

doi:10.6041/j.issn.1000-1298.2015.S0.022

基于近红外光谱的核桃仁品种快速分类方法^{*}

马文强^{1,2} 张漫¹ 李忠新²

(1. 中国农业大学现代精细农业系统集成研究教育部重点实验室, 北京 100083;

2. 新疆农业科学院农业机械化研究所, 乌鲁木齐 830091)

摘要: 采用傅里叶变换近红外光谱仪,采集了4个不同品种的200份核桃仁样本的近红外漫反射光谱,建立了核桃仁品种分类模型。光谱范围为 $3\ 800\sim 9\ 600\text{ cm}^{-1}$,预处理方法采用多元散射校正法和标准正态化方法;通过主成分分析法优选出5个主成分因子,光谱信息累计贡献率达到99.21%;采用随机抽取法建立建模集和验证集,以主成分因子为输入变量,建立了基于支持向量机分类模型,并采用网格搜索法对RBF核函数参数 λ 和 δ 进行寻优。分析结果表明,建立的核桃仁分类识别模型对4个核桃仁品种的总体正确识别率达到96%,为核桃仁品种的快速无损识别提供了一种可行的方法。

关键词: 核桃仁 品种分类 支持向量机 近红外光谱

中图分类号: O657.33; S664.1 **文献标识码:** A **文章编号:** 1000-1298(2015)S0-0128-06

Rapid Classification Method of Walnut Kernel Varieties Based on Near-infrared Diffuse Reflectance Spectra

Ma Wenqiang^{1,2} Zhang Man¹ Li Zhongxin²

(1. Key Laboratory of Modern Precision Agriculture System Integration Research, Ministry of Education, China Agricultural University, Beijing 100083, China

2. Agricultural Mechanization Institute, Xinjiang Academy of Agricultural Sciences, Urumqi 830091, China)

Abstract: Walnut is an important dry fruit and woody oil crop in China, and it has significant meaning to establish a rapid, nondestructive testing method for identification and classification of walnut kernel varieties in walnut processing industry. Near-infrared diffuse reflection spectroscopies of 200 walnut samples of four species were adopted to establish models for rapid and nondestructive classification. The spectral region of walnut samples was ranged from $3\ 800\text{ cm}^{-1}$ to $9\ 600\text{ cm}^{-1}$. The spectra data of walnut were processed using the multiplicative scatter correction (MSC) and the standard normalized variate (SNV) methods. Principal component analysis (PCA) was used to reduce the dimensionality of spectra data. The cumulative contribution rate of the first five main components reached 99.21%, which were selected as variables for modeling. Totally 100 walnut samples were selected as training set by random sampling method. The NIR classification model of walnut kernel varieties was built based on support vector machine (SVM) method, and grid search method was used for searching the best parameter. The built model was tested by the rest 100 walnut samples of four species, and the results showed that the correct recognition rate of the model reached 96%. The analyzed results indicated that the NIR classification model could provide a feasible method for rapid and nondestructive identification of walnut kernel varieties.

Key words: Walnut kernel Variety classification Support vector machine Near infrared spectrum

收稿日期: 2015-10-28 修回日期: 2015-11-20

^{*} 新疆自治区成果转化项目(201454122)

作者简介: 马文强, 博士生, 新疆农业科学院助理研究员, 主要从事农产品智能化检测技术研究, E-mail: mwq4530@163.com

通讯作者: 张漫, 副教授, 博士生导师, 主要从事智能检测技术研究, E-mail: cauzm@cau.edu.cn

引言

中国核桃种植面积和产量均居世界首位,2013 年仅新疆地区核桃产量已达 40.28 万 t^[1]。我国核桃种植区域分布广泛、品种繁多,不同品种的核桃仁加工适应性各不相同^[2]。目前,核桃仁加工过程中,主要依靠人工或机械方式根据大小、色泽、完整程度等指标对其进行分级处理,难以对核桃仁的品种进行分辨^[3-4]。因此,对不同品种的核桃仁进行快速、准确的鉴别是核桃仁生产加工中的一项亟待解决的问题。

