

基于 DPLS 特征提取的 LDA 方法在玉米近红外光谱定性分析中的应用

覃 鸿, 王徽蓉, 李卫军*, 金小贤

中国科学院半导体研究所, 北京 100083

摘 要 提出了一种基于 DPLS+ LDA 的玉米近红外光谱定性分析新方法。该方法在训练时, 首先用包含 30 个玉米品种每个品种 20 个近红外光谱样本的训练集进行 DPLS 回归, 确定最佳 DPLS 主成分数为 28; 然后对训练集光谱进行 DPLS 特征提取后再进行 LDA 分析, 确定最佳 LDA 主成分数为 26, 并提取 LDA 特征。识别时, 测试样本经过 DPLS+ LDA 特征提取后, 用最小距离分类器进行识别。实验比较了 DPLS+ LDA 方法与传统的 DPLS 回归预测方法及 DPLS 特征提取方法的判别结果, DPLS+ LDA 方法的性能最优, 等识率达到了 96.18%, 而传统 DPLS 预测方法只有 85.38%, DPLS 特征提取方法为 95.76%。实验结果说明 DPLS+ LDA 方法是一种有效的玉米近红外光谱定性分析方法, 且具有很强的推广能力。

关键词 近红外光谱; 线性判别分析; 判别式偏最小二乘; 定性分析; 玉米

中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2011)07-1777-05

引 言

我国是玉米生产和消费大国, 未来对优良玉米品种的需求在不断增加, 品种的选育、种子的生产经营以及政府部门的管理监督等各个环节都迫切需要对玉米品种纯度和真实性进行快速鉴定的技术。传统的形态学检验法简单、直观, 但要求检验人员具有系统的农业科技知识, 同时必须熟悉品种之间的形态特征和生理特性的差异, 知道被检验品种特殊的形态特性, 因而不能满足种子生产、运输和加工等各个环节对检验的要求。随着种子科学与技术的不断发展, 品种鉴定由外观形态检验向室内生理指标和分子基因水平上的鉴定发展, 如电泳技术、分子标记技术^[1]以及近红外光谱分析^[2,3]等。室内鉴定周期短、费用低, 因而研究出简便、准确、快捷的种子室内鉴定方法成为近年来的热点。

近红外光谱(near infrared spectroscopy, NIRS) 包含有丰富的物质信息, 光谱与物质本身的组成及含量是密切相关的^[2], 因此, 同类光谱之间存在相似性, 异类光谱之间存在差异性。利用品种内的相似性和品种间的差异性, 可以对物质的近红外光谱进行定性分析, 通过对光谱特征的分析研究来确定样品的归属。并且, 近红外光谱信息丰富、容易获取, 借助于计算机技术和化学计量学方法, 可以对样品进行无损分析。简而言之, 近红外光谱分析方法具有简易、快速、无

损等特点^[3], 近几年来开始成为品种识别的一个研究方向^[4-6]。

由于近红外光谱数据量通常很大, 并且光谱吸光度差异通常很小, 光谱具有复杂性和多变性, 有效信息率低, 因而在近红外光谱定性分析中通常需要对光谱进行数据压缩以及信息提取, 最常用的方法有主成分分析(principal component analysis, PCA)^[6]和偏最小二乘法(partial least squares, PLS)。Fisher 线性判别分析(linear discriminant analysis, LDA), 是模式识别中一种较为普遍的用于特征提取的方法^[7-10], 但 LDA 方法在实际应用中经常遇到小样本问题(small sample size, SSS), 即样本的维数远远大于样本的数量, 从而导致 LDA 求解困难, 一个常用的解决办法是对样本进行降维。LDA 在近红外光谱定性分析方面鲜有应用, 近红外光谱通常维数很高, 而可获取的样本数量相对较少, 同样存在小样本问题。本文提出了一种 DPLS+ LDA 的特征提取方法对玉米近红外光谱进行定性分析, 先利用判别式偏最小二乘法(discriminant partial least squares, DPLS) 进行对玉米近红外光谱降维, 克服小样本问题, 然后采用线性鉴别分析的方法提取特征, 用最小距离分类器进行分类。实验中, 比较了 DPLS 回归预测方法、DPLS 特征提取方法和 DPLS+ LDA 特征提取方法对已训练玉米品种的正确识别率以及对未训练玉米品种的正确拒识率。实验结果表明 DPLS+ LDA 方法对已训练玉米品种样品的正确识别率最高, 同时

