文章编号 1004-924X(2009)08-1839-06

# 特征谱区筛选在近红外光谱检测茶叶游离氨基酸含量中的应用

郭志明,赵杰文,陈全胜,黄星奕
(江苏大学 食品与生物工程学院,江苏 镇江 212013)

**摘要**:为了提高近红外光谱快速检测茶叶游离氨基酸含量的精度和稳定性,利用特征谱区结合偏最小二乘法建立了预测模型。研究了利用联合区间偏最小二乘法和遗传偏最小二乘法等筛选特征谱区的方法,通过交互验证法确定偏最小二乘模型的主成分因子数和筛选区间,并以预测均方根误差(RMSEP)和相关系数 $R$ 作为模型的评价指标。实验结果表明:两种方法建立模型的预测能力都好于传统 PLS 模型;利用联合区间偏最小二乘法建立的预测模型最佳,预测的相关系数 $R$ 和 RMSEP 分别为 0.954 2 和 0.256 0。研究结果表明,近红外光谱结合特征谱区筛选方法可以快速准确地测定茶叶中游离氨基酸含量。

**关 键 词**:近红外光谱法;特征谱区筛选;偏最小二乘法;茶叶
**中图分类号**:O657.33 **文献标识码**:A

# Application of selecting wavelength regions to determination of free amino acid content in tea by FT-NIR spectroscopy

GUO Zhi-ming, ZHAO Jie-wen, CHEN Quan-sheng, HUANG Xing-yi

(*School of Food and Biological Engineering, Jiangsu University, Zhenjiang* 212013, *China*)

**Abstract**: In order to improve the detecting precision and robustness in determination of the content of free amino acid in tea by the FT-NIR spectroscopy, a predictive model was established by selecting efficient spectral regions combined with the Partial Least-squares(PLS) algorithm. In this research, the synergy interval Partial Least-squares (si-PLS) and the Genetic Algorithm Partial Least-squares (GA-PLS) were applied to select the efficient spectral regions. The number of PLS components and the number of intervals were optimized according to the cross-validation in a calibration set. The performance of the final model was evaluated according to the Root Mean Square Error of Prediction (RMSEP) and the Correlation coefficient ($R$) in prediction and calibration sets. Experimental results show that the two methods are able to produce better prediction models in relation to the full-spectrum model, and the performance of siPLS model is better than that of the GA-PLS. The optimal model offers its $R$ in 0.954 2 and RMSEP in 0.256 0 respectively by a prediction set. This study demonstrates that

**收稿日期**:2008-09-24;**修订日期**:2008-10-23.
**基金项目**:国家自然科学基金资助项目(No. 30800666);江苏省自然科学基金重大资助项目(No. BK2007087;No. BK2009216)

NIR spectroscopy combined with efficient selection wavelength region algorithm can be used successfully to analyze the content of free amino acid in tea.

**Key words**: near infrared spectroscopy; selection wavelength regions; partial least-squares; tea

## 1 引　言

随着近红外光谱技术和化学计量方法的发展,近红外光谱技术已经在国民经济发展的各个领域得到应用[1-3],在茶叶品质检测方面也逐渐得到人们的重视,尤其在实现茶叶品质快速检测和预测茶叶有效成分等方面取得了良好的效果[4-7],具有十分广阔的应用前景。借助先进的近红外光谱仪,研究者可以在短时间内很方便地获得大量光谱数据。但由于仪器所采集的数据除样品的自身信息外,还包含了其它无关信息和噪音,如电噪音、样品背景等,这些信息很难在预处理中全部消除;其次有些区域样品的信息很弱,与样品的组成或性质缺乏相关关系。如果将这些数据都参与建模,不但计算量大、模型复杂,而且精度也受到影响。因此,通过特定方法对自变量进行优选来简化模型,并通过剔除不相关或非线性变量,得到预测能力强、稳健性好的校正模型很有必要。

