基于核局部保持投影的近红外光谱玉米单倍体识别研究

刘文杰[1,2]，李卫军[1,2*]，覃 鸿[1,2]，李浩光[1,2]，宁 欣[1,2]

1. 中国科学院半导体研究所，高速电路与神经网络实验室，北京 100083
2. 中国科学院大学微电子学院，北京 100049

摘 要 实现快速、精确地鉴别玉米单倍体籽粒对玉米单倍体育种技术十分重要。近红外光谱分析技术可在线分析、监测，且无损、分析速度快、操作简便、测试成本低，对实现自动化的大规模鉴定并分拣玉米单倍体非常有帮助。通过美国JDSU的近红外光谱仪进行玉米近红外光谱的数据采集，交叉采集玉米单倍体、多倍体数据。数据处理时，将数据分为训练集和测试集两部分。依次对数据做预处理以消除噪声影响，做核变换将其投射到更高维度空间中增强可分性并进行特征提取，最后建立分类模型鉴别分析。分别统计采用不同的特征提取算法并建立模型鉴别测试的正确识别率。实验结果表明，采用核局部保持投影(KLPP)的特征提取算法的正确识别率更高、稳定性更好，在两组测试集上的正确识别率的均值分别达到95.71%和96.43%。通过分析可以得出，玉米种子的近红外光谱数据经过非线性变换(为高斯核变换)投影到更高维度的空间后，表现出更易于分类的分布特点，保持数据的局部特性也更利于后续的分类。这为玉米单倍体鉴定进一步研究提供了新的方向。

关键词 近红外光谱；特征提取；核局部保持投影(KLPP)；玉米单倍体

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)08-2574-04

引 言

玉米单倍体育种技术以其育种周期短、效率高、操作简便等特点，在玉米的育种中有非常广阔的前景[1-3]。然而在自然条件下产生玉米单倍体概率非常低[4]，因此如何快速、精准地鉴别出玉米单倍体对玉米单倍体育种技术显得至关重要。

目前，国内外的玉米单倍体检测鉴定的方法很多，主要有：遗传标记法、形态学方法、解剖学方法、射线照射法、分子生物学方法等[5-7]。而这些方法鉴定时间长、过程复杂、鉴定成本高、有损耗且需要大量的专业人才，无法实现玉米单倍体鉴定的自动化。近红外光谱分析技术可在线分析、监测，且无损、分析速度快、操作简便、测试成本低[8]。这对于实现玉米单倍体鉴定的自动化分拣非常有帮助。

将近红外光谱分析技术与玉米单倍体鉴定相结合，并采用核局部保持投影(kernel locality preserving projection，KLPP)[9-14]的特征提取算法，取得了较好的效果。通过实验说明可以采用分析近红外光谱的方法进行玉米单倍体鉴定，而且玉米种子的近红外光谱数据经过非线性变换后在更高维度的空间中表现出更易于分类的分布特点，保持数据的局部特性也更利于后续的分类。这为玉米单倍体鉴定进一步研究提供了新的方向。

1 实验部分

1.1 仪器设备

实验采用美国JDSU公司的MicroNIR-1700系列的微型近红外光谱仪。波长范围：950～1 650 nm，分辨率：12.5 nm，测量时间(典型值)：0.25 s。

1.2 样品与光谱获取

实验所用的玉米品种为：国家玉米改良中心提供的导入Navajo遗传标记后杂交诱导产生的郑单958玉米单倍体和多倍体籽粒。

近红外光谱数据的获取：为了验证模型的鲁棒性，分别于2014年7月2日和7月3日的上午下午，在外部条件(光源电压、测试样本与光源的距离等)不变的情况下，于室内采用漫透射的方式采集的玉米种子单倍体、多倍体的近红外

收稿日期：2016-06-13，**修订日期**：2016-12-18

基金项目：国家重大科学仪器设备开发专项(2014YQ470377)资助

作者简介：刘文杰，1989年生，中国科学院半导体研究所博士研究生 e-mail：liuwenjie@semi.ac.cn

*通讯联系人 e-mail：wjli@semi.ac.cn

光谱数据。在采集近红外光谱数据时，采用单倍体和多倍体交叉采集的方法，这是为了减小仪器参数的漂移对实验的影响，而且实际生产检测中无法事先知道待检测玉米种子的单倍体、多倍体情况，更接近自动化检测。每次采集的玉米种子样本数为 50 个，每次每个样本采集一条光谱，即每次采集的玉米单倍体、多倍体的光谱数目为 50 条。

