性固形物含量测定

供试样品均产自太谷县有机栽培枣园，无外部缺陷，外表全部呈红色。2009 年采摘壶瓶枣样本 180 个，随机分为校正集 150 个和验证集 30 个样本。样品的光谱采集和可溶性固形物含量标准值测定当天完成。

将光谱仪置于枣样本的上方，距鲜枣表面 30 mm，对每一个样本在其赤道上选取两个部位来采集光谱，各扫描 30 次，以减少随机误差。对鲜枣样本进行光谱数据采集后，按 BS ISO 2173—2003《水果和蔬菜制品 可溶性固态物的测定 折射仪法》，逐一对样本进行可溶性固形物（SSC）含量的分析。为对应样本的光谱采集，在每个样本的赤道上测试两个部位。所测样本可溶性固形物含量的分布范围为 14.1 ~ 27.2°Brix，其均值为 21.23°Brix，标准差为 2.29。

1.3　光谱数据预处理

在原始光谱数据采集过程中常会受到高频随机噪声、光散射、样本不均匀等因素的影响，所以需要进行光谱预处理。经过对多种不同预处理方法的比较，本实验采用多元散射校正（MSC）方法进行壶瓶枣样本光谱数据的预处理。

1.4　光谱数据信息提取方法

采用连续投影算法和逐步回归方法进行光谱数据信息提取。连续投影算法能够从光谱信息中充分寻找含有最低限度冗余信息的变量组，使得变量之间的共线性达到最小。同时可大大减少建模所用变量的个数，提高建模的速度和效率。逐步回归法是多元线性回归法中选择回归变量的一种常用数学方法，它自动地从大量可供选择的变量中，选择那些与因变量相关性大的建立回归方程，达到压缩自变量的目的。

1.5　可溶性固形物近红外光谱分析建模方法

采用偏最小二乘（PLS）和最小二乘-支持向量机（LS－SVM）方法建模。偏最小二乘法是一种新型的多元统计数据分析方法。它是多因变量对多自变量的回归建模方法，在解决普通多元回归无法解决的问题上能达到良好的效果。最小二乘-支持向量机是一种新型的统计学习方法，它以结构风险最小化原则取代传统机器学习中的经验风险最小化原则，从而提高泛化能力。该算法通过非线性映射函数建立回归模型，将输入向量映射到高维空间，并在此高维空间内利用结构风险最小化原则构造最优决策函数。本文采用径向基函数（RBF）作为核函数，建立预测的 LS－SVM 模型。

2　实验结果与分析

2.1　壶瓶枣近红外光谱测定

180 个壶瓶枣的可见/近红外吸收光谱如图 1 所示，光谱曲线的趋势非常相似。波长范围为350 ~ 2 500 nm。

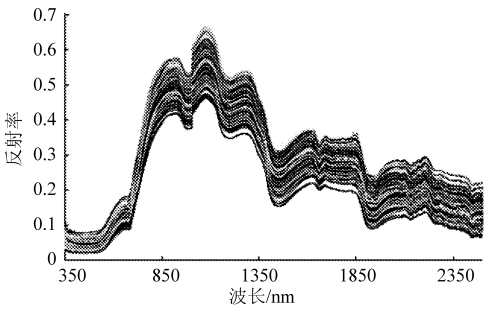

图 1 壶瓶枣近红外原始光谱

Fig. 1 Original Vis/NIR spectra of Huping jujube

2.2 光谱数据信息的定量提取

2.2.1 基于连续投影算法提取特征波长

利用连续投影算法选取特征波长。对经 MSC 预处理后的建模集 150 个样本的光谱数据进行有效波长提取。设定所提取的有效波长最大数为 30,在 Matlab V2009 中进行连续投影运算,得到 5 个特征波长,分别为 1 374、2 428、1 718、574 和 2 487 nm,其中不同波长点的顺序即为提取特征波长的重要性排序,即 1 374 nm 最重要,其次为 2 428 nm。