在测定茶叶中某类有效成分含量时,如游离氨基酸,由于不是单一组分,其特征谱区或特征波长很难选择,所以确定茶叶中游离氨基酸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,茶叶近红外光谱的总体走势比较平缓,波峰和波谷没有剧烈的起伏,因此,通过一些强有力的数据挖掘方法从这些烦杂的数据中提取有用的信息,筛选特征光谱区域和特征波长是一个值得研究的问题。

在近红外光谱结合偏最小二乘法(Partial Least Squares, PLS)方法建模中,传统观点认为PLS具有较强的抗干扰能力,可全波长参与多元校正模型的建立。但对PLS方法的深入研究和应用表明,通过特定方法筛选特征波长或波长区间有可能得到更好的定量校正模型。鉴于此,本文研究了利用联合区间偏最小二乘和遗传偏最小二乘等算法筛选茶叶近红外光谱中的特征谱区或特征波长,然后结合偏最小二乘建立游离氨基酸定量分析模型来提高近红外光谱预测模型的精度和稳定性的方法。

## 2 近红外光谱特征谱区筛选方法

研究了利用特征谱区筛选结合偏最小二乘法检测茶叶中游离氨基酸含量的方法。用交互验证法确定主成分数,以含量的实测值与预测值的相关系数$R$和交互验证均方根误差(Root Mean Square Error of Cross-validation, RMSECV)及预测均方根误差(root mean squared error of prediction, RMSEP)作为评价各种方法的有效指标。数据分析在Matlab V7.0 (Mathworks, Natick, USA)的软件平台上完成。

### 2.1 联合区间偏最小二乘法

区间偏最小二乘波长筛选法是由Lars Nφrgaard[8-9]于2000年提出的一种波长筛选法,该法主要用于筛选偏最小二乘建模的波长区域。联合区间偏最小二乘法(Synergy interval PLS, siPLS)是建立在常规区间偏最小二乘法的基础上的一种方法[10],它将同一次区间划分中精度较高的几个局部模型所在的子区间联合起来,共同预测农产品品质指标。但目前尚不能从理论上确定参加联合建模的子区间数目。由于茶叶是一种非常复杂的天然生物体,其内部品质指标并不是以某种纯的化学成分存在,单独的一个小区间不能提供足够的信息来预测茶叶的内部品质指标。联合几个子区间建立的预测模型,测量精度会更高。

### 2.2 遗传偏最小二乘法

遗传偏最小二乘法(Genetic Algorithm-partial Least Squares, GA-PLS)利用遗传算法全局快速搜索的优点,将遗传算法和PLS方法有机地结合起来,发挥各自的长处,可建立更加稳定、预测能力更强的模型[11-12]。其基本思想是将PLS交互验证中因变量的预测值和实际值间相关系数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PLS方法对筛选后的波长变量建立分析校正模型。遗传偏最小二乘法包括遗传编码、适用度函数设计、产生初始群体、选择、交叉、变异等。研究中,以遗传迭代次数达到设定的RMSECV值为收敛终止条

