江西农业大学学报 2010,32(5):1075 – 1080 http://xuebao.jxau.edu.cn
Acta Agriculturae Universitatis Jiangxiensis E – mail: ndxb7775@sina.com

# 鸡蛋蛋白 pH 可见/近红外光谱在线检测信息变量提取研究

刘燕德,彭彦颖,孙旭东

(华东交通大学 机电工程学院,光机电技术及应用研究所,江西 南昌 330013)

**摘要:** 利用可见/近红外光谱在线检测鸡蛋品质中的蛋白 pH,采用漫反射方式进行光谱采集。采用反向区间偏最小二乘法(BiPLS)和蒙特卡罗无信息变量消除法(MC – UVE)分别优化鸡蛋蛋白 pH 可见/近红外光谱的信息区间组合及筛选有效建模变量数。经过最优预处理方法一阶导数对光谱进行预处理校正后,BiPLS 方法筛选的区间分隔最优数为 25,对应信息区间为 598.33 ~ 617.55 nm、636.63 ~ 655.58 nm、783.25 ~ 800.72 nm 和 852.24 ~ 885.82 nm。利用 MC – UVE 方法筛选出来的最佳建模变量数为 250 个,BiPLS 模型的 *Rp* 为 0.827 和 *RMSEP* 值为 0.094; MC – UVE – PLS 模型的 *Rp* 为 0.833 和 *RMSEP* 值为 0.086。结果表明利用蒙特卡罗无信息变量消除方法可以有效选择建模变量,既克服了复杂样品各信息区间对 PLS 建模贡献率不一样的问题,又能提高模型的稳定性和多元校正的预测精度。

**关键词:** 可见/近红外光谱; 在线检测; 蒙特卡罗无信息变量消除法; 蛋白 pH
**中图分类号:** TP75; TP274$^{+}$.5 **文献标志码:** A **文章编号:** 1000 – 2286(2010)05 – 1075 – 06

# A Study on Variable Selection of Vis – NIR Spectral Information for Online Detection Albumen pH of Eggs

LIU Yan-de, PENG Yan-ying, SUN Xu-dong

(Institute of Optics – Mechanics – Electronics Technology and Application (OMETA), East China Jiaotong University, Nanchang 330013, China)

**Abstract:** Method for albumen pH of eggs online detection by using Visible – NIR diffuse reflectance spectroscopy was studied. Backward interval partial least squares (BiPLS) and Monte Carlo Uninformative Variables Elimination (MC – UVE) were proposed to search for an optimized combination of information spectral intervals and number of variables about albumen pH from Vis – NIR spectra of egg. The spectra were preprocessed by first derivative which was the best spectrum preprocessing. It was found that the selected result was the best when the interval number was 25 used by BiPLS, and the information intervals were 598.33 ~ 617.55 nm, 636.63 ~ 655.58 nm, 783.25 ~ 800.72 nm and 852.24 ~ 885.82 nm. The optimized effective

**收稿日期:** 2010 – 08 – 10
**基金项目:** 国家科技支撑计划项目(2008BAD96B04)、江西省主要学科学术和技术带头人培养对象计划项目(2009DD00700)、江西省自然科学基金项目(2008GQN0029)和江西省对外科技合作计划项目(2009BHB15200)
**作者简介:** 刘燕德(1967 – ),女,江西泰和人,博士,教授,博士生导师。1990 年 6 月本科毕业江西农业大学工学院; 2006 年 6 月博士毕业于浙江大学农业机械化工程。现为华东交通大学首席教授,江西省主要学术学科带头人。主要从事光机电技术和信息化技术的基础理论与应用研究。先后主持或完成国家及省部级项目 16 项,申请或获得国家发明专利 8 项,完成科技项目鉴定 4 项。发表论文 80 篇,其中 SCI 收录 20 篇、EI 收录 25 篇,主参编著作和教材 3 部。E-mail: jxliuyd@163.com。

number of variables was 250 used by MC – UVE. For BiPLS and monte carlo uninformative variables elimination partial least squares (MC – UVE – PLS) models of the combination intervals and the effective number of variables, the *R* values of prediction set were 0.827 and 0.833, respectively. And the root mean square errors of prediction (RMSEP) were 0.094 and 0.086, respectively. The results reveal that the best results were obtained by MC – UVE – PLS. The proposed method overcomes the difficulties that different information intervals of complicated samples have different contribution to PLS model. And it makes the prediction more robust and accurate in quantitative analysis of albumen pH.