近红外光谱(Near infrared spectroscopy, NIRS)技术可以对样品进行定性或定量分析,具有分析速度快、结果稳定、对物料无破坏性、适合在线处理等特点,非常适合应用于农产品品质检测、分级等方面,目前已在蔬菜、水果、肉类、谷物等多种农产品中得到了广泛的应用^[5-9]。农产品的近红外漫反射光谱包含了大量农产品内部有机物官能团信息,通过合理的数据处理,可将其应用于农产品品种产地、品质分类等方面^[10-13]。Eva 等^[14]利用近红外漫反射光谱对杏仁内氰糖苷进行无损检测,并实现了苦杏仁与甜杏仁的分类。Jiang 等^[15]采用了 K 近邻算法、BP 神经网络、支持向量机 3 种方式成功实现了对 4 种不同品种大豆近红外光谱的分类。Khanmohammadi 等^[16]利用近红外反射光谱,通过最小二乘支持向量机模型对来自西班牙不同产地的柿子实现了分类。郑田甜等^[17]采集了 3 种不同花生种子的近红外光谱,通过线性分类模型实现了花生种子的分类。目前将近红外光谱技术应用于核桃仁无损检测和分类方面的研究,国内外尚未见相关的研究报道。

本文分析来自 4 个不同品种的核桃仁近红外漫反射光谱信息,采用主成分分析法提取光谱特征信息后,建立基于支持向量机的核桃仁品种分类识别模型,探讨建立的模型对不同品种核桃仁的分类效果,初步探索采用近红外光谱分析技术进行核桃仁品种快速无损分类识别的新方法。

1 材料与方法

1.1 实验材料

实验用核桃样品购自北京新发地农贸市场,选用了 4 个核桃品种,分别是:新疆“温 185”核桃、云南泡核桃、新疆铁核桃和东北山核桃,选购的带壳核桃含水率约为 6%,其核桃仁样本如图 1 所示。其中,新疆“温 185”核桃属于新疆地区主栽核桃品种,主要栽种于南疆地区;新疆铁核桃分布于天山山区,

属于核桃野生原始群类;云南泡核桃主要产于云南海拔高度 1 500 m 以上的山林地区,为云南地区主栽品种;东北山核桃为核桃属核桃楸组作物,主要分布于长白山、大小兴安岭山区,为野生品种^[18-19]。

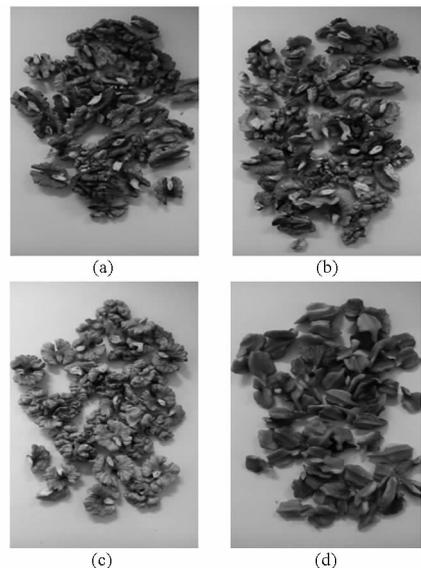


图 1 核桃仁样本

Fig. 1 Walnut kernel sample

(a) 新疆“温 185”核桃 (b) 新疆铁核桃
(c) 云南泡核桃 (d) 东北山核桃

采用手工破壳方式对 4 种核桃进行破壳取仁,并尽可能保持仁的完整性。采用高露仁率(完整程度在 1/2 以上的核桃仁占总核桃仁质量的百分比)描述核桃仁样品的完整性,破壳后的“温 185”核桃、新疆铁核桃、云南泡核桃和东北山核桃高露仁率分别为 94.7%、69.8%、90.5% 和 13.9%。