1.3　方法

将近红外光谱数据，分为训练数据和测试数据。数据处理流程如图 1 所示，在训练阶段，首先对数据预处理，然后对训练数据进行特征提取得到变换矩阵 $\boldsymbol{W}$，最后采用支持向量机建立分类模型；在测试阶段，先对测试数据进行预处理，再利用变换矩阵 $\boldsymbol{W}$ 进行特征提取，最后用训练得到的分类模型进行分类，得到数据的分类结果。在进行数据预处理时，采用平滑(smoothing)、一阶导(first derivative，FD)和矢量归一化(vector normalization，VN)；进行数据分类建模时，采用的是支持向量机(support vector machines，SVM)。

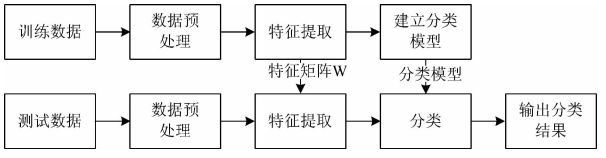

图 1　实验数据处理流程图

Fig. 1　Flow chart of experimental data processing

1.3.1　局部保持投影算法

局部保持投影(locality preserving projection，LPP)[12-16]是一种能很好地保持数据局部结构的线性特征提取降维方法。LPP 利用数据间的相似性构建数据集的邻接图，来保持数据的内在几何性质和局部结构，更加注意数据的局部结构的信息。

设原始样本空间 X 有 n 个样本，每个样本 x_i 为 m 维的数据，则 $X=[x_1, x_2, \cdots, x_n]$，$x_i=[x_{i1}, x_{i2}, \cdots, x_{im}]$，$i=1, 2, \cdots, n$。通过矩阵变换 $\boldsymbol{W}$，将原始样本 X 映射到新的保持了局部结构的低维特征空间 $Y=[y_1, y_2, \cdots, y_n]$，即 $y_i=\boldsymbol{W}^{\mathrm{T}}x_i$。LPP 算法的目标函数为

$$\min\sum_{ij}(y_i-y_j)^2\boldsymbol{S}_{ij} \tag{1}$$

其中，矩阵 $\boldsymbol{S}$ 为原始数据 X 的邻接图，其计算方式有如下 2 种

$$\boldsymbol{S}_{ij}=\begin{cases}\mathrm{e}^{-\frac{\|x_i-y_i\|^2}{t}} & \text{如果 } \|x_i-y_i\|^2<\varepsilon\\ 0 & \text{其他}\end{cases} \tag{2}$$

或者，

$$\boldsymbol{S}_{ij}=\begin{cases}\mathrm{e}^{-\frac{\|x_i-y_i\|^2}{t}} & \text{如果 } x_i \text{ 与 } y_i \text{ 互为 } k \text{ 最近邻}\\ 0 & \text{其他}\end{cases} \tag{3}$$

其中，权值 $\mathrm{e}^{-\frac{\|x_i-y_i\|^2}{t}}$ 也可直接设定为 1，由 $\boldsymbol{S}$ 的定义式可知，$\boldsymbol{S}$ 为对称矩阵。

经过对目标函数进行变换，最终可将目标函数表示为求解(4)式的最小特征值和特征向量。

$$\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{w}=\lambda\boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{w} \tag{4}$$

其中，$\boldsymbol{D}$ 为对角阵，对角线上的元素 $D_{ii}=\sum_j\boldsymbol{S}_{ij}$，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ 为拉普拉斯矩阵。

假定求得的按升序排列后的 l 个特征值为 $\lambda_0<\lambda_1<\cdots<\lambda_{l-1}$，其对应的特征向量分别为 W_0，W_1，$\cdots$，W_{l-1}。取其前 k 个特征向量组成最佳投影矩阵。

1.3.2　核局部保持投影

KLPP 是一种非线性特征提取方法，而且保持了数据的局部结构特性。KLPP 将原数据投影到更高维度的空间中，并保留了数据的局部结构特性，从而使其更容易分类。

假设原始空间为欧式空间 R^n，$\mathscr{H}$ 为映射后的 Hilbert 泛函空间，存在非线性映射 $\Phi: R^n\rightarrow\mathscr{H}$，使得任意 $x_i\in R^n$ 可映射为 $\Phi(X)=[\Phi(x_1), \Phi(x_2), \cdots, \Phi(x_n)]$。

定义核函数

$$K=\Phi(X)^{\mathrm{T}}\Phi(x)=[k(x_i, x_j)]_{m\times m} \tag{5}$$

其中，

$$k(x_i, x_j)=\langle\Phi(x_i)^{\mathrm{T}}, \Phi(x_j)\rangle=\Phi(x_i)^{\mathrm{T}}\Phi(x_j) \tag{6}$$

那么在 $\mathscr{H}$ 空间中，式(4)可写为

$$\Phi(X)L\Phi(X)^{\mathrm{T}}\alpha=\lambda\Phi(X)D\Phi(X)^{\mathrm{T}}\alpha \tag{7}$$