收稿日期: 2010-09-21, 修订日期: 2010-12-18

基金项目: 国家自然科学基金项目(60753001, 90920013)资助

作者简介: 覃 鸿, 女, 1977 年生, 中国科学院半导体研究所助理研究员 e-mail: qinh@semi.ac.cn

* 通讯联系人 e-mail: wjli@semi.ac.cn

能有效拒识未训练玉米品种的样品, 因而 DPLS+ LDA 方法性能最优, 且有较强的推广能力。

1 材料与方法

1.1 光谱获取

试验用的玉米种子由国家科技支撑计划高产优质多抗玉米项目组提供, 玉米籽粒样品为全国各地种植的同年份的 37 个玉米品种, 样品盛放在统一尺寸的样品杯中, 对同一品种的玉米籽粒各自进行 25 次取样。实验所用仪器为德国 BRUKER 公司生产的 MPA 近红外光谱仪, 每条光谱为 64 次扫描的平均值, 扫描区域为: $12\,000 \sim 4\,000\text{ cm}^{-1}$, 分辨率为 16 cm^{-1} , 光谱数据点为 1 037 个。

实验中, 将采集的样品光谱数据分成 3 部分, 第一部分为训练样本集, 选取 30 个品种, 每个品种选取 20 个样本组成, 共 600 个样本; 第二部分为第一部分所选取的 30 个品种的余下的 5 个样本组成, 共 150 个样本, 作为第一测试集; 第三部分由前两部分未包含的其余 7 个品种的所有样本组成, 共 175 个样本, 作为第二测试集。

1.2 光谱标准化预处理

标准化预处理能使光谱中所有波长变量得到相同的权重, 并可校正样品间因散射而引起的光谱的误差, 扣除样品光谱中的线性平移的影响^[11]。本文采用标准化方法来对原始光谱进行预处理, 具体计算步骤为: (1) 对由 n 个样本组成的一组原始光谱 x_i , $i = 1, 2, \dots, n$, 计算其平均光谱 μ ; (2) 计算光谱数据的标准差 σ ; (3) 处理后的光谱值 $x'_i = \frac{x_i - \mu}{\sigma}$ 。

1.3 DPLS 特征提取

1.3.1 偏最小二乘法 (PLS)^[12]

设有 n 个样品 p 个波长点处吸光度矩阵 $X = (x_{ij})_{n \times p}$ 和该 n 个样品 m 个组分的浓度矩阵 $Y = (y_{ij})_{n \times m}$, PLS 首先对 X 和 Y 分解成如下形式

$$\begin{aligned} X &= TP + E \\ Y &= UQ + F \end{aligned} \quad (1)$$

其中 T 和 U 分别为 X 和 Y 矩阵的得分矩阵, P 和 Q 分别为 X 和 Y 矩阵的载荷矩阵, E 和 F 分别为 X 和 Y 的 PLS 拟合残差矩阵。

然后, 建立回归模型

$$U = TB + E_d \quad (2)$$

其中 E_d 为随机误差矩阵, B 为对角回归系数矩阵。

在预测时, 先求得待测样品 x 的得分 t , 然后由下式得到浓度预测值

$$y = tBQ \quad (3)$$

实际的 PLS 算法中, X 和 Y 矩阵的分解同时进行, 并且将 Y 的信息引入到 X 矩阵的分解过程中, 使得 X 主成分直接与 Y 关联, 从而保证获得最佳的预测模型。

1.3.2 判别式偏最小二乘法 (DPLS)

在 PLS 算法中, 如果用表示类别属性的二进制变量替代矩阵中的浓度变量, 以计算光谱向量与类别向量的相关关系, 称为判别式偏最小二乘法。Y 矩阵类似于如下形式

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

其中, $y_{ij} = 1$ 表示第 i 个样本属于第 j 类; 反之 $y_{ij} = 0$ 表示第 i 个样本不属于第 j 类。建立 PLS 回归模型并设定类别属性鉴别阈值 th 。

