文章编号:1001-9014(2007)06-0433-04

基于独立组分分析和BP神经网络的可见/近红外光谱稻谷年份的鉴别

邵咏妮, 曹 芳, 何 勇
(浙江大学 生物系统工程与食品科学学院,浙江 杭州 310029)

摘要:建立了一种基于独立组分分析的可见/近红外光谱反射技术快速鉴别稻谷年份的新方法.首先用独立组分分析方法获取不同年份稻谷的可见/近红外光谱载荷图,将载荷图中相关性最大的波段(特征波段)作为人工神经网络的输入建立稻谷年份的鉴别模型.每个年份40个样本,3个年份共120个样本用来建立BP神经网络模型,剩余的3个年份各20个样本用于预测.预测的结果表明,准确率达到100%.同时通过独立组分分析,得到了稻谷主要成分对应的敏感波段.说明本文提出的基于独立组分分析的方法具有很好的鉴别效果,为稻谷的年份鉴别提供了一种新方法.

关 键 词:可见/近红外光谱;稻谷;独立组分分析;BP神经网络
中图分类号:S511.2+2; S511.3+3;TH744.1 **文献标识码**:A

DISCRIMINATION YEARS OF ROUGH RICE BY USING VISIBLE/NEAR INFRARED SPECTROSCOPY BASED ON INDEPENDENT COMPONENT ANALYSIS AND BP NEURAL NETWORK

SHAO Yong-Ni, CAO Fang, HE Yong
(College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China)

Abstract: A new method for discrimination years of rough rice based on independent component analysis was developed by using visible/near infrared spectroscopy (Vis/NIRS). First, the Vis/NIR loading weight of rough rice with different years was got by using independent component analysis (ICA) and setting the wavelengths corresponding to the maximal correlation as the inputs of artificial neural network (ANN), then the discrimination model was build. 120 samples (40 with each year) from three years were selected randomly as a calibration set; the left 60 samples (20 with each year) were as the prediction set. The discrimination rate of 100% was achieved. Synchronously, the sensitive wavelengths corresponding to the main components in rough rice were obtained with ICA. It indicates that the result for discrimination years of rough rice is very good based on ICA method, and it offers a new approach to the fast discrimination years of rough rice.

Key words: visible/near infrared spectroscopy (Vis/NIRS); rough rice; independent component analysis (ICA); BP neural network (BP-NN)

引言

我国稻谷主要产在秦岭淮河以南,而长江、珠江流域产量尤多.产区广,产量大,种类多.按稻谷籽粒特性的不同,一般可将稻谷分为粳稻和籼稻两大类.粳稻谷:稻粒一般呈椭圆形,谷壳组织松而薄,米粒强度大,耐压性很好,加工时碎米少,出米率高.晚粳稻谷,糙米腹白较小,硬质粒多.

稻谷的工艺品质,主要是指与加工工艺效果直接有关的物理性质和化学性质,如稻谷的品种、形状、结构、水分、气味及化学成分等内容,不同品种和不同等级、年份的稻谷,具有不同的工艺品质,它直接影响到成品质量的优劣和出米率的高低.

目前在稻谷的年份鉴别中,国内外标准方法主

收稿日期:2006-11-27,**修回日期**:2007-05-14 **Received date**: 2006-11-27, **revised date**: 2007-05-14
基金项目:国家自然科学基金(30671213)、国家十一五科技支撑计划(2006BAD10A07)和高等学校博士学科点专项科研基金(20040335034)资助项目
作者简介:邵咏妮(1983-),女,浙江慈溪人,浙江大学生物系统工程与食品科学学院博士生,主要从事数字农业和多光谱检测技术研究.

要是形态学方法、荧光扫描鉴定法、化学鉴定法和电泳鉴定法. 这些方法的缺点是可操作性不强,且仪器成本高,操作复杂,不适合非专业人士操作,不适合于无损在线的快速检测. 因此,研究一种简单、快速、无损的稻谷年份鉴别技术和开发相应的检测仪器很有必要.