2.2.2 基于逐步回归法提取特征波长

利用逐步回归法选择壶瓶枣可溶性固形物的特征波长。基于光谱仪的条件,每个样本光谱有 2 151 个数据点,将全谱数据对各校正组样本的可溶性固形物含量用 SPSS 统计软件进行逐步回归分析,引入对可溶性固形物含量影响比较显著的波长点,剔除影响不显著的波长点,这样就可得到 32 条对可溶性固形物有显著影响的波长,即 1 686、2 486、2 240、1 374、2 428、549、1 372、963、841、478、481、1 310、2 280、1 689、2 314、2 423、387、437、2 454、2 223、1 679、2 137、1 827、443、2 339、2 333、2 335、2 392、2 356、2 352、2 395、2 311 nm。

2.2.3 提取的特征波长分析

SPA 和 SRA 两种方法提取的特征波长数量不同,从波长点来看,1 374 nm 和 1 718 nm 相同,其余波长点也分别有相近的分布。由图 1 可知,鲜枣可溶性固形物光谱图主要吸收峰位于 960、1 180、1 450和 2 000 nm,这些吸收峰均是由于水的吸收所导致。其中 2 000 nm 是 O—H 键的合频吸收峰,1 450 nm是 O—H 键一级倍频峰,960 nm 和 1 180 nm 是 O—H 键的二级倍频峰[1]。由于水的吸收很强,因此对其组成成分的检测干扰很大。在 SPA 提取的特征波长中没有明显的水吸收峰对应的波长点。在 SRA 提取的特征波长中有 963 nm 位于 960 nm 波长点附近,但这也是可溶性固形物分子中 C—H 键的三级倍频所在[1],因此没有去掉该特征波长点。鲜枣可溶性固形物分子中的 C—H、O—H 键的合频近红外谱带位于 2 000 ~ 2 500 nm[1],SPA 提取的特征波长点 2 428 nm 和 2 487 nm 都在这个范围内;SRA 提取的 32 个特征波长点中有 17 个都在这个范围内。

2.3 可溶性固形物含量近红外光谱建模与分析

对经过 MSC 预处理的 150 个样本光谱数据,建立全波段(350 ~ 2 500 nm)的 MSC - PLS 和 MSC - LS - SVM 模型;对经过连续投影算法提取的 5 个特征波长建立 MSC - SPA - PLS 和 MSC - SPA - LS - SVM 模型;对经过逐步回归提取的 32 个特征波长建立 MSC - SRA - PLS 和 MSC - SRA - LS - SVM 模型。采用建立的各模型分别对 30 个预测样本进行预测,结果见表 1。

表 1 不同模型预测结果比较

Tab. 1 Comparison of prediction results for different models

模型	预测相关系数	预测均方根误差
MSC - PLS	0.887 4	1.088 9
MSC - LS - SVM	0.675 6	1.188 5
MSC - SPA - PLS	0.799 0	1.407 8
MSC - SPA - LS - SVM	0.796 3	1.145 8
MSC - SRA - PLS	0.822 4	1.385 1
MSC - SRA - LS - SVM	0.592 8	2.691 2

由表 1 可见,对经预处理之后的 30 个预测集样本的全波段建模结果为:MSC - PLS 模型预测的相关系数和 RMSEP 分别为 0.887 4 和 1.088 9,预测效果如图 2 所示,预测样本均分布在回归线周围。MSC - PLS 模型的预测精度远好于 MSC - LS - SVM 模型(图 3),这个结果与众多的研究结果不同。

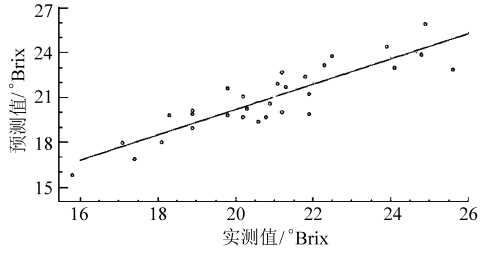

图 2 壶瓶枣可溶性固形物 MSC - PLS 模型预测结果

Fig. 2 Prediction result of soluble solids content of Huping jujube by MSC - PLS model