由式(5)，式(6)和式(7)可得到

$$\boldsymbol{KLK}\alpha=\lambda\boldsymbol{KDK}\alpha \tag{8}$$

其中，$\boldsymbol{K}$ 为数据经过核变换后的矩阵。

本文采用的核函数为高斯核函数，也称作径向基核函数(radial basis function)

$$\boldsymbol{K}(x, y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

求得式(8)中的 l 个按升序排列的特征值和其对应的特征向量，取前 k 个特征向量组成的投影矩阵即为所求的最佳投影矩阵。

2　结果与讨论

分别选取每天上午单倍体、多倍体各 30 条近红外光谱进行特征提取、建立分类模型，采用每天未参与训练的上午剩余的 20 条光谱和下午的 50 条光谱进行统一测试。

在实验中：(1)数据的预处理和建模分类部分，采用的方法都相同；(2)数据的特征提取部分，采用了不同的特征提取算法进行对比实验，采用的特征提取算法包括：主成分分析(principal component analysis，PCA)，正交化线性判别分析(orthogonal linear discriminant analysis，OLDA)，PCA＋OLDA，LPP，核主成分分析(kernel PCA，KPCA)和 KLPP。其中 PCA，OLDA，PCA＋OLDA 和 KPCA 为全局特征提取算法，对于数据的全局特征有较好的处理作用；LPP 和 KLPP 为局部保持特征提取算法，对数据的局部特性有很好的保持作用；PCA，OLDA，PCA＋OLDA 和 LPP 为较低维度空间上的特征提取算法；KPCA 和 KLPP 通过核变换将数据映射到更高维度上，是高维空间中的特征提取算法。采用不同的特征提取算法做对比实验，实验结果如表 1 所示。

由表1的实验数据可以得出：(1)基于KLPP特征提取

表 1 基于不同特征提取算法的识别率结果统计

Tabel 1 The recognition rate results statistics of extraction algorithm based on different characteristics

特征提取算法	日期	上午			下午			上午+下午
		单倍体	多倍体	均值	单倍体	多倍体	均值	均值
PCA	7.2	0.85	0.95	0.9	0.78	0.96	0.87	0.878 6
	7.3	0.75	0.85	0.8	0.82	0.92	0.87	0.85
OLDA	7.2	0.9	0.9	0.9	0.86	0.72	0.79	0.821 4
	7.3	0.65	0.65	0.65	0.86	0.38	0.62	0.628 6
PCA+OLDA	7.2	0.95	0.95	0.95	0.8	0.94	0.87	0.892 9
	7.3	0.8	0.85	0.825	0.88	0.82	0.85	0.842 9
LPP	7.2	0.95	0.95	0.95	0.84	0.96	0.9	0.914 3
	7.3	0.8	0.85	0.825	0.86	0.88	0.87	0.857 1
KPCA	7.2	0.9	0.95	0.925	0.98	0.94	0.96	0.95
	7.3	0.95	0.95	0.95	0.92	0.94	0.93	0.935 7
KLPP	7.2	0.95	1	0.975	0.96	0.94	0.95	0.957 1
	7.3	0.95	0.95	0.95	0.98	0.96	0.97	0.964 3

算法的识别率(7 月 2 日为 95.71%、7 月 3 日为 96.43%)和基于 KPCA 特征提取算法的识别率(7 月 2 日为 95%、7 月 3 日为 93.57%)明显高于基于 LPP 特征提取算法的识别率(7 月 2 日为 91.43%、7 月 3 日为 85.71%)、基于 PCA 特征提取算法的识别率(7 月 2 日为 87.86%、7 月 3 日为 85%)、基于 OLDA 特征提取算法的识别率(7 月 2 日为 82.14%、7 月 3 日为 62.86%)和基于 PCA+OLDA 特征提取算法的识别率(7 月 2 日为 89.29%、7 月 3 日为 84.29%)。(2)基于 KLPP 特征提取算法的识别率和基于 LPP 算法的识别率略高于基于 KPCA 特征提取算法的识别率和基于 PCA，OLDA 和 PCA+OLDA 特征提取算法的识别率。(3)基于 KLPP 和 KPCA 特征提取算法识别率的波动性较小，且明显小于基于 PCA，OLDA，PCA+OLDA 和 LPP 特征提取算法的识别率的波动性。

通过对上述实验数据的分析，可以得出：(1)基于核函数特征提取算法的识别率要明显高于传统特征提取算法的识别率，识别率较高；(2)基于核函数特征提取算法的识别率的稳定性较高；(3)采用局部保持特征提取算法的识别率要略高于基于全局特征提取算法的识别率，说明基于数据的局部结构特性的特征提取算法更为有效。基于以上分析，可以得出玉米近红外光谱数据在线性空间内难以找出较好的分类界面；在保持了原始数据的局部结构特性的情况下，通过核变换将其非线性的映射到高维空间中后，数据展现出更易于分类的分布特性。