件，迭代终止后将被选用频数最多的前 150 个波数点按频数高低逐一顺序加入 PLS 模型中，以最小的交互验证均方根误差 RMSECV 值确定出最佳的建模变量。

## 3　实验方法及数据

试验选用市售茶叶 18 种，分别来自江苏、浙江、福建、云南、安徽、江西、河南、四川、湖南和广东，均产于 07 年 4 月到 5 月。从中随机地选取 12 种作为校正集，余下的 6 种作为预测集，选取每种茶叶 5 个样本，试验前先将茶叶粉碎并过 40 目筛，每个样本称取 1 g，这样校正集共 60 个样本，预测集共 35 个样本，然后将它们分别编号后置于 4 ℃冰柜中贮藏。光谱采集试验在温度(25 ℃)稳定的实验室内进行。试验前，将冰柜中取出的茶叶置于实验室中 12 h，以使茶叶整体温度与环境温度一致。试验时，将茶叶样本放于样品杯中，由近红外光谱仪(Antaris Ⅱ，美国 Thermo Scientific 公司)进行光谱扫描(扫描波数范围为 9 999.10～3 999.64 $cm^{-1}$，扫描次数 32 次，分辨率 4 $cm^{-1}$)，然后将样本倒出样品杯再重新倒入，重复采集 4 次，并取 4 次的平均光谱作为该样本的原始光谱。采集光谱后，按照 GB/T 8314-2002 来测定茶叶样本中游离氨基酸的含量。测定结果如表 1 所示。

**表 1　茶叶样本氨基酸含量实测值的统计结果**

Tab.1　Statistics of amino acid contents for calibration and prediction sets of tea

| 成分 | 样本数 | 范围% | 均值% | 方差 |
|---|---|---|---|---|
| 校正集 | 60 | 0.976 2～5.126 8 | 2.988 8 | 0.808 9 |
| 预测集 | 30 | 1.236 3～4.898 9 | 2.986 0 | 0.747 2 |

## 4　实验结果与讨论

### 4.1　联合区间偏最小二乘模型

应用联合区间偏最小二乘法对茶叶的近红外光谱进行筛选时，模型的主成分因子取为 10；将整个光谱区域分别划分为 10、11、12、…、25 个子区间，以考查不同数目的子区间划分对模型性能以及最佳波长区间的影响。在数据处理过程中，划分为相同子区间的情况下，又尝试分别联合 2 个、3 个和 4 个子区间。表 2 为用联合区间偏最小二乘模型分析茶叶中氨基酸含量的结果。

**表 2　选择不同子区间数的联合区间偏最小二乘分析模型的结果**

Tab.2　Results of siPLS calibration model with different spectral regions

| 区间划分数 | 被选区间 | 主因子数 | 交互验证均方根误差 | 校正集相关系数 |
|---|---|---|---|---|
| 10 | [1, 7, 8, 10] | 9 | 0.279 4 | 0.950 9 |
| 11 | [4, 8] | 8 | 0.250 2 | 0.960 6 |
| 12 | [4, 5, 8] | 10 | 0.260 8 | 0.957 2 |
| 13 | [1, 4 | ]9 | 0.260 2 | 0.957 9 |
| 14 | [1, 8, 12] | 9 | 0.263 2 | 0.956 4 |
| 15 | [5, 6, 13,14] | 9 | 0.252 1 | 0.960 0 |
| 16 | [5, 6, 7, 14] | 10 | 0.259 6 | 0.957 6 |
| 17 | [1, 9, 13, 15] | 9 | 0.265 6 | 0.955 5 |
| 18 | [6, 7, 13] | 8 | 0.241 5 | 0.963 3 |
| 19 | [6, 7, 8, 9] | 9 | 0.248 5 | 0.961 2 |
| 20 | [6, 8, 16, 17] | 9 | 0.255 5 | 0.958 9 |
| 21 | [7, 8, 9, 17] | 10 | 0.259 7 | 0.957 6 |
| 22 | [7, 8, 9, 18] | 8 | 0.232 2 | 0.966 1 |
| 23 | [1, 4] | 10 | 0.260 9 | 0.957 4 |
| 24 | [7, 8, 9, 22] | 10 | 0.256 9 | 0.958 5 |
| 25 | [8, 9, 11, 23] | 8 | 0.246 9 | 0.961 7 |

