近红外光谱相似性评估结合局部回归方法无损检测苹果糖度

夏阿林[1], 周新奇[2], 叶华俊[2], 张学锋[2], 陈英斌[2]

(1. 杭州电子科技大学 电子信息学院, 浙江 杭州 310018; 2. 聚光科技(杭州)股份有限公司, 浙江 杭州 310052)

摘 要: 基于 Bayesian相似性评估方法结合偏最小二乘局部回归, 对苹果近红外数据库进行数据挖掘。通过相似性计算方法搜索出与预测样品相近的近红外光谱, 形成校正子集后采用局部回归方法获得待测样品的相关信息。该方法所建立局部模型的平均检验标准偏差(SEV)约为 0.57, 分析 30个预测样品的预测标准偏差(SEP)约为 0.61; 基于马氏距离的传统方法建立的偏最小二乘局部模型的平均 SEV为 0.59, 分析 30个待测样品的预测 SEP为 0.64; 而采用整个数据库建立的全局偏最小二乘模型的 SEV约为 0.65, 分析 30个预测样品 SEP约为 0.70。基于 Bayesian相似性评估的局部回归方法在苹果糖度的近红外无损定量分析中获得较好的应用结果, 在实际应用中该方法比全局回归方法具有更强的适用性, 为近红外光谱分析提供了新的分析工具。

关键词: 近红外光谱; 相似度; 局部回归; 糖度; 苹果

中图分类号: O657.33 文献标识码: A 文章编号: 1004-4957(2010)12-1173-05

doi: 10.3969/j.issn.1004-4957.2010.12.010

Non-destructive Determination of Sugar Content in Apple by Near Infrared Spectroscopy with Similarity Evaluation Combined with Local Regression Method

XIA A-lin[1], ZHOU Xin-qi[2], YE Hua-jun[2], ZHANG Xue-feng[2], CHEN Ying-bin[2]

(1. Electronic Information College, Hangzhou Dianzi University, Hangzhou 310018, China; 2. Focused Photonics (Hangzhou), Inc., Hangzhou 310052, China)

Abstract A novel local regression method combined with similarity evaluation for near infrared spectra was proposed. In this method, the similarity evaluation based on Bayesian statistics was utilized to compare the NIR spectra. The calibration subsets were then selected to construct the model by partial least square method according to the similarity. The sugar content of apple samples was predicted with the model. The mean values of the standard errors of validation (SEV) and prediction (SEP) for the partial least square local regression method based on Bayesian statistics (B-PLS) were 0.57 and 0.61, respectively. Those for the partial least square local regression method based on Mahalanobis distance (M-PLS) were 0.59 and 0.64, respectively, and those for the partial least square global regression method (G-PLS) were 0.65 and 0.70, respectively. The results showed that B-PLS could accurately predict the sugar content in apple and its performance was superior to that of G-PLS and a little superior to that of M-PLS. Thus the proposed method possesses higher accuracy compared with G-PLS and could be widely applied in the rapid and nondestructive analysis of internal qualities such as the sugar content of apple. Furthermore, the method could provide a new tool for near infrared analysis.

Key words near infrared spectra; similarity; local regression; sugar content; apple

现代分析仪器产生了多种多样的谱图, 形成了巨大的数据库, 如何对数据库进行数据挖掘, 从中提取知识已成为当前比较热门的研究课题[1]。近红外光谱分析已广泛应用于各领域[2-4], 当近红外建模样本数目巨大时, 因样品的品种、产地等因素影响, 其光谱特征会存在较大差异, 建立的分析模型

收稿日期: 2010-07-14; 修回日期: 2010-08-24

基金项目: 国家 863项目资助(2009AA04Z129)

第一作者: 夏阿林(1974-), 男, 湖南邵阳人, 讲师, 博士, Tel: 0571-86919135, E-mail: alinxia@hdu.edu.cn