3 结 论

采用多种算法做特征提取，通过对比试验得出：基于核局部保持投影的特征提取算法能有效的改善玉米种子近红外光谱原始数据的分布方式，使其更易于分类，相比于未采用核变换和未保持局部结构的特征提取算法，识别准确率高，达到 95%以上，且稳定性较好。

本文从保持数据的局部结构特性和对数据进行非线性变换即核变换两方面验证了其特征提取的有效性，为玉米种子单倍体鉴别研究提供了新的方向。

References

[1] ZHANG Qiang(张 强). Heilongjiang Agricultural Sciences(黑龙江农业科学), 2014, (9): 150.

[2] CHEN Shao-jiang, SONG Tong-ming(陈绍江, 宋同明). Acta Agronomica Sinica(作物学报), 2003, 29(4): 587.

[3] DU He-wei, DAI Jing-rui, LI Jian-sheng(杜何为, 戴景瑞, 李建生). Journal of Maize Sciences(玉米科学), 2010, (6).

[4] CAI Zhuo, XU Guo-liang(才 卓, 徐国良). Journal of Maize Sciences(玉米科学), 2014, (1): 1.

[5] WEI Chang-song, XU Gui-ming, TIAN Pu-huan(魏昌松, 许贵明, 田甫焕). Crops(作物杂志), 2014, (6).

[6] LI Xiang-qun, SONG Bing, FU Yong-ping(李向群, 宋 冰, 付永平). Seed World(种子世界), 2014, (7): 22.

[7] Hartwig H Geiger, G Andrés Gordillo, Silvia Koch. Crop Science, 2013, 53: 2313.

[8] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄, 赵龙莲, 韩东海, 等). Fundamentals and Applications of Near Infrared Spectroscopy(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.

[9] Deng Xiaogang, Tian Xuemin. Chinese Journal of Chemical Engineering, 2013, 21(2): 163.

[10] Su Yu-Chuan, Chiu Tzu-Hsuan, Kuo Yin-His. Multimedia, IEEE Transactions on, 2014, 16(6): 1645.

[11] Luo Lijia, Bao Shiyi, Mao Jianfeng. Journal of Process Control, 2016, 38: 11.

[12] Wonga W K, Author Vitae, ZhaobAuthor Vitae H T. Pattern Recognition, 2012, 45(1): 186.

[13] Shikkenawis Gitam, Mitra K Suman. Neurocomputing, 2016, 173(2): 196.
[14] Jiang Rui, Fu Weijie, Li Wen. Neurocomputing, 2016, 187: 109.
[15] Zhong Fujin, Li Defang, Zhang Jiashu. Journal of Visual Communication and Image Representation, 2014, 25(7): 1676.
[16] Yu Guoxian. Neurocomputing, 2011, 74(4): 598.

Research on Identifying Maize Haploid Seeds Using Near Infrared Spectroscopy Based on Kernel Locality Preserving Projection

LIU Wen-jie[1,2], LI Wei-jun[1,2*], QIN Hong[1,2], LI Hao-guang[1,2], NING Xin[1,2]
1. Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
2. School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Haploid identification plays a key role in the field of maize-haploid breeding. To achieve mass and automated identification, Near-infrared Spectroscopy (NIRS) Analysis Technology is widely used. Its advantages include online monitoring, rapid analysis, easy operation, lossless process, cost-effectiveness, etc. At the beginning of the experiment, NIRS data of haploid and polyploidy maize seeds are cross collected via JDSU's near-infrared spectrometer. To enhance validity, this experiment encompasses a testing set of data besides a training set. After pre-processing, experiment data is subsequently mapped in a higher-dimensional space to enhance its divisibility, and haploid feature is extracted. Then the experiment establishes identification models to predict whether maize seeds are haploid. It needs to point out that the experiment applies different feature extraction algorithms, thus different identification models are established accordingly. The experiment results show that the feature extraction algorithm of Kernel Locality Preserving Projection (KLPP) guarantees accurate recognition in a more stable way. Recognition rate of testing set and training set reaches up to 95.71% and 96.43%. The above experiment proves that NIRS data of maize seeds can be classified more effectively and accurately through non-linear transformation (Gaussian kernel transform in this experiment) and high-dimensional spatial mapping. The above process also maintains partial characteristics of NIRS data. Therefore, this paper may provide some new idea and method for Maize Haploid Identification technology.

Keywords Near infrared spectroscopy; Feature extraction; Kernel locality preserving projection (KLPP); Maize haploid

(Received Jun. 13, 2016; accepted Dec. 18, 2016)

* Corresponding author