经过连续投影算法(SPA)提取特征波长之后,建立 MSC - SPA - PLS 和 MSC - SPA - LS - SVM 模型。MSC - SPA - PLS 模型预测的相关系数和 RMSEP 分别为 0.799 0 和 1.407 8,预测结果见图 4,与全波段的 MSC - PLS 模型相比,相关系数变小,精度降低;MSC - SPA - LS - SVM 模型预测的相关系数和均方根误差 RMSEP 分别为 0.796 3 和 1.145 8,

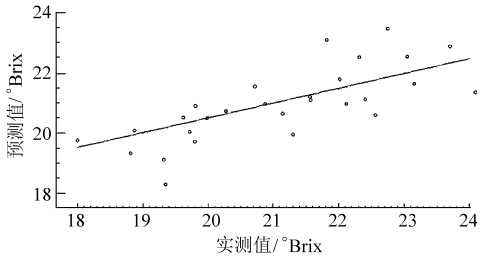

图 3　壶瓶枣可溶性固形物 MSC－LS－SVM 模型的预测结果

Fig. 3　Prediction result of soluble solids content of Huping jujube by MSC－LS－SVM model

预测结果见图 5，与全波段的 MSC－LS－SVM 相比，相关系数变大，精度提高。分析两种简化模型，其效果正好相反，不过对于两种建模方法来说，连续投影算法（SPA）均还可以用来简化建模，只是精度不高。

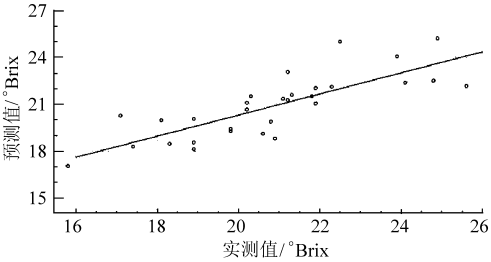

图 4　壶瓶枣可溶性固形物 MSC－SPA－PLS 模型的预测结果

Fig. 4　Prediction result of soluble solids content of Huping jujube by MSC－SPA－PLS model

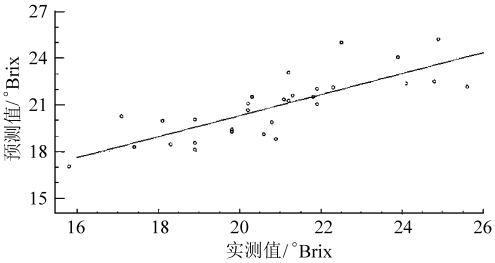

图 5　壶瓶枣可溶性固形物 MSC－SPA－LS－SVM 模型的预测结果

Fig. 5　Prediction result of soluble solids content of Huping jujube by MSC－SPA－LS－SVM model

对经过逐步回归法（SRA）提取的特征波长，建立 MSC－SRA－PLS 和 MSC－SRA－LS－SVM 模型。MSC－SRA－PLS 模型预测的相关系数和均方根误差（RMSEP）分别为 0.822 4 和 1.385 1，预测结果见图 6，与全波段 PLS 相比，相关系数变小，精度降低；MSC－SRA－LS－SVM 模型预测的相关系数和均方根误差 RMSEP 分别为 0.592 8 和 2.691 2，预测结果见图 7，与全波段的 MSC－LS－SVM 相比，相关系数大大减小，精度大大降低。分析两种简化模型可以看出，逐步回归方法适用于壶瓶枣可溶性固形物的 PLS 模型的简化建模，并不适用于 LS－SVM 模型的简化建模。