可见/近红外光谱法是一种快速简便无损的分析方法. 它已被广泛应用于食品、石油、化工、医药等行业[1~2]. 在稻谷品质检测方面,也有相关报导,如 Kwon and Cho 应用近红外光谱技术对稻米品种进行鉴别[3]. Kim et al. 利用近红外光谱技术对国产及进口的稻谷品种进行鉴别[4].

独立组分分析(ICA)是近几年来伴随着盲信号分离问题发展起来的一种新的信号处理技术,该方法的基本思路是以非高斯信号为研究对象,在独立性假设的前提下,对多路观测信号进行盲源分离,并已成功地应用于特征提取、图像处理、生物医学信号处理、金融等领域[5~8]. 在对近红外光谱分析方面,也有相关的报导[9~10].

人工神经网络是处理非线性问题的三大方法之一[11]. 其基本思想是模拟人脑细胞(神经元)工作原理,以建立模型进行分类和预测. 目前使用最多的是多层结构的误差反向传播学习算法(BP)[12].

本研究通过对不同年份晚粳谷的可见/近红外反射光谱进行分析,同时将独立组分分析与 BP 神经网络相结合,目的在于寻找一种稻谷年份鉴别的快速有效的方法.

1 材料与方法

1.1 仪器设备

实验使用美国 ASD(Analytical Spectral Device)公司的 Handheld FieldSpec 光谱仪,其光谱测定范围 325 ~ 1075nm,采样间隔为 1.5nm,分辨率为 3.5nm,,探头视场角为 20°,光谱扫描次数设定为 30 次. 光源采用 14.5V 卤素灯. 分析软件采用 ASD View Spec Pro ,Unscrambler V9.5, Matlab 7.0.

1.2 样品制备及光谱的扫描

从当地粮食储备公司购得 2003,2004 和 2005 年 3 个不同年份的晚粳谷. 实验前稻谷被置于冷库中存放. 从完整的,未发芽的,无霉变的稻谷中挑选出 180 个样本. 样本都盛放在统一尺寸的玻璃培养皿中. 培养皿的直径是 95mm,高度是 14mm,样本以盛满培养皿为准. 光谱仪探头位于样本的正上方,距离样本高度为 100mm,探头视场角为 20°. 光源距离培养皿中心 300mm,与水平面成 45°角. 对于每一个样本,3 个光谱记录被保存下来,每个光谱记录是 30 次扫描的平均值. 从 3 个年份的 180 个样本中,随机选择 40 个样本共 120 个作为建模集,剩余的每个年份 20 个样本,共 60 个样本作为预测集.

1.3 光谱数据预处理

为了去除来自高频随机噪音、基线漂移、光散射等影响,需要对原始光谱进行预处理来消除噪音. 采用 Savitzky-Golay 平滑法,选用平滑点数为 7,再进行附加散射校正(MSC)处理. 为消除光谱数据在采集时首端与末端产生的部分噪音,采用 400 ~ 1000nm 波段的光谱数据进行分析.

1.4 独立组分分析

令 s 为 n 个原始独立组分信号,$s=(s_1, s_2, \cdots, s_n)^T$, 假定观察到 m 个($m \geqslant n$) 原始信号的线性组合 x, $x=(x_1, x_2, \cdots, x_m)^T$,则有 $x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n (j=1,2,\cdots,m)$. 可以表示为

$$x = As \quad , \tag{1}$$

A 为 $m \times n$ 的混合矩阵. 当这种观察进行了 k 次,则上式为

$$X = AS \quad , \tag{2}$$

式中 X 为 $m \times k$ 矩阵,S 为 $k \times n$ 矩阵,式(2)则为 ICA 模型. 目前最常用的是 FastICA 算法,它又称固定点(Fixed-Point)算法,是由芬兰赫尔辛基大学 Hyvärinen 等人提出来的,是一种快速寻优迭代算法,基本步骤如下:

a. 对观测数据 X 进行中心化,使它的均值为 0;

b. 对数据进行白化,$X \rightarrow Z$;

c. 选择一个初始权重矢量 W;

d. 令 $W^* = E\{Zg(W^TZ)\} - E\{g'(W^TZ)\}W$,$g$ 为奇次非二次函数;

e. 令 $W = W^* / \| W^* \|$;

f. 如果没有收敛,返回步骤 d.