在预测过程中, 对未知样本 x , 得到该 PLS 模型预测值 y , 求

$$c = j, y_j = \max(y_1, y_2, \dots, y_m) \quad (4)$$

若 $y_c \geq th$ 则认为该样本属于第 c 类, 反之, 若 $y_c < th$ 则认为该样本不属于已知的任何类。因而, 可以应用 DPLS 进行定性判别分析, 确定未知样本的类归属。

1.3.3 DPLS 特征提取

由 PLS 的算法可知, 所提取的主成分 P 具有对类别综合成分 Q 的最大的相关程度, 而又对类别信息矩阵具有尽可能好的解析能力, 因而, 样品光谱在这些成分上的得分可以用来作为光谱的特征。

DPLS 特征提取的步骤如下:

首先, 对样本数据 X 及类别信息 Y 进行标准化预处理, 得到标准化后的 X_0 及 Y_0 ; 其次, 对 X_0 及 Y_0 进行 PLS 回归, 保存回归过程中的权向量 w_i , w_i 是回归迭代计算过程中 $X'_i Y_i Y'_i X_i$ 的最大特征根对应的特征向量, $T = X_0 W$; 最后, 对未知样本 x , 求其得分 $t = x W$, 作为样本 x 的 DPLS 特征。

t 的维数为 PLS 回归确定的主成分数, 通常该主成分数远小于样品光谱的维数, 因而, DPLS 特征提取还起到了降维的作用。

1.4 Fisher 线性判别分析 (LDA)

设训练样本集 $\{x_i^j\}$ 包含 c 个模式类 $X_1, X_2, \dots, X_c$, 每类有 n_i 个样本。其中, x_i^j 表示第 i 类的第 j 个样本, $i = 1, 2, \dots, c, j = 1, 2, \dots, n_i$ 。样本的类内离散度矩阵 S_W 和类间离散度矩阵 S_B 分别定义为

$$S_W = \frac{1}{N} \sum_{i=1}^c \sum_{x \in X_i} (x - m_i)(x - m_i)' \quad (5)$$

$$S_B = \sum_{i=1}^c (m_i - \bar{m})(m_i - \bar{m})'$$

其中, $m_i = \frac{1}{n_i} \sum_{x \in X_i} x$, 为第 i 类所有样本的均值; $\bar{m} =$

$\frac{1}{c} \sum_{i=1}^c \sum_{j=1}^{n_i} x_i^j$, 为所有样本的均值。

Fisher 线性判别分析基于 Fisher 准则, 寻找将高维样本投影到低维空间的最佳判别投影矩阵 W , 使得类间离散度和类内离散度的比值达到最大

$$J(W_{\text{opt}}) = \arg \max_W \frac{|W' S_B W|}{|W' S_W W|} \quad (6)$$

因而, W 必须满足

$$S_B W = \lambda S_W W \quad (7)$$

义 Y 变量为

如果 S_W 是非奇异的, 就能得到通常的本征问题

$$S_W^{-1} S_B W = \lambda W \quad (8)$$

W 是由 $S_W^{-1} S_B$ 的前 m 个较大特征值对应的特征向量。但在很多模式识别问题中, 样本的维数往往远大于训练样本的总数, 这将造成 S_W 不可逆, 使得式(7)不能方便求解, 这就是小样本问题。可以通过降维来解决小样本问题。本文的试验用 1.3.3 节的 DPLS 特征提取方法来进行降维, 再对降维后的数据进行线性判别分析。

2 试验结果与分析

2.1 玉米的近红外光谱特点

图 1 所示为训练样本集中的随机选取的三个玉米品种的近红外漫反射光谱图。从图 1 可见, 不同玉米品种的近红外光谱形态十分相似, 并混杂在一起。对训练样本集进行 PCA 分析, 第一主成分和第二主成分的累积贡献率达到了 99.97%, 然而, 从图 1 的三个玉米品种近红外光谱在第 1 和第 2 主成分空间分布图上看, 见图 2, 不同品种的样本仍混杂在一起, 品种的分布不存在显著规律性。