**Key words:** visible – near infrared spectrum; online detection; monte carlo uninformative variables elimination (MC – UVE); albumen pH

## 0 引 言

鸡蛋作为一种高质量食品,是人们日常饮食的重要组成部分。随着人们生活水平的提高及鸡蛋生产的规模的产业化,鸡蛋品质的提高和改良倍受关注,鸡蛋物理品质评价指标包括蛋重量、蛋白高度、蛋白质量、蛋壳强度、哈氏单位(HU)、蛋型指数、蛋壳颜色、气室高度、蛋比重等[1]。通常以蛋品质及新鲜度的测定结果作为鸡蛋分级的主要依据,传统的检测手段、精度、效率都不高。亟需建立一种新的能够应用于蛋品质量监督和市场分级的快速无损检测方法。

近红外光谱(NIRS)法以其快速、简便、无损等特点在农产品品质检测中占有重要地位[2-5],目前国内外已有一些利用近红外光谱技术检测鸡蛋内部品质及新鲜度的相关报道。吴瑞梅等运用紫外光谱等技术研究了鸡蛋白高度的变化[6]。侯卓成等利用傅里叶近红外反射技术对鸡蛋的蛋品质进行了检测研究[7]。刘燕德等利用可见近红外透射光谱检测鸡蛋新鲜度[8]。Kemps 等利用可见近红外透射光谱检测鸡蛋的蛋白品质问题[9]。Giunchi 等利用近红外反射光谱定性分析鸡蛋在不同储存时间下的新鲜度[10]。Nicolas 等利用可见近红外光谱预测鸡蛋的新鲜度和鸡蛋蛋白质量[11]。

由于近红外光谱产生于分子振动,吸收较弱,吸收峰严重重叠,且多组份复杂样品的近红外光谱往往不是各组份光谱的简单叠加。因此,近红外光谱分析法是一种间接分析技术,必须借助化学计量学方法才能进行定性或定量分析,偏最小二乘法(PLS)具有良好的处理光谱数据共线性能力,常用于解析近红外光谱信息。NIRS 分析过程中,常进行信息区间及变量筛选选择以简化模型和提高模型预测稳定性和精确度,如遗传算法、交互信息变量选择、移动窗口偏最小二乘法、反向区间偏最小二乘法(BiPLS)、无信息变量消除 PLS (UVE – PLS) 方法及其衍生方法等[12-15]。

本文以鸡蛋为研究对象,利用反向区间偏最小二乘(BiPLS)法和蒙特卡罗无信息变量消除(MC – UVE)法两种不同的模型简化方法对鸡蛋蛋白 pH 可见近红外光谱若干信息进行对比定位分析,探讨建模变量选择方法对鸡蛋蛋白 pH 指标预测模型的精度和稳定性的影响,选择一种更简便可靠的鸡蛋蛋白检测技术,实现鸡蛋蛋白 pH 值的快速无损分析。

## 1 材料与方法

### 1.1 样品

本试验中所采用的样品为褐色和白色两种不同蛋壳颜色的鸡蛋样品,总共 100 个样品,样品随机选购于江西省农贸市场及超市。将购买的鸡蛋表面的污染排泄物清理干净,依次在鸡蛋尖端头编号,并置于 18 ℃、76% 湿度的实验室内 2 h,待测样品温度达到室温后,采集在线光谱,光谱采集时尽量避免擦伤或伤疤等缺陷部位。共 100 个试验样品分为校正集和预测集,其中 65 个样品作为校正样品集,剩余 35 个未参与建模的样品作为预测样品集,用于评价模型的预测能力和稳定性。