表 2 是把光谱区间划分为 22 个并联合 4 个子区间数时获得最优氨基酸含量的 siPLS 模型，此时选取的子区间为 7、8、9 和 18，共 567 个变量，如图 1 所示。该模型采纳的主成分因子数为 8 个，其校正集的相关系数 $R_c$ 和交互验证均方根误差 RMSECV 分别为 0.966 1 和 0.232 2，预测集的相关系数$R_p$和预测均方根误差RMSEP分别为 0.954 2 和 0.256 0。与其它模型相比，该 siPLS 模型精度最高，预测性能最佳。而建立的传统偏最小二乘模型预测茶叶中氨基酸含量的精度不高，校正集相关系数 $R_c$ 为 0.927 0，预测集相关系数 $R_p$ 为 0.924 8，模型的主因子数为 10 个。与传统偏最小二乘模型相比，联合区间偏最小二乘法不仅能有效地减少建模所用的变量数，而且能有效提高茶叶氨基酸含量模型的测量精度。将精度较高的几个局部模型所在的子区间联合起来建立一个茶叶氨基酸含量的预测模型是有效的。分析其原因，可能是几个精度较高的局部模型与茶叶品质的相关性较高，含有较少的噪声信息，可以充分表征茶叶的某一品质特征。图 2 为该模型校正集样本和预测集样本的预测值与实测值之间的关系图，可以看出预测值与实测值有很好的相关性。

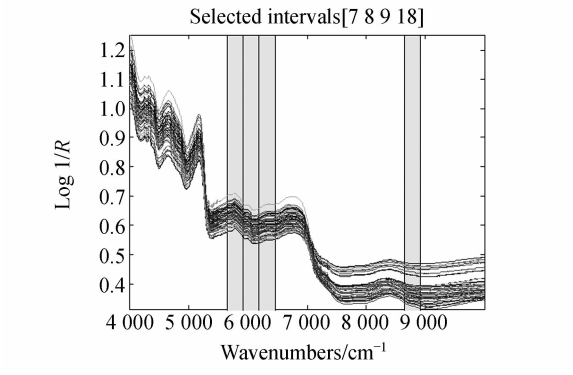


图 1　联合区间偏最小二乘模型选择的最佳子区间[7 8 9 18]

Fig. 1　Optimal spectral region selected by siPLS with intervals 7, 8, 9 and 18

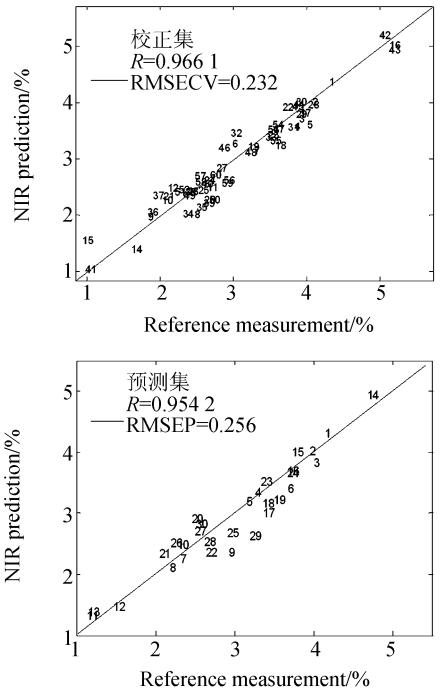


图 2　氨基酸含量 siPLS 最佳模型的校正集和预测集样本的预测值与实测值之间的关系

Fig. 2　Reference measured versus NIR predicted by siPLS in prediction and calibration sets

## 4.2　遗传偏最小二乘模型

以全光谱范围内的 3 112 个波数点作为筛选对象。遗传算法的控制参数设定为：初始群体大小为 50，交叉概率 $P_c=0.5$，变异概率 $P_m=0.01$，遗传迭代次数为 100，迭代终止后将被选用频数最多的前 150 个波数点按频数高低逐一顺序加入 PLS 模型中，以最小的交互验证均方根误差 RMSECV 值确定出最佳的建模波数点数。图 3 显示了其中一次运算过程中各波数点被选用的频次，从该图中可看出，被选用频次较多的波数点主要集中在 4 232.985 $cm^{-1}$、4 126.919 $cm^{-1}$、5 702.477 $cm^{-1}$以及 4 323.623 $cm^{-1}$等几个波数点附近，特别是 4 232.985 $cm^{-1}$左右的几个波数点被选用的频次最高，这说明，这几个波数点与茶叶中氨基酸含量具有高度的相关性。