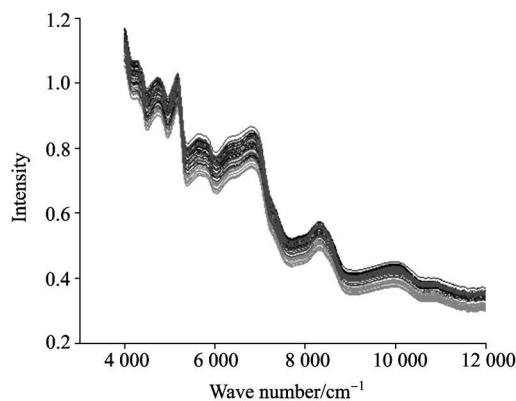


Fig 1 Near infrared spectra of 3 corn species

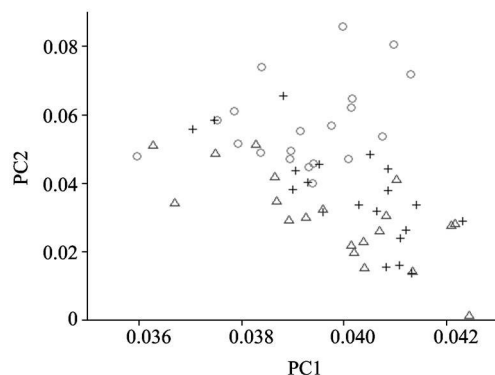


Fig 2 The distribution of 3 corn species samples in PC1 and PC2 space

2.2 DPLS 估计与 DPLS 特征提取

将训练样本集的全区光谱数据作为 PLS 的 X 变量, 定

$$Y_{600 \times 30} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \text{-----} \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \text{-----} \\ 0 & 1 & 0 & \dots & 0 \\ \text{-----} \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

建立 PLS 回归模型, 并保存特征矢量 W 。

对于 DPLS 估计, 若有未知样本 x , 则根据式(3)可得类别预测值 y , 再根据式(4), 确定 x 的类归属。经过交叉验证, 在 PLS 成分数为 28 时, 模型具有最佳预测性能。

而对于 DPLS 特征提取, 并不要求出预测值 y , 而是在建立 PLS 回归模型时, 确定最佳成分数 d , 并保存其权重向量 W , 并将得分矩阵 T 作为训练样本的特征保存。识别过程中, 未知样本 x 的 DPLS 特征即为 $t: t = xW$, 设定阈值, 用最小距离判别方法, 求 t 与所有训练样本特征的距离, 确定其类别归属。

从三个玉米品种近红外光谱在 DPLS 回归模型得分 T 的前两维图上看, 见图 3, 不同品种样本的分布与图 2 相比, 已呈现一定的类别规律性。

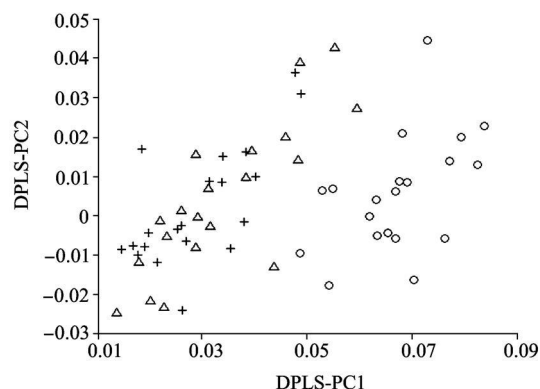


Fig 3 The distribution of 3 corn species samples in DPLS-PC1 and DPLS-PC2 space

2.3 DPLS+ LDA

根据 2.2 节, 一组 1037 维的原始光谱样本 X 经过 DPLS 特征提取, 得到一组 d 维的得分向量 T , 根据式(8)对 T 进行线性判别分析, 求出 W , 经过多次实验确定, 在 DPLS 特征数 $d = 28$ 、LDA 特征数 $m = 26$ 时, 获得最佳识别性能。