通常会导致预测准确性下降[5]。为保证分析模型具有较好的准确性，目前已报道了多种局部建模方法[6-10]。该类方法通过一定的规则从校正集中选取与待测样品尽可能相似的子集建立局部模型，对不同的待测样品分别采用各自的局部模型进行预测，研究的重点一般集中于选择合适的光谱特征参数和阈值以选择适当的子集，并采用不同的方式建立校正模型。

对物质表达信号进行相似性判断的方法称为相似性分析，该方法已成为对数据库进行数据挖掘，进而提取知识解决实际问题的重要方法[11]。本文采用 Bayesian相似性统计方法[12-13]评估苹果近红外漫反射光谱的相似性，在苹果近红外光谱库中搜索与预测样品光谱相似的校正子集，建立局部偏最小二乘模型，获得预测样品的定量分析结果。该方法相对于全局模型和基于马氏距离的偏最小二乘局部模型，提高了预测准确度。

1 理论部分

1.1 Bayesian相似性计算

将近红外光谱记为向量形式 $\boldsymbol{r}(r_1, r_2, \cdots r_n)$，其中 n表示 n个分析通道(波长)。任意一个样品的光谱则标记为 $r(m)$。光谱信号差值 Δr可用式(1)表示:

$$\Delta r = r(m_1) - r(m_2) = e(m_1) - e(m_2) \tag{1}$$

$e(m)$表示仪器的噪声信号，一般假设 e服从标准正态分布函数。

μ_i'为不同样品光谱之差 $\Delta r'$的平均值，则 μ_i'服从 $N(0, 2)$分布。对 μ_i'进行统计:

$$T = \frac{1}{n}(\mu_1' + \mu_2' + \cdots + \mu_n') \tag{2}$$

此时可用式(3)的假设判断光谱是否相似。当 H_0成立时，则说明两条光谱在统计上没有差别；反之则说明两条光谱存在统计上的显著差异。

$$\begin{cases} H_0 : T = 0 \\ H_1 : T \neq 0 \end{cases} \tag{3}$$

对于 $T = 0$的估计，因为 T的方差 $= 2/n$，则其标准差 $\sigma = \sqrt{2/n}$，3σ表示 99% 的置信度，故 $p(T \mid H_0)$和 $p(T \mid H_1)$分别满足 $N(0, 2/n)$分布和 $N = (3\sqrt{2/n}, 2/n)$分布，那么谱图相似性系数(LR)可用式(4)表示。

$$LR = \frac{p(T \mid H_0)}{p(T \mid H_1)} = \frac{\left[\frac{n}{4\pi}\right]^{\frac{1}{2}} e^{-\frac{n}{4}T^2}}{\left[\frac{n}{4\pi}\right]^{\frac{1}{2}} e^{-\frac{n}{4}(T - 3\sqrt{\frac{2}{n}})^2}} = e^{-\frac{n}{4}(6T\sqrt{\frac{2}{n}} - \frac{18}{n})} \tag{4}$$

对于光谱间是否相似可用式(5)进行评判:

$$\mathrm{POR} = \frac{p(H_0 \mid T)}{p(H_1 \mid T)} = \alpha e^{-(\frac{n}{4})(6T\sqrt{\frac{2}{n}} - \frac{18}{n})} \tag{5}$$

如果 POR > 1，则接受 H_0，认为光谱相似；否则光谱是不相似的，因此光谱相似性度量转化成假设检验。本文通过计算 POR 判断光谱是否相似并以此作为阈值选择待测样品的子校正样品集。式(5)中的 α是未知数，可通过重复采集同一样品的光谱进行计算。本文重复采集同一样品的 10条光谱，以10条光谱预处理后平均值的最大差值的 3倍为基础计算获得 α为 0.011 1。

1.2 偏最小二乘局部回归

局部建模方法的基本思想是选择与待测样品相似的样品作为建模的校正子集建立局部回归模型。本文根据待测样品与数据库中各样品光谱的相似性，为每个待测样品选择校正子集，使用偏最小二乘算法[14]进行回归计算，建立局部模型并对待测样品进行预测。