### 1.2 光谱采集及设备

图 1 为可见近红外光谱在线检测装置示意图,光谱采集采用漫反射方式,利用光谱范围 350 ~ 1 100 nm 的可见近红外光谱仪(USB 2000 +, Oceanoptics INC, Florida, USA),1 个 100 W 高强度的卤素灯作为光源,光纤探头采集方向与入射光夹角为 45°。在测量水果光谱前先采集参比和暗电流光谱,以聚四氟乙烯材质的白板(6.5 mm 厚度)为标准参比,每个鸡蛋水平放置在移动托盘上,波长采集范围为 500

~950 nm。样品随机放置在均匀成单列的输送台上,每个样品重复采集 3 次光谱,最终取 3 次的平均光谱作为每个样品的最终光谱,样品、参比和暗电流的积分时间均为 25 ms,扫描次数为 10 次,且光谱仪参数设置、数据收集和存储均利用 Spectrasuite( Ocean optics INC, USA) 软件。本试验获取的光谱以漫反射光谱比值表示,转换公式如下:

$$R_\lambda = \frac{R_s - R_d}{R_e - R_d} \quad (1)$$

$R_\lambda$ 为波长 $\lambda$ 下样品的漫反射比值; $R_s$ 为波长 $\lambda$ 下样品光谱的强度; $R_e$ 为波长 $\lambda$ 下参比光谱的强度; $R_d$ 为波长 $\lambda$ 下暗电流光谱的强度。

100 个鸡蛋样品对应的可见/近红外漫反射光谱如图 2 所示。

### 1.3　蛋白 pH

对光谱采集完毕的鸡蛋样品进行破坏性试验,将鸡蛋样品的蛋白和蛋黄进行分离。将分离出来的蛋白放置于玻璃器皿中,使用 pH 温度仪( model: testo205, Testo AG, Lenzkirch Germany),将探头至于玻璃器皿中对蛋白 pH 值进行测定。检测之前,先要利用 pH 缓冲液对仪器进行 pH 值校正,本试验中的缓冲液采用 pH 4 和 pH 7。

### 1.4　数据处理数据处理及模型评价

反向区间偏最小二乘法工具包由 NΦrgaard 等提供的网络共享获得。MC - UVE - PLS 工具箱,在 MATLAB10.0 中实现。光谱预处理方法在 unscrambler8.0( CAMO, Trondheim, Norway) 中实现。由建模集及预测集相关系数( $R_c$ 和 $R_p$) 、建模均方根误差( *RMSEC*) 、交叉验证均方根误差( *RMSECV*) 、预测均方根误差( *RMSEP*) 进行评价。