从三个玉米品种近红外光谱数据在进行 (DPLS+ LDA) 分析后的前两维图上看, 见图 4, 不同品种样本的分布与图 2、图 3 相比, 类别规律性更为明显。

2.4 结果比较

实验采用交叉验证的方式进行 7 次测试, 实验结果取 7 次测试的平均值。

由 1.1 节可知, 第一测试集的样本品种为已训练品种, 用来测试不同方法对已训练品种的正确识别率; 第二测试集

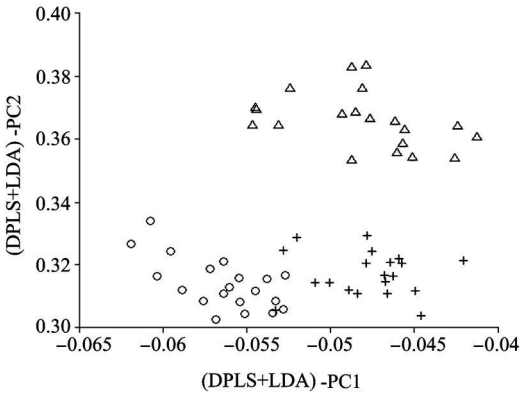


Fig 4 The distribution of 3 corn species samples in (DPLS+LDA)-PC1 and (DPLS+LDA)-PC2 space

的样本不属于任何已训练品种,该样本集用来测试不同方法对于未训练品种的正确拒识率。不同方法的结果如表 1 所示。由表 1 可以看出,设定阈值使得第一测试集都能被正确识别的情况下(正确识别率= 100%),DPLS 预测方法在第二测试集上的正确拒识率只有 23.27%;DPLS 特征提取方法的正确拒识率为 66.04%,远大于 DPLS 预测的结果;而 DPLS+LDA 方法的正确拒识率最高,为 73.47%,可见,DPLS+LDA 方法的推广能力最强。通过调整阈值降低正确识别率,可以在一定程度上提高正确拒识率,如表 1 所示,调整阈值使三种方法的正确识别率降低,相应的正确拒识率都有不同程度的提高。若将阈值调整到使第二测试集能被全

Table 1 Results comparison of the three methods in terms of recognition rate on test set1 and test set2

DPLS 估计方法		DPLS 特征提取方法		DPLS+LDA 方法	
第一测试集的正确识别率/%	第二测试集的正确拒识率/%	第一测试集的正确识别率/%	第二测试集的正确拒识率/%	第一测试集的正确识别率/%	第二测试集的正确拒识率/%
100	23.27	100	66.04	100	73.47
99	45.39	99	75.35	99	81.71
90	79.1	90	98.78	90	99.18
26.29	100	81.62	100	88	100

部正确拒识(正确拒识率= 100%),DPLS 预测方法的正确识别率急剧下降,只有 26.29%;DPLS 特征提取方法的正确识别率为 81.62%;DPLS+LDA 方法的正确识别率达到了 88%。

DPLS 特征提取方法的性能优于 DPLS 预测方法,这是由于在建立 DPLS 回归模型时,由于存在光谱矩阵 X 的拟合残差矩阵 E 及类别矩阵 Y 的拟合残差矩阵 F ,因而,DPLS 模型对未知光谱的预测输出受到 E 和 F 的双重影响。DPLS 特征提取方法在建立 DPLS 回归模型后,只利用了光谱矩阵 X 的 PLS 拟合成分得分,保留精确的类别信息,用最小距离法做分类判别,获得了更高的识别率。而 DPLS+LDA 方法在 DPLS 特征提取的基础上进行 LDA 运算,使类间离散度和类内离散度的比值最大化,进一步提高了识别率,因而又优于 DPLS 特征提取方法。

调整阈值,可以使得在某一阈值下,正确识别率与正确拒识率相等,称为等识率。表 2 为不同方法的等识率比较。可见 DPLS+LDA 方法的等识率最高,为 96.18%,比 DPLS 估计方法提高了 12.65%。