FastICA 算法通过 Matlab7.0 实现.

1.5 BP 神经网络模型

人工神经网络是一种模拟人脑功能的信息处理系统,它主要借鉴了人脑神经系统处理信息的过程,以数学网络拓扑结构为理论基础. 因其突出的非线形映照能力,倍受化学计量学家的关注. 目前最常用的是反向传输(Back Propagation ,BP)模型.

本文建立了一个三层的 BP 神经网络模型,各层传递函数采用 S 型(Sigmoid)函数. 网络输入层节点数为 8,经多次实验确定隐含层节点数为 10,输出层节点数为 3. 目标误差设为 0.0001,网络学习速率

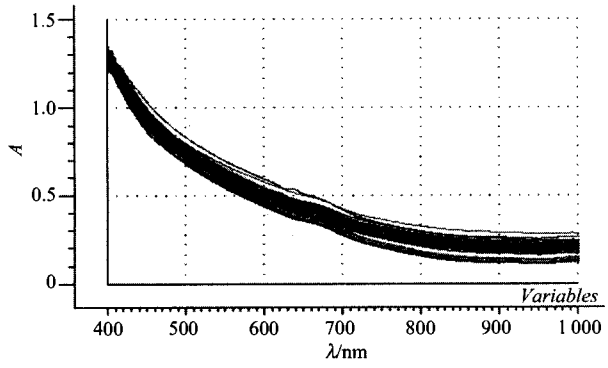

图 1　不同年份晚粳谷的可见/近红外光谱吸光度图
Fig. 1　Vis/NIR absorbance spectra of three different years of late medium to short-grain nonglutinous rice

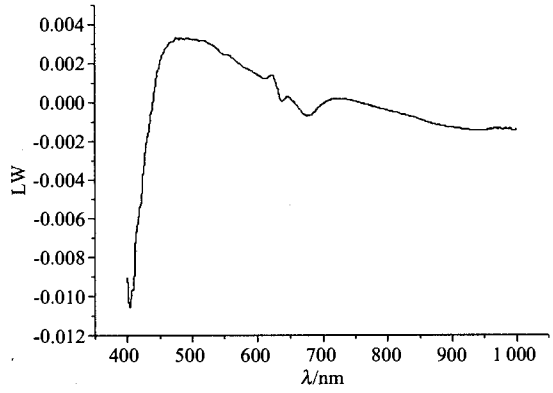

图 2　独立成分 1 在 400 ~ 1000nm 波段的载荷图
Fig. 2　The loading plot of IC1 across the spectral region from 400 ~ 1000nm

为 0.2，训练迭代次数为 1000 次.

2　试验结果与分析

2.1　不同年份晚粳谷的可见/近红外光谱图分析

3 种年份晚粳谷的原始近红外光谱如图 1 所示. 图 1 中横坐标为波长范围 400 ~ 1000nm，纵坐标为光谱吸光度值. 从图 1 中可以看出，不同年份晚粳谷的原始光谱曲线没有明显的差别，必须通过一定的光谱预处理及数学模型才能实现不同年份的鉴别. 我们采用 ASD View Spec Pro 软件，将得到的光谱曲线进行 log(1/R)处理，并转换成 ASCII 码，得到吸光度矩阵，再用独立组分分析法提取有用的光谱信息.

2.2　建立 BP 模型

全波段从 400 ~ 1000nm 共有 601 个点，但是采用全谱计算时，计算量大，而且有些区域样品的光谱信息很弱，与样品的组成或性质间缺乏相关关系. 所以在独立组分分析的基础上，选取敏感波段作为输入建立 BP 模型. 独立组分的选取对模型的好坏有一定的影响，因此本文将独立组分从 8 变化到 20，根据模型最终的预测结果，确定独立组分数为 10.

从图 2 和图 3 的载荷图中可以看出独立成分 1、2 与各个变量的相关程度，独立成分 3 ~ 10 略. 图 2 和图 3 中横坐标表示波长范围 400 ~ 1000nm，纵坐标表示各波长变量对于独立成分的载荷值，即是各个变量与独立成分的相关性大小. 从图 2 可以看出独立成分 1 与 475nm 波长处的吸光度相关性最大；从图 3 可以看出独立成分 2 与 400nm 波长的吸光度值的相关性最强. 同理可得，独立成分 3 ~ 10 的对应波长为 770nm，970nm，972nm，996nm，880nm，922nm，970nm 和 996nm.