2 实验部分

实验方法流程如图 1所示。

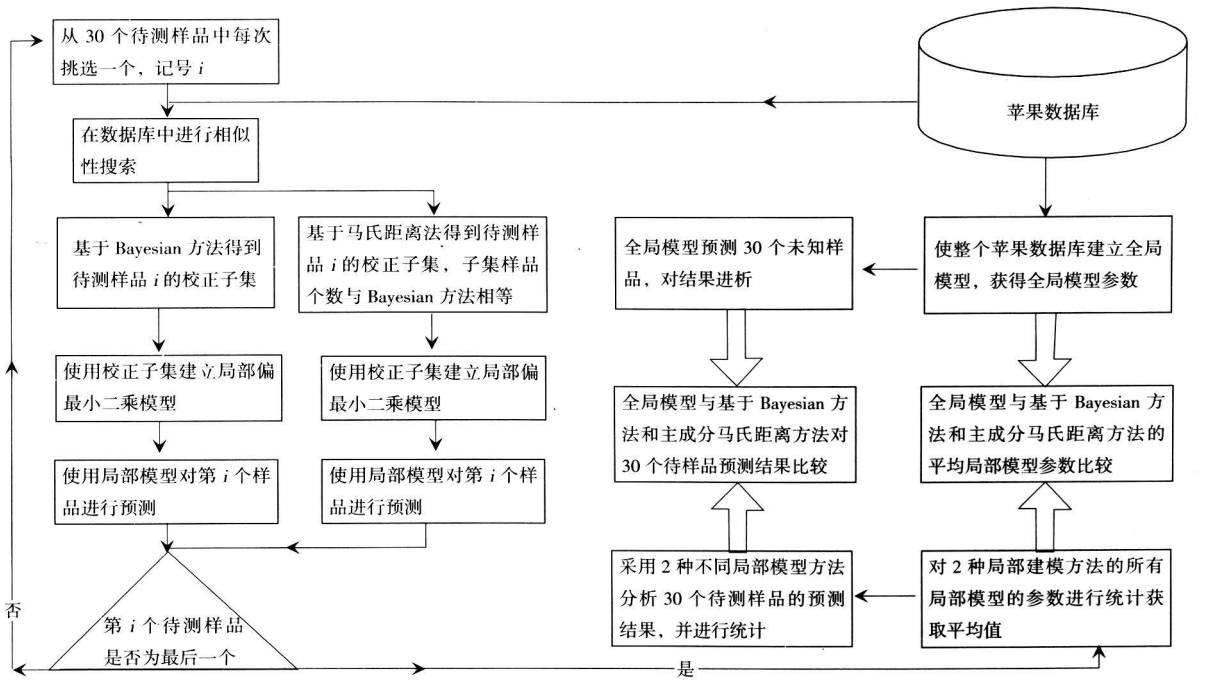

图 1　实验方法流程

Fig. 1　Flow chart of experiment

2.1　苹果数据库的建立

数据库中的样品为陕西红富士和秦冠 2个品种苹果。于 2009年 10月上旬至 11月下旬依据苹果成熟程度分 3批采收，总计约 1 000个样品，其中 1/3来自白水，1/3来自韩城，其余的来自旬邑。每次将采集的样品分成数量相等的两份，其中一份采摘后 2 d内完成光谱采集和糖度测量，剩余一批冻库冷冻 15 d后采集光谱并进行糖度测量。

使用光栅阵列近红外光谱分析仪(聚光科技研制，型号 SupNIR-1100)采集苹果光谱。光谱检测系统参数：测量波长 600~ 1 100 nm，光谱平均次数为 5次，仪器分辨率 6 nm，硅阵列传感器。每个苹果均测量其赤道部位，且测量部位尽可能避免明显的表面缺陷。