1.2 实验数据获取

实验仪器使用德国 BRUKER 公司的 MPA 型傅里叶变换漫反射近红外光谱仪,如图 2 所示。实验时取每种核桃仁各 500 g 采用最大积分球漫反射方式进行光谱测量,样品池为内径 51 mm 旋转样品池,测量光谱范围为 3 800 ~ 9 600 cm^{-1} ,分辨率为 8 cm^{-1} ,扫描次数为 64。测量步骤为:打开



图 2 BRUKER MPA 型漫反射近红外光谱仪

Fig. 2 BRUKER MPA NIR spectrometer

光谱仪预热 30 min 以确保仪器的测量稳定性,进行测量参数的设置;每种核桃仁取 15 ~ 20 g 样品放入样品池进行一次测量;测量后放回该品种的总体样品中进行充分混合,再重新抽取下一组测量样品。每种核桃仁样品反复测量 50 次,共获取 200 组光谱数据。

1.3 光谱数据预处理

为了消除散射对光谱信息的影响,首先采用多元散射校正方法对光谱数据进行处理。多元散射校正就是通过计算所有光谱的平均光谱,并以此作为标准光谱对原始光谱的线性平移量和倾斜量进行修正的方法。通过多元散射校正可以有效消除散射影响,强化与成分含量相关的光谱吸收信息,提高光谱的信噪比^[20]。

为进一步降低样品颗粒大小和表面散射的影响,对多元散射校正后的光谱数据进行标准正态化处理(Standard normal variate transformation, SNV)^[21]。

1.4 主成分分析

核桃仁全光谱信息数据量较大,且数据间相关性较强,因此采用主成分分析方法(Principal component analysis, PAC)对预处理后的光谱数据进行降维处理。PAC 通过多元统计分析方法将具有相关关系的多维变量线性组合后,降维为互不相关的几个可反映原有信息的综合变量。

1.5 分类模型的建立

分类模型以支持向量机算法为基础进行建模。支持向量机是一种以结构风险最小化为基础思想的模式识别算法,在小样本数据集的分类中有显著优势^[22-23]。在支持向量机的核函数选择方面,径向基函数(Radial basis function, RBF)核在紧致性和模型复杂度方面的综合优势较好^[24]。RBF 核函数具有 λ 和 δ 两个参数,其中 λ 用于权衡模型的精度和复杂度, δ 用于调节核函数的紧致性。

对于 λ 和 δ 的寻优,本文采用网格搜索算法完成。具体做法是在设定好的网格范围内,遍历所有

λ 和 δ 的组合建立 SVM 分类模型,并采用训练集数据计算模型精度,最终选出达到最高模型精度的 λ 和 δ 组合。参数 λ 的搜索范围设定为 $[2^8, 2^{18}]$,参数 δ 的搜索范围为 $[2^{-18}, 2^{-8}]$,网格节点数为 19×19 ,采用 3 折交叉验证过程进行优化。采用 Matlab R2014a 软件和 Libsvm 工具箱完成建模仿真计算。

2 结果分析与讨论

2.1 光谱数据预处理

核桃仁的主要成分包括脂肪、蛋白质和少量糖分以及多种微量元素。从本质上来看,应用近红外光谱进行核桃仁品种分类是由于不同品种核桃仁内部组成成分存在差异,通过对其近红外光谱信息的处理和分类从而达到对核桃仁品种进行分类的目的。图 3a 为 4 种共 200 组核桃仁的近红外漫反射吸光度原始光谱图,核桃仁的近红外光谱吸收范围主要分布在 $4\ 000 \sim 9\ 000\ \text{cm}^{-1}$ 频率范围内。由图 3a 可以看出 4 种核桃仁吸光度原始光谱趋势相似,但吸收峰并不明显,不利于进行分类处理。