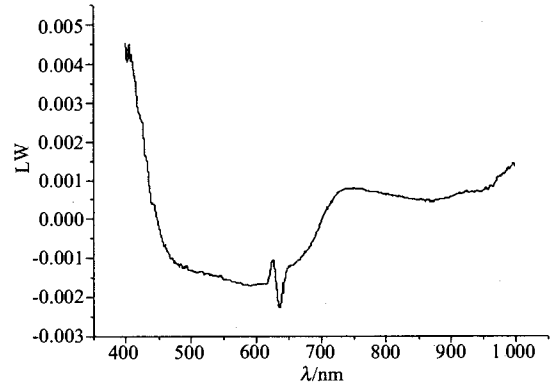

图 3　独立成分 1 在 400 ~ 1000nm 波段的载荷图
Fig. 3　The loading plot of IC2 across the spectral region from 400 ~ 1000nm

从 400 ~ 1000nm 范围的 601 个采样点中选取上述 8 个特征波长采样点的吸光度值作为 BP 网络的输入，并建立模型，隐含层结点数依据式(3)分别定为 5 ~ 14.

$$n_1 = \sqrt{n+m} + a \quad , \tag{3}$$

式(3)中 n_1 为隐含层结点数，n 为输入神经元个数，m 为输出神经元个数，a 为 1 ~ 10 之间的常数. 其它人工神经网络参数设置为：最小训练速度为 0.2，数据进行标准化转化，Sigmoid 参数为 0.9，动态参数为 0.6，允许误差为 0.0001. 通过计算模型的相关系数(r)，预测标准偏差(SEP)的值对隐含层节点数进行调整. 当隐含层的节点数为 10，即模型结构为 8-10-3 时，预测结果最佳，得模型拟合残差为 5.694235×10^{-5}，预测准确率为 100%. 预测结果好于最小二乘法回归和主成分分析回归方法，图 1 为 3 种模型的预测结果比较. 独立组分结合神经网络对 60 个未知样本的实测 vs 预测值见图 4.

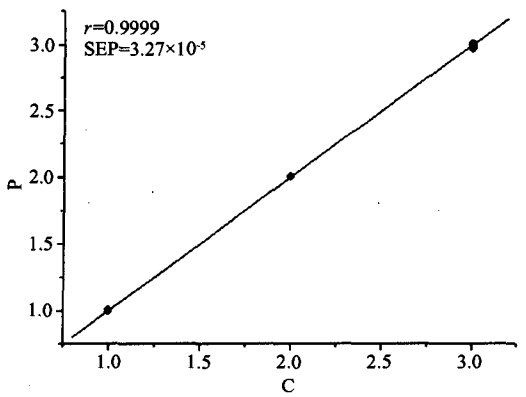

图4 60个未知样本的IC-BP预测模型

Fig. 4 Prediction results for the unknown 60 samples from the IC-BP model

表1 3种模型对60个未知样本的预测结果比较

Table 1 The comparison of three models for 60 unknown samples

	独立组分—神经网络(IC-BP)	最小二乘法回归(PLS)	主成分分析回归(PCR)
相关系数r	0.9999	0.9834	0.9810
预测标准偏差	3.27×10^{-5}	0.0330	0.0366
准确率(%)	100	96.7	95.0

2.3 晚粳谷主要成分对应的敏感波段

为了研究稻谷不同年份的鉴别本质,我们将对不同年份晚粳谷的主要成分进行分析.一般地,组成稻谷的主要成分有水,蛋白质,脂肪等.不同年份稻谷的主要成分差别在于含水量的不同.本文通过对采集的光谱数据进行独立组分分析,分析并得到晚粳谷主要成分对应的敏感波段.由上述的独立成分1~10中,分析得到独立成分3(770nm)和4,9(970nm)对应晚粳谷中的水分含量[13],独立成分7(880nm)对应脂肪含量[14],独立成分5(972nm),6,10(996nm),8(922nm)对应蛋白质含量[13].对早籼谷的研究也得到类似的分析结果,因此本文的研究对不同种类的稻谷具有普适性.