上述收集的样品扫描光谱后经切片榨汁，使用数字式折光仪(日本爱拓，型号 PR-32a)检测果汁的糖度(可溶性固形物)。糖度性质统计如表 1所示。两种苹果糖度性质基本呈正态分布，如图 2所示。将样品的近红外光谱和糖度性质进行关联，形成苹果基础数据库。

表 1　样品糖度性质统计结果

Table 1　Statistical results of the brix nature in apples

Sample	Number	Maximum	Minimum	Mean
Fuji apple	527	11.40	21.65	15.65
Qinguan apple	471	11.80	19.80	15.75

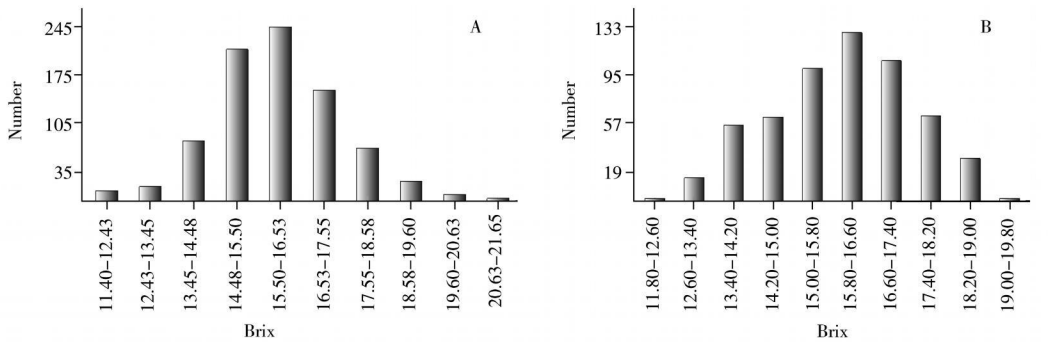

图 2　红富士(A)和秦冠(B)苹果的糖度分布

Fig. 2　Brix distributions for Fuji apple(A) and Qinguan apple(B)

2.2 待测样品的分析

采用相同的近红外分析仪扫描 30个待测样品，获取其光谱；同时采用数字式折光仪检测其糖度参考值，与近红外光谱预测值进行比较。

3 结果与讨论

3.1 苹果近红外光谱分析

数据库中苹果的吸收光谱如图 3A所示。由图可知，样品的吸收光谱存在显著基线漂移现象，故在相似度计算前需对样品光谱进行预处理，以消除基线干扰，本文采用一阶导数法消除该现象。求导后的光谱如图 3B所示。由图 3可知，样品在 600~ 700 nm存在较大色素吸收峰，在 850 nm附近出现弱吸收峰信号，该处为 C—H键的吸收峰，表达了有机化合物的含量信息；在 950 nm附近存在明显的吸收峰，该处为 O—H键的吸收峰，表达了水分子和其他含羟基化合物(如糖分)的含量信息。

将待测样品光谱与数据库中各样品光谱进行相似性匹配，搜索出与待测样品最相似的校正子集光谱。某待测样品光谱与数据库中光谱的相似度为 89.877~ 90.015(仅列出相似度最大的前 20个数据)，说明预测样品与数据库中光谱相似程度较高。表明近红外光谱差异性小，采用全局模型进行分析则可能较难提取信息。