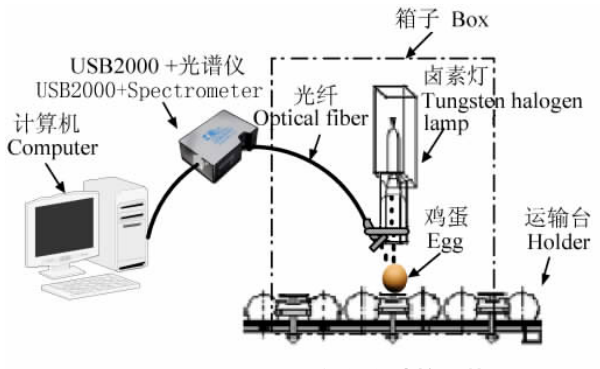


图 1　可见/近红外在线检测装置

Fig. 1　Schematic diagram of visible - NIR diffuse reflectance spectra collection

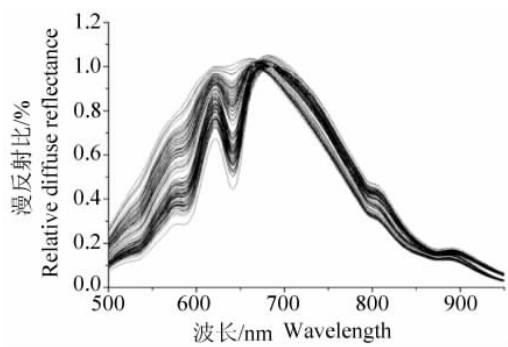


图 2　鸡蛋可见/近红外漫反射原始光谱

Fig. 2　Original Vis - NIR spectra of eggs

## 2　结论与讨论

### 2.1　测量值正态分布

由表 1 可知,样品的蛋白 pH 含量为 8.99 ~ 9.68,且预测集样本范围都处在校正集样本之内,可见样品的校正集所建立的模型能较好地使用于预测集。

表 1　校正集和预测集样品对应蛋白 pH 值的标准值

Tab. 1　Statistics of the albumen pH value in calibration and prediction data sets

| 数据集 Data set | 数量 Number of samples | 范围 Range | 平均值 Mean | 标准差 Standard deviation | 变异系数/% Coefficient of variation |
|---|---|---|---|---|---|
| 校正集 Calibration set | 65 | 8.99 ~ 9.68 | 9.40 | 0.17 | 1.81 |
| 预测集 Prediction set | 35 | 9.02 ~ 9.60 | 9.37 | 0.16 | 1.70 |

### 2.2　光谱数据预处理

光谱实际测量过程中不仅包括了鸡蛋品质相关的各种信息,还包括仪器噪声、外界干扰等,需要对原始光谱进行适当的预处理,将鸡蛋品质信号和噪声分离,最大程度地提取鸡蛋品质的有效信息。本试验分别运用了平滑、多元散射校正( MSC) 、导数处理等手段,对不同的光谱数据形式和数据平滑方式以

表 2 不同预处理方法下对应蛋白 pH 值处理结果

Tab. 2 Results of different prediction with different preprocessing methods

| 测量指标 Eggs quality | 预处理方法 Spectrum pretreatment | 因子数 *LVs* | 校正集 Calibration set | | 预测集 Prediction set | |
|---|---|---|---|---|---|---|
| | | | $R_c$ | *RMSEC* | $R_p$ | *MSEP* |
| 蛋白 pH | 原始光谱 | 8 | 0.795 | 0.101 | 0.781 | 0.103 |
| | 平滑(5 点) | 8 | 0.804 | 0.102 | 0.794 | 0.105 |
| Albument pH | MSC | 3 | 0.803 | 0.102 | 0.793 | 0.106 |
| | 一阶导数(17 点) | 3 | 0.821 | 0.096 | 0.797 | 0.101 |
| | 二阶导数(25 点) | 3 | 0.763 | 0.106 | 0.640 | 0.115 |

及每种数据平滑方式相应的最佳参数组合进行了选择,用每次剔除一个光谱点的方法进行交互验证,然后分析光谱异常点和杠杆系数,进一步优化模型(表 2)。最佳预处理方法处理后的光谱图如图 3 所示。

### 2.3 BiPLS 反向间隔偏最小二乘法选择信息区间

BiPLS 信息变量区间选择方法首先将整个光谱分割成 $k$ 个等宽子区间,然后在每个区间进行最小二乘回归。采用留一交互验证法计算交互验证均方根误差(*RMSECV*),依次减少信息量最差或共线性变量最多的 $i$($i=0,1,2,\cdots k$)个区间,即去除 *RMSECV* 值最大的区间,在剩余的($k-1$)个区间上建立最优 PLS 模型,并给出相应的 *RMSECV* 值。当 *RMSECV* 最小时所对应的多个区间即为所优化的组合区间,且对应的因子数为最佳因子数。