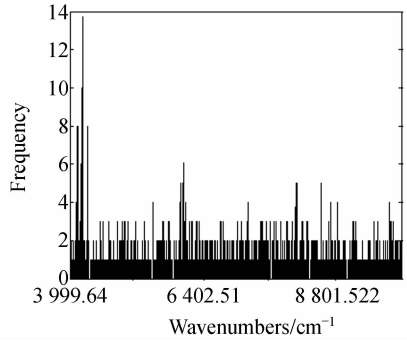


图 3　各波数点被选用的频次图

Fig. 3　Histogram of selected times of each variable

**表 3　各次遗传偏最小二乘回归建模的结果**

Tab. 3　Results of GA-PLS model for prediction of amino acid in tea

| 模型序号 | 波长点数 | 主成分因子数 | 交互验证均方根误差 RMSECV | 校正集相关系数 $R_c$ | 预测均方根误差 RMSEP | 预测集相关系数 $R_p$ |
|---|---|---|---|---|---|---|
| 1 | 48 | 8 | 0.246 9 | 0.961 7 | 0.279 6 | 0.947 8 |
| 2 | 54 | 8 | 0.245 3 | 0.962 2 | 0.277 9 | 0.948 5 |
| 3 | 46 | 8 | 0.248 2 | 0.961 3 | 0.280 5 | 0.947 4 |
| 4 | 51 | 8 | 0.241 6 | 0.963 4 | 0.274 6 | 0.949 8 |
| 5 | 43 | 6 | 0.243 7 | 0.962 8 | 0.277 2 | 0.948 9 |
| 6 | 60 | 8 | 0.246 5 | 0.961 8 | 0.278 1 | 0.948 4 |

表 3 为各次遗传偏最小二乘建模的结果，综合考虑模型对校正集及预测集的预测能力，可看出第 4 号模型为最佳模型，其交互验证均方根误差 RMSECV 为 0.241 6，校正集相关系数 $R_c$ 为 0.963 4，预测均方根误差 RMSEP 为 0.274 6，预测集相关系数 $R_p$ 为 0.949 8，该模型采用的波数点由 3 112 减少到了 51，主因子数由 10 个减少到 8 个，在保证精度的前提下大大简化了模型。GA-PLS 模型的结果稍差于 siPLS 模型的结果，但采用的波数点更少，预测性能明显优于全光谱模型。由于遗传算法是一种随机搜索算法，其初始群体的选取以及遗传操作算子的执行过程都带有较强的随机性，可能会影响模型预测的稳定性。

## 5　结　论

用联合区间偏最小二乘和遗传偏最小二乘法对茶叶近红外光谱进行了特征谱区或特征波长的筛选。与传统 PLS 模型相比，联合区间偏最小二乘和遗传偏最小二乘法都能有效地减少建模所用的变量数，并且都能有效提高茶叶氨基酸含量模型的预测精度。其中，联合区间偏最小二乘法模型优于遗传偏最小二乘模型，联合区间偏最小二乘法模型的预测集相关系数 $R_p$ 和预测均方根误差 RMSEP 分别为 0.954 2 和 0.256 0。通过选取合适的光谱区间或波长进行建模，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近红外光谱模型的预测能力和精度更高。

**参考文献：**


[1] 丁海泉，卢启鹏，朴仁官，等. 土壤有机质近红外光谱分析组合波长的优选[J]. 光学 精密工程，2007，15(12)：1946-1951.
DING H Q, LU Q P, PIAO R G, *et al*.. Optimum choice of combination wavelengths in near infrared analysis for soil organic matte [J]. *Opt. Precision Eng.*, 2007,15(12):1946-1951. (in Chinese)