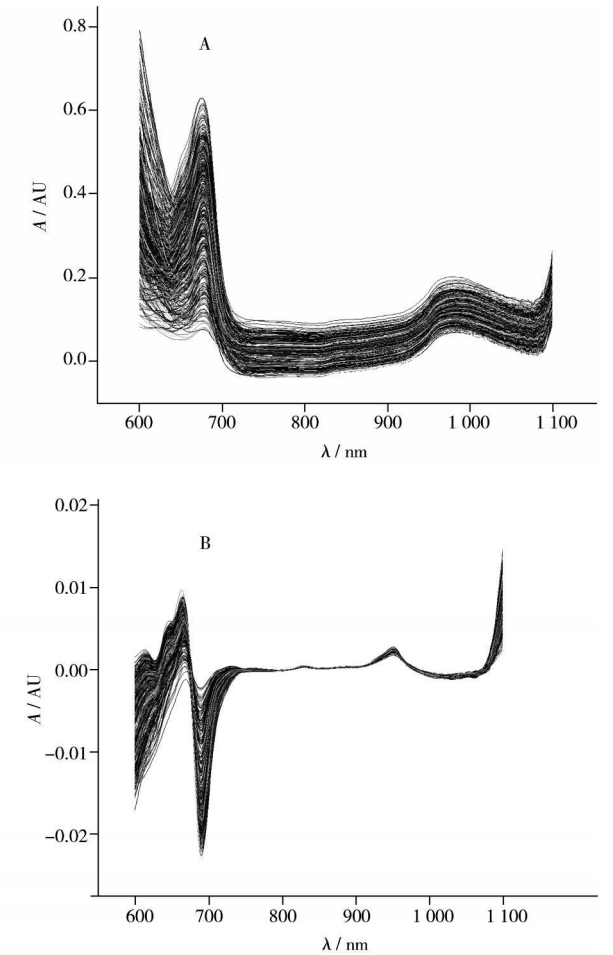

图 3 数据库中样品的原始吸收光谱(A)与一阶导数光谱(B)

Fig. 3 Original(A) and first derivative(B) absorption spectra of samples in database

3.2 局部回归比较分析

通过 Bayesian相似性计算方法，依据 POR判定阈值从数据库搜索到与待测样品光谱相似的样品，并建立校正子集，采用局部偏最小二乘方法建立模型分析待测样品，获得预测结果。同时采用传统的基于马氏距离局部偏最小二乘方法针对相同的待测样品搜索到与 Bayesian方法相同个数的样品，并建立模型获得待测样品预测值。

确定局部建模的模型参数一般有内部检验和外部检验两种方法。本文采用外部检验方法，在建模过程中利用检验集对有关参数进行选择。首先采用 K- S方法[15]在搜索到的校正子集中选择 80% 样品用于建立模型，剩余 20% 样品用于检验并优化模型。然后采用优化后的分析模型对待测样品进行分析。

分别对 30个待测样品采用两种局部建模方法建立了 30个相应的局部模型。基于 Bayesian相似性评估方法的偏最小二乘局部模型(B-PLS)的平均校正偏差(SEC)为 0.52，平均检验标准偏差(SEV)为 0.57，对 30个待测样品进行预测分析，其预测值与参考值的预测标准偏差(SEP)为 0.61；基于马氏距离方法的偏最小二乘局部模型(M-PLS)的平均 SEC为 0.55，平均 SEV为 0.59，30个待测样品的 SEP为 0.64，说明基于 Bayesian相似性评估方法建立的偏最小二乘局部模型的检测结果略优于传统基于马氏距离的偏最小二乘局部建模方法。

3.3 局部回归模型与全局回归模型的比较

为了对基于 Bayesian相似性评估的局部回归方法和全局偏最小二乘建模方法进行比较，本文对整个数据库进行模型分析，建模参数参照局部模型的参数。全局偏最小二乘模型(G-PLS)校正集的 SEC

为 0.68，检验集的 SEV 为 0.65，30个待测样品的 SEP为 0.70。将该结果与基于 Bayesian方法的平均局部模型结果进行比较可知，局部模型检验集的平均 SEV 小于全局模型检验集的 SEV，说明局部模型优于全局模型。

采用 Bayesian局部模型和全局模型同时分析 30个待测样品，待测样品糖度的参考值与预测值残差分布见图 4。由图 4可知，全局模型预测 A样品的残差较大，而此时采用基于相似性分析的局部模型进行预测时，其预测偏差显著小于全局模型，说明在实际应用中局部模型比全局模型具有更好的准确性。Bayesian局部模型的预测标准差 (SEP)小于全局模型的预测标准差，也说明局部模型的准确性优于全局模型。