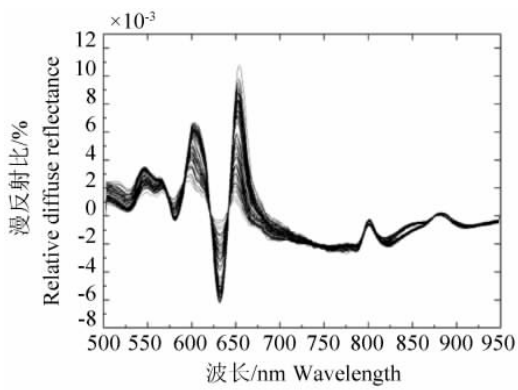


图 3 经一阶导数处理后的光谱图

Fig. 3 Spectra after preprocessed by first derivative

可见近红外光谱既表征目标信息,同时也受到非目标信息或仪器噪声干扰,而且目标信息区间分布复杂,信息区间宽度不等。因此,在运行 BiPLS 选择信息区间时,应该考察区间分割数对选择结果及模型的影响。在 PLS 计算中,最大因子数应该不大于所包含的变量数,将整条光谱(1 337 个数据点)分为 15、20、25 个子区间,最大因子数设为 8。

表 3 不同区间数下的 BiPLS 优化结果

Tab. 3 Optimization result of BiPLS for different numbers of interval

| 区间数 Number of intervals | 所选区间 Selected intervals | 光谱范围/nm Spectral range | *RMSECV* | 变量数 Number of variables |
|---|---|---|---|---|
| 15 | [13, 5, 4, 10] | 597.98 ~ 629.58, 629.58 ~ 660.82, 781.93 ~ 811.2, 868.47 ~ 896.46 | 0.093 | 356 |
| 20 | [2,16,17,5,13] | 524.94 ~ 549.36, 597.62 ~ 621.45,782.26 ~ 804.33, 847.76 ~ 869.11,869.11 ~ 890.21 | 0.101 | 335 |
| 25 | [20,8,21,6,16] | 598.33 ~ 617.55, 636.63 ~ 655.58, 783.25 ~ 800.72, 852.24 ~ 885.82 | 0.091 | 321 |

由最优的 BiPLS 运行结果(表 3)可知,当模型具有最小的 *RMSECV* 值时,对应的子区间分隔数为 25,此时所选择的波长范围是 598.33 ~ 617.55 nm、636.63 ~ 655.58 nm、783.25 ~ 800.72 nm、852.24 ~ 885.82 nm,数据点数为 321 个。BiPLS 是一种区间定位法,可以对近红外光谱信息区间进行初步定位,剔除信息量较差的区间。

### 2.4　基于 MC – UVE PLS 方法模型的变量选择

蒙特卡罗无信息变量消除 PLS( MC – UVE – PLS) 方法是基于 PLS 回归系数提出的一种波长选择方法。该方法通过一定的变量筛选标准,利用变量的稳定性值来评价模型中每个变量的可靠性,从而决定每个变量的取舍。计算预测验证均方根误差( *RMSEP*) ,当 *RMSEP* 值最小时所对应的变量数为最佳变量数,采用最佳的变量数替代整体光谱变量数建立 PLS 模型。

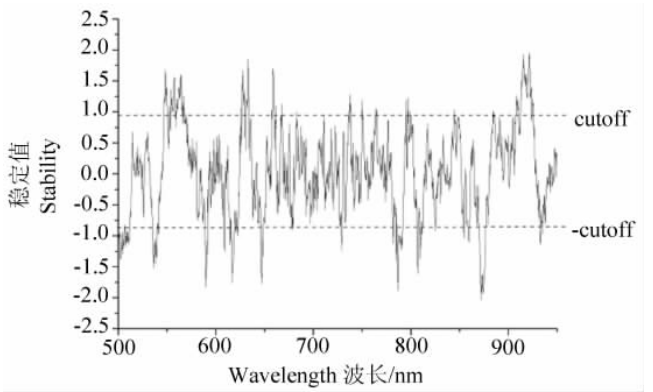


图 4　光谱经 MC – UVE 处理后建模变量稳定性分布

Fig. 4　The stability distribution of each variable for prediction by MC – UVE method

图 4 给出了波长 500 ~ 950 nm 范围内,鸡蛋蛋白 pH 通过 MC – UVE 方法得到的每个变量的稳定性值。图 4 中,虚线表示变量筛选的阈值,位于 2 个虚线之间的稳定性值所对应的变量将被舍弃,保留虚线之外的变量用于 PLSR 模型的构建。