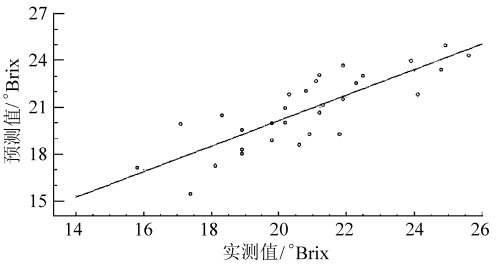

图 6　壶瓶枣可溶性固形物 MSC－SRA－PLS 模型的预测结果

Fig. 6　Prediction result of soluble solids content of Huping jujube by MSC－SRA－PLS model

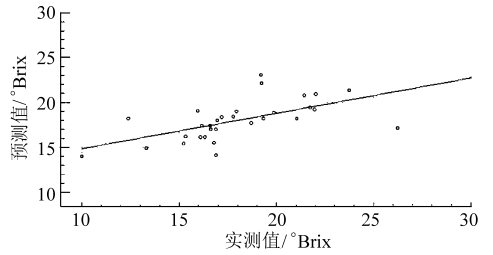

图 7　壶瓶枣可溶性固形物 MSC－SRA－LS－SVM 模型的预测结果

Fig. 7　Prediction result of soluble solids content of Huping jujube by MSC－SRA－LS－SVM model

本文中研究的各种建模方法，除了 MSC－SRA－LS－SVM 建模方法之外，与在主成分分析的基础上，采用 BP 神经网络建模方法预测鲜枣可溶性固形物含量均方根误差（RMSEP 为 1.990 3）[11] 比较，均大大减小。

3　结论

（1）壶瓶枣可溶性固形物含量的全波段 MSC－PLS 模型预测的相关系数和预测均方根误差（RMSEP）分别为 0.887 4 和 1.088 9，预测效果最好。

（2）采用连续投影算法（SPA）和逐步回归算法（SRA）提取特征波长，都不同程度损失了一些有效的线性信息，与 PLS 建模效果比较可以看出：MSC－SPA－PLS 模型预测的相关系数和 RMSEP 分别为 0.799 0 和 1.407 8，MSC－SRA－PLS 模型预测的相关系数和均方根误差（RMSEP）分别为 0.822 4 和 1.385 1，预测效果均不如全波段 PLS 建模；SPA 方法可以更有效地解决光谱数据中的非线性信息问题，这一点在连续投影算法（SPA）简化 LS－SVM 建模的效果中得到了验证：MSC－SPA－LS－SVM 模型预测的相关系数和均方根误差 RMSEP 分别为 0.796 3 和 1.145 8，与全波段 LS－SVM 模型相比，预测能力得以提高。

(3) 采用连续投影算法(SPA)和逐步回归算法(SRA)提取特征波长,再依据近红外光谱理论进行分析和确定,使提取的特征波长点更具有理论基础。

(4) 研究表明,对于同样的研究对象,不同的建模方法结果会有很大差别,因此要进行多种建模方法的比较。

参 考 文 献

1 陆婉珍. 现代近红外光谱分析技术[M]. 北京:中国石化出版社,2007:6~7.

2 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):162~165.

3 刘燕德,孙旭东,陈兴苗. 近红外漫反射光谱检测梨内部指标可溶性固形物的研究[J]. 光谱学与光谱分析,2008,28(4):797~800.
Liu Yande, Sun Yandong, Chen Xingmiao. Research on the soluble solids content of pear internal quality index by near infrared diffuse reflectance spectroscopy [J]. Spectroscopy and Spectral, 2008,28(4):797~800. (in Chinese)

4 李东华,纪淑娟,重滕和明. 南果梨糖、酸度近红外光谱模型适用的贮藏期研究[J]. 农业工程学报,2009,25(4):270~275.

5 Esmé D Louw, Karen I Theron. Robust prediction models for quality parameters in Japanese plums (*Prunus salicina* L.) using NIR spectroscopy [J]. Postharvest Biology and Technology, 2010,58:176~184.