图 3b 为采用多元散射校正和标准正态化组合方法对原始光谱信息预处理后的光谱图,可以看出其中光谱图中基线偏移量得到较好修正,并且出现较为明显的吸收峰:在 $4\ 338\ \text{cm}^{-1}$ 和 $4\ 264\ \text{cm}^{-1}$ 附近为脂肪吸收带, $4\ 587\ \text{cm}^{-1}$ 和 $4\ 878\ \text{cm}^{-1}$ 附近为蛋白质吸收带, $5\ 372\ \text{cm}^{-1}$ 处吸收带由 $\text{C}=\text{O}$ 键二倍频伸缩振动引起, $5\ 797$ 、 $5\ 935$ 、 $6\ 079\ \text{cm}^{-1}$ 处吸收带由 $\text{C}-\text{H}$ 键一倍频伸缩振动引起, $8\ 547\ \text{cm}^{-1}$ 处吸收带由 $\text{C}-\text{H}$ 键二倍频伸缩振动引起。

2.2 主成分分析

对预处理后的光谱信息进行主成分分析,前 5 个主成分的方差贡献率和累积贡献率如表 1 所示,第 1、2、3、4、5 主成分方差贡献率分别为 87.45%、8.94%、1.52%、1.04% 和 0.26%,这前 5 个主成分的累积贡献率为 99.21%,可以有效地代表原有光谱信息。

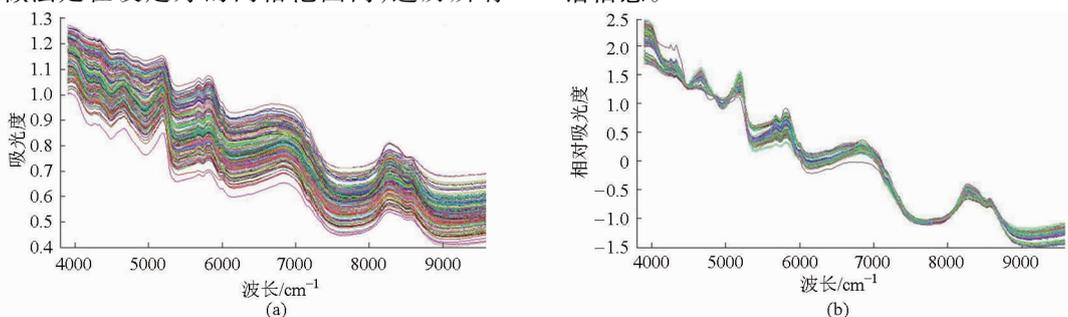


图 3 核桃仁近红外光谱预处理

Fig. 3 Preprocessing results of walnut kernel NIR spectra

(a) 原始光谱 (b) 预处理后的光谱

表1 核桃仁近红外光谱数据主成分分析数据

Tab.1 Principal component analysis histogram of walnut kernel NIR spectra data

主成分	特征值	方差贡献率/%	累积贡献率/%
1	7.441 0	87.45	87.45
2	0.761 0	8.94	96.39
3	0.129 3	1.52	97.91
4	0.088 5	1.04	98.95
5	0.021 6	0.26	99.21

由于第1主成分和第2主成分的方差贡献率远大于其他主成分,为了初步确定4个核桃仁品种间的差异,本文以第1主成分和第2主成分为变量绘制散点分布图,如图4所示。

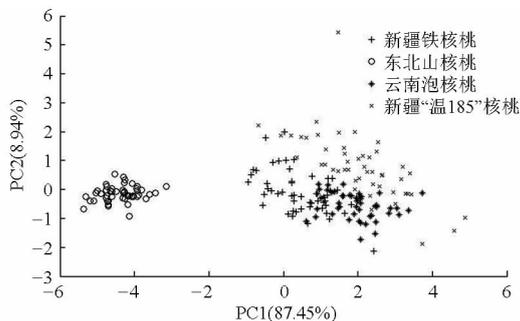


图4 第1和第2主成分散点分布图

Fig.4 The first and second principal components scatter diagram

由主成分散点分布图可以看出,东北山核桃的主成分散点分布与其他3种核桃具有显著的差别,而其他3种核桃的主成分散点分布则无明显的差异,说明东北山核桃与其他3种核桃在品种上存在较为明显的差异,而其他3种核桃仁的分类需要进一步采用分类识别模型。