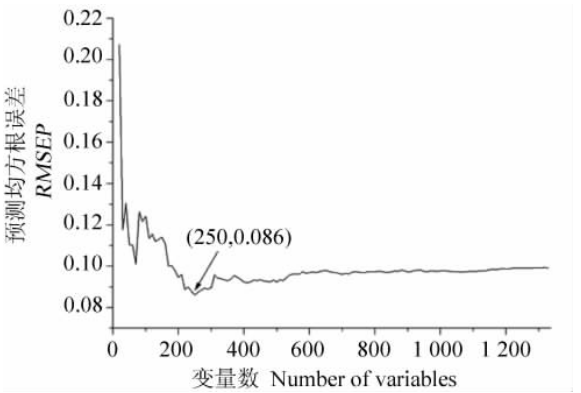


图 5　测试集的 *RMSEP* 随着保留变量数目的变化分布

Fig. 5　Variation of *RMSEP* with the number of selected wavelengths

图 5 给出了变量数目从 20 ~ 1 337 之间每隔 10 个数目所得的 *RMSEP* 值,并显示了测试集 *RMSEP* 值随保留变量数目变化的情况。采用每一组保留的变量建立一个 PLS 模型,用来计算测试集的 *RMSEP*。这是因为在 MC – UVE 方法中,保留变量的数目决定着模型的预测稳定性和精确度。如果保留的变量个数过少,可能会造成有用信息变量的丢失; 相反,如果保留的变量个数过多,多余的无用信息变量会使得模型质量变差。从图 5 可以看出,最初的数值很大,随着保留变量数目的增加,数值均急剧降低; 当保留变量数目为 250 时,*RMSEP* 的值最小( 为 0. 086) ; 当保留变量数目继续增多时,*RMSEP* 的值逐渐增大。这表明当保留较少变量时,有用信息变量不能全部被模型所包含; 而当选用过多变量时,无用的变量也会影响预测的结果。

### 2.5　PLS 模型预测结果

经过一阶导数处理后,在采用 BipLS 所选择的组合区间和 MC – UVE 方法所选取出的变量数,分别建立 PLS 模型并进行预测。从表 4 可知,最佳建模方法 MC – UVE – PLS 的预测集 *R* 值为 0. 833,*RMSEP* 值为 0. 086。对应相关系数如图 6 所示。

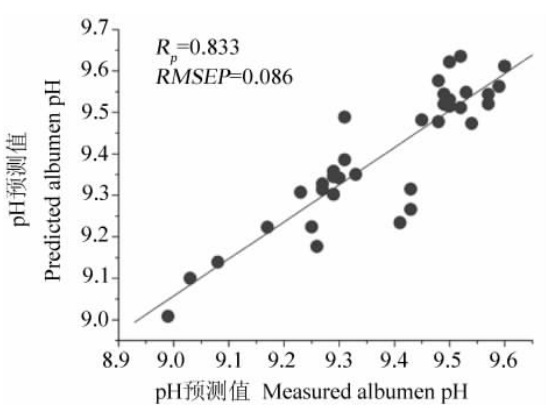


图 6　建模下预测集的预测值与实际值的相关关系

Fig. 6　Correlation of predicted values and actual values of pear in prediction set

## 3　结　论

采用可见/近红外漫反射技术在 500 ~ 950 nm 波段内对鸡蛋的蛋白 pH 进行了在线检测研究,通过

表4 PLS，BiPLS 和 MC－UVE－PLS 3 种方法建模结果比较

Tab. 4 Comparison of the results obtained by PLS，BiPLS and MC－UVE－PLS

| 评价模型 Models | 变量数目 Number of variables | $R_p$ | *RMSEP* |
|---|---|---|---|
| PLS | 1 337 | 0.797 | 0.101 |
| BiPLS | 321 | 0.827 | 0.094 |
| MC－UVE－PLS | 250 | 0.833 | 0.086 |