6 黄康,汪辉君,徐惠荣,等. 基于最小二乘-支持向量机的番茄汁糖酸度分析研究[J]. 光谱学与光谱分析,2009,29(4):931~934.
Huang Kang, Wang Huijun, Xu Huirong, et al. NIR spectroscopy based on least square support vector machines for quality prediction of tomato juice [J]. Spectroscopy and Spectral, 2009,29 (4):931~934. (in Chinese)

7 齐小明,张录达,杜晓林,等. PLS-BP 法近红外光谱定量分析研究[J]. 光谱学与光谱分析,2003,23(5):870~872.
Qi Xiaoming, Zhang Luda, Du Xiaolin, et al. Quantitative analysis using NIR by building PLS-BP model [J]. Spectroscopy and Spectral Analysis,2003,23(5):870~872. (in Chinese)

8 吴迪,金春华,何勇. 基于连续投影算法的光谱主成分组合优化方法研究[J]. 光谱学与光谱分析,2009,29(10):2 734~2 737.
Wu Di, Jin Chunhua, He Yong. Study on combinatorial optimization of spectral principal components study on combinatorial optimization of spectral principal components using successive projections algorithm [J]. Spectroscopy and Spectral, 2009,29 (10): 2 734~2 737. (in Chinese)

9 黄凌霞,吴迪,金航峰,等. 基于变量选择的蚕茧茧层量可见-近红外光谱无损检测[J]. 农业工程学报,2010,26(2):231~236.
Huang Lingxia, Wu Di, Jin Hangfeng, et al. Non-destructive detection of cocoon shell weight based on variable selection by visible and near infrared spectroscopy [J]. Transactions of the CSAE, 2010, 26(2): 231~236. (in Chinese)

10 孙旭东,章海亮,欧阳爱国,等. 南丰蜜桔可溶性固形物近红外特征波段选择[J]. 农业机械学报,2009,40(7):139~142.
Sun Xudong, Zhang Hailiang, Ouyang Aiguo, et al. Selection of NIR characteristic wavelength bands for soluble selection of NIR characteristic wavelength bands for soluble [J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(7):129~132. (in Chinese)

11 张淑娟,王凤花,介邓飞,等. 鲜枣品种和可溶性固形物的近红外光谱无损检测[J]. 农业机械学报,2009,40(4):139~142.
Zhang Shujuan, Wang Fenghua, Jie Dengfei, et al. Non-destructive of the fresh jujube varieties and SSC by NIR spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(4): 139~142. (in Chinese)

12 Wang J, Nakano K, Ohashi S, et al. Comparison of different modes of visible and near-infrared spectroscopy for detecting internal insect infestation in jujubes [J]. Journal of Food Engineering,2010,101(1):78~84.

13 Zheng Hong, Lu Hongfei, Zheng Yueping, et al. Automatic sorting of Chinese jujube using chlorophyll fluorescence and support vector machine [J]. Journal of Food Engineering, 2010,101(4): 402~408.

14 徐惠荣,陈晓伟,应义斌. 基于多元校正法的香梨糖度可见/近红外光谱检测[J]. 农业机械学报,2010,41(12):126~129.
Xu Huirong, Chen Xiaowei, Ying Yibin. Multivariate approach to determinate sugar content of fragrant pears with temperature variation by visible/NIR spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2010,41(12):126~129. (in Chinese)

15 代芬,洪添胜,岳学军,等. 砂糖橘可溶性总糖可见-近红外光谱无损检测[J]. 农业机械学报,2011,42(4):133~138.
Dai Fen, Hong Tiansheng, Yue Xuejun, et al. Nondestructive examination of sugar content of intact 'Shatangju' with visible-near infrared spectroscopy based on variables selection [J]. Transactions of the Chinese Society for Agricultural Machinery, 2011,42(4):133~138. (in Chinese)