2.3 分类模型建立

为了研究利用近红外漫反射光谱进行核桃仁品种分类的可行性,以4种核桃仁光谱信息的前5个主成分为输入变量,采用支持向量机算法建立多分类识别模型。通过随机抽样的方法从每种核桃仁光谱的前5个主成分数据中随机抽取25组数据,共100组数据作为模型训练集;每个核桃仁品种剩余的25组光谱数据形成4个验证集,用于验证分类模型对每个核桃仁品种的分类率;并将4个验证集合并后进行随机排列,用于验证分类模型对不同品种核桃仁的总体识别率。

为了便于进行分类,4种核桃仁分别被赋予标号。“温185”核桃、新疆铁核桃、云南泡核桃和东北山核桃分别被赋予标号1、2、3、4。建立模型时,首先对训练集样品数据按照标号进行编码,再利用编

码后的训练集数据建模;验证模型时,模型将验证集样品数据分类为不同的标号,再解码为核桃仁样品类型。

建立分类模型时,采用网格搜索法寻优RBF核函数参数 λ 和 δ ,得到的最优参数分别为 $e^{-5.89}$ 和 $e^{10.05}$,建模精度达到98%。使用得到的分类识别模型对训练集、多品种验证集和各单一品种验证集分别进行验证,验证结果如表2所示。从表2可以看出,建立的分类识别模型对于4种核桃仁的总体正确分类识别率为96%。对于新疆铁核桃、东北山核桃、“温185”核桃和云南泡核桃的识别正确率分别为100%、100%、88%和96%,其中将“温185”核桃识别为云南泡核桃的误识率为12%,将云南泡核桃识别为“温185”核桃的误识率为4%。分析其原因在于,新疆铁核桃为核桃原始野生品种,与“温185”核桃和云南泡核桃的亲缘关系较远;东北山核桃与其他3种核桃虽然都属于核桃属作物,但在更加具体的品种分类中,东北山核桃属于核桃楸组而其他3种核桃属于核桃组,因此这2种核桃识别率较高;“温185”核桃和云南泡核桃属人工培育品种,二者亲缘关系相对比较近,从而对分类效果有影响。

表2 分类模型识别结果

Tab.2 Recognition results of classification model

品种	训练集	验证集	
	识别率/%	识别率/%	误识数 误识类型
“温185”核桃	100	88	3 云南泡核桃
云南泡核桃	100	96	1 “温185”核桃
新疆铁核桃	100	100	0
东北山核桃	100	100	0
总体识别率	100	96	

3 结束语

核桃仁品种分类是核桃加工处理中的一项难题。本文采用了近红外光谱分析方法对来自4个不同品种的核桃仁样本的近红外漫反射光谱进行了分类研究。以多元散射校正和标准正态化的组合方法对原始光谱预处理后,采用主成分分析与支持向量机相结合的方式建立了核桃仁品种分类识别模型,对于4种核桃仁的总体正确分类识别率为96%。实验结果表明,将近红外光谱分析方法应用于核桃仁品种分类是可行的。本文采用了主成分分析与支持向量机算法相结合的建模方式,成功实现了数据降维,降低了建模复杂度,同时具有较高的识别准确度。