反向区间偏最小二乘法(BiPLS)和蒙特卡罗无信息变量消除法(MC－UVE)对鸡蛋蛋白 pH 可见/近红外光谱若干信息进行定位分析，分别选出了最优优化的组合区间下对应变量数及有效变量数；利用所选变量数进行 PLS 模型建立，克服了复杂样品各信息区间对 PLS 建模贡献率不同而影响预测模型的精度和稳定性的问题。2 种方法所建立模型的对比结果表明：利用 MC－UVE PLS 模型获得了最佳预测精度，其 $R_p$ 值为 0.833，*RMSEP* 值为 0.086。即利用可见近红外漫反射光谱快速检测蛋白 pH 的方法是可行的，同时应用蒙特卡罗无信息变量消除法减小信息维数，提高预测模型精度和稳定性，具有良好的应用前景。

## 参考文献：


[1] North M O，Bell D. Commercial Chichken Production Manual(4rd Edition) [M]. New York：Van Nostrand Reinhold，1990.

[2] Flores－Rojasa K，Sáncheza M T，Pérez－Marínb D，et al. Quantitative assessment of intact green asparagus quality by near infrared spectroscopy [J]. Postharvest Biology and Technology，2009，52：300－306.

[3] Liu Y D，Sun X D，Zhang H L，et al. Nondestructive measurement of internal quality of Nanfeng mandarin fruit by charge coupled device near infrared spectroscopy [J]. Computers and Electronics in Agriculture，2010，71(S1)，S10－S14.

[4] 吴桂芳，何勇. 应用可见/近红外光谱进行纺织纤维鉴别的研究 [J]. 光谱学与光谱分析，2010，30(2)：331－335.

[5] 马本学，饶秀勤，应义斌，等. 基于近红外漫反射光谱的香梨类别定性分析 [J]. 光谱学与光谱分析，2009，29(12)：3288－3290.

[6] 吴瑞梅，严霖元，乔振先. 不同品种鸡蛋新鲜度与其光特性的相关关系 [J]. 江西农业大学学报，2004，24(5)：781－784.

[7] 侯卓成，杨宁，李俊英，等. 傅里叶变换近红外反射用于鸡蛋蛋品质的研究 [J]. 光谱学与光谱分析，2009，29(8)：2063－2068.

[8] Liu Y D，Ying Y B，Ouyang A G，et al. Measurement of internal quality in chicken eggs using visible transmittance spectroscopy technology. Food Control，2007，18，18－22.

[9] Kemps B J，Ketelaere B D，Bamelis F R，et al. Albumen freshness assessment by combining visible near－infrared transmission and low－resolution proton nuclear magnetic resonance spectroscopy [J]. Poultry Science，2007，86：752－759.

[10] Giunchi A，Berardinelli A，Ragni L，et al. Non－destructive freshness assessment of shell eggs using FT－NIR spectroscopy [J]. Journal of Food Engineering，2008，89：142－148.

[11] Nicolas A，Michael N，Shiv P，et al. Prediction of Egg Freshness and Albumen Quality Using Visible/Near Infrared Spectroscopy [J]. Food and Bioprocess Technology. 2009，10：265－274.

[12] 鸿明坚，温志渝. 一种多模型融合的近红外波长选择算法 [J]. 光谱学与光谱分析，2010，30(8)：2088－2092.

[13] 林颢，赵杰文，陈全胜，等. 近红外光谱结合一类支持向量机算法检测鸡蛋的新鲜度 [J]. 光谱学与光谱分析，2010，30(4)：929－932.

[14] 王加华，李鹏飞，曹楠宁，等. 基于 iPLS 原理最优化信息区间的桃糖度组合权重 PLS 模型研究 [J]. 红外与毫米波学报，2009，28(5)：386－391.

[15] 李鹏飞，王加华，曹楠宁，等. BiPLS 结合 GA 优选可见/近红外光谱 MLR 变量 [J]. 光谱学与光谱分析，2009，29(10)：2637－2641.