[2] 陈洁梅，潘涛，陈星旦. 二阶导数光谱预处理在用 FTIR/ATR 方法定量测定葡萄糖-6-磷酸和果糖-6-磷酸中的应用[J]. 光学 精密工程，2006，14(1)：1-7.
CHEN J M, PAN T, CHEN X D. Application of second derivative spectrum prepares in quantification measuring glucose-6-phosphate and fructose-6-phosphate using a FTIR/ATR method[J]. *Opt. Precision Eng.*, 2006, 14(1):1-7. (in Chinese)

[3] CHEN Q SH, ZHAO J W, ZHANG H D, *et al*.. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration [J]. *Analytica Chimica Acta*, 2006,572(1):77-84

[4] 陈华才，吕进，陈星旦，等. 基于径向基函数网络的茶多酚总儿茶素近红外光谱检测模型的研究[J]. 光学 精密工程，2006，14(1)：58-62.
CHEN H C, LU J, CHEN X D, *et al*.. Near infrared spectroscopic model for determinating total catechins in tea polyphenol powder based on radical basis function network [J]. *Opt. Precision Eng.*, 2006,14(1):58-62. (in Chinese)

[5] CHEN Q SH, ZHAO J W, HUANG X Y, *et al*.. Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy[J]. *Microchemical Journal*, 2006,83(1):42-47.

[6] LUYPAERT J, ZHANG M H, MASSART D L. Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.)[J]. *Analytica Chimica Acta*,2003,478(2):303-312.

[7] SCHULZ H, JOUBERT E, SCHUTZE W. Quantification of quality parameters for reliable evaluation of green rooibos (Aspalathus linearis) [J]. *Eur. Food Res. Technol.*, 2003,216(6):539-543.

[8] NORGAARD L, SAUDLAND A, WAGNER J. Interval Partial Least Squares Regression (iPLS): a

comparative chemometric study with an example from near-infrared spectroscopy[J]. *Applied Spectroscopy*, 2000, 54: 413-419.

[9] LEARDI R, NORGAARD L. Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions[J]. *Journal of Chemometrics*, 2004, 18(11): 486-497.

[10] CHEN Q SH, ZHAO J W, LIU M H, *et al*.. Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms [J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2008, 46: 568-573.

[11] GHASEMI J, NIAZI A, LEARDI R. Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc ixture [J]. *Talanta*, 2003, 59: 311-317.

[12] ZOU X B, ZHAO J W, HUANG X Y, *et al*.. Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji' apple based on different PLS models [J]. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87: 43-51.

**作者简介：**

**郭志明**(1982－)，男，山东济宁人，2006 年于山东轻工业学院获得学士学位，2008 年于江苏大学获得硕士学位，现为江苏大学食品与生物工程学院博士研究生，主要从事近红外光谱分析方面的研究。Email: guozm_ujs@yahoo.com.cn

**赵杰文**(1945－)，男，江苏苏州人，博士，教授，博士生导师，分别于 1968 年、1980 年、1988 年在江苏工学院获学士、硕士、博士学位，主要从事食品和农产品无损检测的研究。Email: chaojw@ujs.edu.cn

**陈全胜**(1973－)，男，安徽桐城人，博士，2004 年于安徽农业大学获得硕士学位，2007 年于江苏大学获得博士学位，主要从事近红外光谱及高光谱图像技术在食品和农产品品质检测中的应用研究。E-mail: chenjiang0518@yahoo.com.cn

**黄星奕**(1963－)，女，江苏常州人，博士，教授，博士生导师，分别于 1985 年、1988 年、1999 年在江苏理工大学获得学士、硕士、博士学位，现任江苏大学食品与生物工程学院副院长，食品科学与工程专业学术带头人，主要从事计算机图像处理技术，电子舌技术及多信息融合技术在农产品、食品品质检测中的应用。Email: h_xingyi@163.com