参 考 文 献

- 1 新疆维吾尔自治区统计局. 新疆统计年鉴[M]. 北京:中国统计出版社,2014.
- 2 刘慧. 核桃品质介绍与评述[J]. 西北园艺,2012(2):34-36.
- 3 李忠新,杨莉玲,阿不力孜·巴斯提,等. 中国核桃产业发展研究[J]. 中国农机化学报,2013,34(4):23-28.
Li Zhongxin, Yang Liling, Abulizi Basiti, et al. Research on the development of walnut in China[J]. Journal of Chinese Agricultural Mechanization, 2013, 34(4):23-28. (in Chinese)
- 4 新疆农业科学院农业机械化研究所. 核桃仁分级机:CN,ZL201120508537.0[P]. 2013-05-01.
- 5 Jie Dengfei, Xie Lijuan, Rao Xiuqin, et al. Using visible and near infrared diffuse transmittance technique to predict soluble solids content of watermelon in an on-line detection system [J]. Postharvest Biology and Technology, 2014,90:1-6.
- 6 Dimitris Alexandrakis, Gerard Downey, Amalia G M Scannell. Rapid non-destructive detection of spoilage of intact chicken breast muscle using near-infrared and fourier transform mid-infrared spectroscopy and multivariate statistics [J]. Food and Bioprocess Technology, 2012,5(1):338-347.
- 7 郭文川,王铭海,岳绒. 基于近红外漫反射光谱的损伤猕猴桃早期识别[J]. 农业机械学报,2013,44(2):142-146.
Guo Wenchuan, Wang Minghai, Yue Rong. Early recognition of bruised kiwifruit based on near infrared diffuse reflectance spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2013, 44(2):142-146. (in Chinese)
- 8 傅霞萍,应义斌. 基于 NIR 和 Raman 光谱的果蔬质量检测研究进展与展望[J]. 农业机械学报,2013,44(8):148-164.
Fu Xiaping, Ying Yibin. Application of NIR and Raman spectroscopy for quality and safety inspection of fruits and vegetables: a review [J]. Transactions of the Chinese Society for Agricultural Machinery, 2013, 44(8):148-164. (in Chinese)
- 9 Nicola Berardo, Vincenza Pisacane, Paola Battilani, et al. Rapid detection of kernel rots and mycotoxins in maize by near-infrared reflectance spectroscopy[J]. Agricultural and Food Chemistry,2005,53(21):8128-8134.
- 10 汪庆平,黎其万,董宝生,等. 近红外光谱法快速测定山核桃品质性状的研究[J],西南农业学报,2009,22(3):873-875.
Wang Qingping, Li Qiwan, Dong Baosheng, et al. Study on the prediction of oils of carya cathayensis sargent using near infrared reflectance spectroscopy [J]. Southwest China Journal of Agricultural Sciences, 2009,22(3):873-875. (in Chinese)
- 11 唐兴田,邹文锦,郭婷婷,等. 近红外光谱漫反射和漫透射测量方法在玉米品种真实性鉴定中的应用研究[J]. 光谱学与光谱分析,2012,32(6):1531-1534.
Tang Xingtian, Wu Wenjin, Guo Tingting, et al. The utility research on NIR diffuse reflectance and transmittance measurements mode in authenticity identification of maize population samples [J]. Spectroscopy and Spectral Analysis, 2012, 32(6):1531-1534. (in Chinese)
- 12 王姣姣,刘浩,任贵兴. 豌豆品质性状近红外模型建立及区域差异分析[J]. 植物遗传资源学报,2014,15(4):779-787.
Wang Jiaojiao, Liu Hao, Ren Guixing. Using Fourier transform near infrared spectroscopy for the evaluation and regional analysis of pea [J]. Journal of Plant Genetic Resources, 2014, 15(4):779-787. (in Chinese)
- 13 He Yong, Li Xiaoli, Deng Xunfei. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model [J]. Journal of Food Engineering, 2007, 79(4):1238-1242.
- 14 Eva Borrás, Jose Manuel Amigo, Frans van den Berg, et al. Fast and robust discrimination of almonds (*Prunus amygdalus*) with respect to their bitterness by using near infrared and partial least squares-discriminant analysis[J]. Food Chemistry, 2014,153:15-19.
- 15 Jiang Hui, Liu Guohai, Xiao Xiahong, et al. Classification of Chinese soybean paste by Fourier transform near-infrared (FT-NIR) spectroscopy and different supervised pattern recognition [J]. Food Analytical Methods, 2012, 5(4):928-934.
- 16 Khanmohammadi M, Karami F, Mir-Marqués A, et al. Classification of persimmon fruit origin by near infrared spectrometry and least squares-support vector machines [J]. Journal of Food Engineering, 2014, 142:17-22.
- 17 郑田甜,孙腾飞,曹增辉,等. 花生种子品质可见-近红外光谱分析[J]. 光谱学与光谱分析,2015,35(3):622-625.
Zheng Tiantian, Sun Tengfei, Cao Zenghui, et al. Quality analysis of peanut seed by visible/near-infrared spectra [J]. Spectroscopy and Spectral Analysis, 2015, 35(3):622-625. (in Chinese)
- 18 王国安,阿卜杜许库尔·牙合甫,张强,等. 新疆普通核桃起源研究[J]. 新疆农业科学,2015,52(3):572-579.
Wang Guoan, Abduxukar Yakup, Zhang Qiang, et al. Preliminary study on the origin of walnuts(*Juglans regia* L.) in Xinjiang [J]. Xinjiang Agricultural Sciences,2015,52(3):572-579. (in Chinese)
- 19 张雨,董润泉,习学良. 云南核桃种资源现状及开发利用[J]. 西北林学院学报,2004,19(2):38-40.
Zhang Yu, Dong Runquan, Xi Xueliang. Germplasm resource of walnut in Yunnan and its exploitation and utilization [J]. Journal of Northwest Forestry University, 2004, 19(2):38-40. (in Chinese)
- 20 李民赞. 光谱分析技术及其应用[M]. 北京:科学出版社,2007.
- 21 尼珍,胡昌勤,冯芳. 近红外光谱分析中光谱预处理方法的作用及其发展[J]. 药物分析杂志,2008,28(5):824-829.
Ni Zhen, Hu Changqin, Feng Fang. Progress and effect of spectral data pretreatment in NIR analytical technique [J]. Chinese Journal of Pharmaceutical Analysis, 2008, 28(5):824-829. (in Chinese)
- 22 王海青,姬长英,顾宝兴,等. 基于机器视觉和支持向量机的温室黄瓜识别[J]. 农业机械学报,2012,43(3):163-167.