3 结语

应用可见/近红外光谱结合独立组分分析与BP神经网络对稻谷年份进行了初步的判别分析,通过提取样本的敏感波段作为神经网络的输入,进而得到了稻谷年份的鉴别模型,该模型的识别率达到100%,鉴别结果令人满意.同时通过独立组分分析,找到了晚粳谷主要成分对应的敏感波段,其中770nm,970nm对应水分含量,880nm对应脂肪含量,922nm,972nm,996nm对应稻谷中蛋白质的含量.说明利用可见/近红外光谱技术结合化学计量学的方法对稻谷年份进行快速鉴别是可行的.它为稻谷年份的快速检测提供了一种新方法.

REFERENCES

[1] YAN Yan-Lu, ZHAO Long-Lian, HAN Dong-Hai, *et al. The Foundation and Application of the Near Infrared Spectros Copy Analysis*[M]. Beijing: China Light Industry Press(严衍禄,赵龙莲,韩东海,等. **近红外光谱分析基础与应用**北京:中国轻工业出版社),2005.

[2] SHAO Yong-Ni, HE Yong. Method for prediction acidity of bayberry juice by using Vis/near infrared spectra [J]. *J. Infrared Millim. Waves*(邵咏妮,何勇.可见/近红外光谱预测杨梅汁酸度的方法研究.**红外与毫米波学报**),2006,**25**(6):478—480.

[3] Kwon Y K, Cho R K. Identification of rice variety using near infrared spectroscopy [J]. *Journal of Near Infrared Spectroscope*,1998,**6**:67—73.

[4] Kim S S, Rhyu M R, kim J M, *et al.* Authentication of rice using near-infrared reflectance spectroscopy [J]. *Cereal Chem.*,2003,**80**(3): 346—349.

[5] Hyvarinen A. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation [J]. *Neural Computation*,1999,**11**(7):1739—1768.

[6] Hoyer P O, Hyv rinen A. Independent component analysis applied to feature extraction from colour and stereo images [J]. *Network: Computation in Neural Systems*, 2000, **11**(3):191—210.

[7] Hyv rinen A, Hoyer P O. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces [J]. *Neural Computation*, 2000,**12**(7):1705—1720.

[8] Shao X G, Wang G Q, Wang S F, *et al.* Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background [J]. *Anal. Chem.*, 2004, **76**(17):5143—5148.

[9] Chen J, Wang X Z. A new approach to near-infrared spectral data analysis using independent component analysis [J]. *J. Chem. Inf. Comput. Sci.*,2001,**41**:992-1001.

[10] BI Xian, LI Tong-Hua, WU Liang. Application of independent component analysis to the IR spectra analysis [J]. *Chemical Journal of Chinese Universities*(毕贤,李通化,吴亮.独立组分分析在红外光谱分析中的应用.**高等学校化学学报**),2004,**25**(6):1023-1027.

[11] HE Yong, LI Xiao-Li. Discrimination of varieties of waxberry using near infrared spectra [J]. *J. Infrared Millim. Waves*(何勇,李晓丽.近红外光谱杨梅品种鉴别方法的研究.**红外与毫米波学报**),2006,**25**(3):192——194,212.

[12] HUANG Min, HE Yong, HUANG Lin-Xia, *et al.* Discrimination of varieties of silkworm egg based on visible-near infrared spectra[J]. *J. Infrared Millim. Waves*(黄敏,何勇,黄凌霞,等. 基于可见-近红外光谱技术的家蚕蚕种鉴别方法的研究.**红外与毫米波学报**), 2006, **25**(5): 342—344.

[13] Osborne B G, Fearn T. Near Infrared Spectroscopy in Food Analysis[J]. *Longman Scientific and Technical*, Essex, U. K.,1986.

[14] Sasic S, Ozaki Y. Short-wave near-infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein, and lactose in raw milk by partial least-squares regression and band assignment [J]. *Anal. Chem.*, 2001,**73**:64—71.