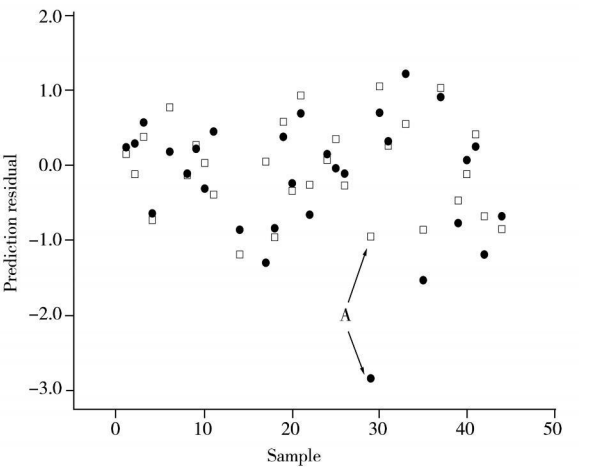

图 4　局部和全局模型预测样品预测值与参考值的残差

Fig. 4　Residual of local and global regression models for sample prediction

□ local model; • global model

参考文献:

[1] WANG J S, LAI L H, TANG Y Q. Data mining of toxic chemicals: structure patterns and QSAR[J]. Molecular Modeling, 1999, 5(11): 252-262.

[2] 陆婉珍. 现代近红外光谱分析技术[M]. 2版. 北京: 中国石化出版社, 2007: 1.

[3] 严衍禄, 赵龙莲, 韩东海, 杨曙明. 近红外光谱分析基础与应用[M]. 北京: 中国轻工业出版社, 2005.

[4] 冯红年, 黎庆涛, 卢家炯, 黎世文, 刘立鹏, 夏阿林, 王健. 近红外光谱技术用于白砂糖质量的实时监测研究[J]. 分析测试学报, 2009, 28(12): 1460-1463.

[5] BERZAGHI P, SHENK J S, WESTERHAUS M O. Local prediction with near infrared multiproduct databases[J]. J Near Infrared Spectrosc, 2000, 8(1): 1-9.

[6] DAVIES A M C, FEARN T. Quantitative analysis via near infrared databases: comparison analysis using restructured near infrared and constituent data-deux(CARNAC-D) [J]. J Near Infrared Spectrosc, 2006, 14(6): 403-411.

[7] WANG Z, ISAKSSON T, KOWALSKI B R. New approach for distance measurement in locally weighted regression[J]. Anal Chem, 1994, 66(2): 249-260.

[8] CENTNER V, MASSART D L. Optimization in locally weighted regression[J]. Anal Chem, 1998, 70(19): 4206-4211.

[9] 石雪, 蔡文生, 邵学广. 基于小波系数的近红外光谱局部建模方法与应用研究[J]. 分析化学, 2008, 36(8): 1093-1096.

[10] ATKESON C G, MOORE A W, SCHAAL S. Locally weighted learning[J]. Artificial Intelligence Rev, 1997, 11(1/5): 11-73.

[11] GAN F, YE R Y. New approach on similarity analysis of chromatographic fingerprint of herbal medicine[J]. J Chromatogr A, 2006, 1104(1/2): 100-105.

[12] NIKOLOVA N, JAWORSKA J. Approaches to measure chemical similarity—— a review[J]. QSAR Comb Sci, 2003, 22(9/10): 1006-1026.

[13] PRESS S J. Bayesian statistics: Principles, models and applications[M]. USA, New York: Wiley & Sons, Inc, 2002.

[14] ASTM E 1655-05. Standard practices for infrared multivariate quantitative analysis[S]. United States, 2005: 10.

[15] KENNARD R W, STONE L A. Computer aided design of experiments[J]. Technometrics, 1969, 11(1): 137-148.