Table 1 Recognition rate comparison of the three methods/ %

DPLS 估计方法	DPLS 特征提取方法	DPLS+LDA 方法
85.38	95.76	96.18

3 结 论

应用近红外光谱可以快速、准确地对玉米品种进行鉴别。采用 DPLS 特征提取而不使用 DPLS 的类别估计输出,能避免对未知样本类别估计的误差。LDA 利用了样本光谱的类别信息,寻找能对样本光谱进行有效分类的变换空间,因而是有效的特征提取算法。本文提出了一种基于 DPLS 特征提取的 LDA 方法,并应用在玉米近红外光谱的定性分析中,取得了优异的鉴别效果。因此,用 DPLS+LDA 方法和近红外光谱技术对玉米品种鉴别是可行的,其鉴别准确度高,可以实现简易、快速、无损且有效的玉米定性分析,有良好的应用前景。本研究亦可作为其他物质定性分析的参考。

References

[1] LUO Cheng-gang, HA Jun-li, ZHOU Yi-he, et al(罗成刚,哈君利,周义和,等). Acta Tabacaria Sinica(中国烟草学报), 1999, 5(3): 42.

[2] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄,赵龙莲,韩东海,等). Foundation and Application of Near-Infrared Spectroscopy Analysis(近红外光谱分析基础与应用). Beijing: China Light Industry Press(北京:中国轻工业出版社), 2005.

[3] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍,袁洪福,徐广通,等). Modern Near-Infrared Spectroscopy Analytical Technology (Second Edition)(现代近红外光谱分析技术,第2版). Beijing: China Petrochemical Press(北京:中国石化出版社), 2007.

[4] CHEN Quan-sheng, ZHAO Jie-wen, ZHANG Hai-dong, et al(陈全胜,赵杰文,张海东,等). Food Science(食品科学), 2006, 27(4): 186.

[5] WU Wen-jin, WANG Hong-wu, CHEN Shao-jiang, et al(邬文锦,王红武,陈绍江,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2010, 30(5): 1248.

[6] ZHAI Ya-feng, SU Qian, WU Wen-jin, et al(翟亚锋,苏谦,邬文锦,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2010, 30(4): 924.

- [7] Duda Richard O, Hart Peter E, et al. Pattern Classification(模式分类). Beijing: China Machine Press(北京: 机械工业出版社), 2003.
- [8] Zhao W, Chellappa R, et al. Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. 336.
- [9] Kawatani T, Shimizu H. Proceedings of Fourteenth International Conference on Pattern Recognition, 1998, 2: 1301.
- [10] Fan Bin, Lei Zhen, et al. Proceedings of 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008. 1.
- [11] LU Wan-zhen, YUAN Hong-fu, CHU Xiao-li(陆婉珍, 袁洪福, 褚小立). Near Infrared Spectrometer(近红外光谱仪器). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2010.
- [12] Svante Wold, Michael Sjostrom, et al. Chemometrics and Intelligent Laboratory Systems, 2001, 58: 109.

Application of DPLS-Based LDA in Corn Qualitative Near Infrared Spectroscopy Analysis

QIN Hong, WANG Huirong, LI Weijun*, JIN Xiaoxian

Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China

Abstract NIR technology is a rapid, nondestructive and user-friendly method ideally suited for Qualitative analysis. In this paper the authors present the use of discriminant partial least Squares (DPLS)-based linear discriminant analysis (LDA) in corn qualitative near infrared spectroscopy analysis. Firstly, a training set including 30 corn varieties (each variety has 20 samples) was used to build the DPLS regression model, and 28 principal components (DPLS-PCs) were obtained from original spectrum. Secondly, the DPLS-PCs scores of the training set were extracted as DPLS features. Thirdly, LDA was applied to the DPLS features, determining 26 principal components (LDA-PCs). A test sample was first projected onto the DPLS-PCs and then onto the LDA-PCs, and finally 26 DPLS+LDA features were obtained. The recognition results were obtained by minimum distance classifier. DPLS+LDA method achieved 96.18% recognition rate, while traditional DPLS regression method and DPLS feature extraction method only achieved 85.38% and 95.76% recognition rate respectively. The experiment results indicated that DPLS+LDA method is with better generalization ability compared with traditional DPLS regression method and NIRS analysis by DPLS+LDA method is an efficient way to discriminate corn species.

Keywords Near infrared spectroscopy; Linear discriminant analysis; Discriminant partial least square; Qualitative analysis; Corn

(Received Sep. 21, 2010; accepted Dec. 18, 2010)

* Corresponding author