- Wang Haiqing, Ji Changying, Gu Baoxing, et al. In-greenhouse cucumber recognition based on machine vision and least squares support vector machine [J]. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(3):163 - 167. (in Chinese)
- 23 钟波,刘兆科. 基于偏最小二乘的支持向量机多分类方法[J]. 计算机工程与应用,2007,43(17):174 - 179.
Zhong Bo, Liu Zhaoke. Support vector machine multi-classified method based on partial least-squares[J]. Computer Engineering and Applications,2007,43(17):174 - 179. (in Chinese)
- 24 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2 - 10.
Ding Shifei, Qi Bingjuan, Tan Hongyan. An overview on theory and algorithm of support vector machines [J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1):2 - 10. (in Chinese)
- 25 於海明,李石,吴威,等. 稻谷千粒质量近红外光谱预测模型的波长选择方法[J]. 农业机械学报,2015,46(11):275 - 279.
Yu Haiming, Li Shi, Wu Wei, et al. Wavelength selecting methods of NIRS predicting model of paddy 1 000-grain weight[J]. Transactions of the Chinese Society for Agricultural Machinery,2015,46(11):275 - 279. (in Chinese)
- 26 刘秀英,申健,常庆瑞,等. 基于可见/近红外光谱的牡丹叶片花青素含量预测[J]. 农业机械学报,2015,46(9):319 - 324.
Liu Xiuying, Shen Jian, Chang Qingrui, et al. Prediction of anthocyanin content in peony leaves based on visible/near-infrared spectra[J]. Transactions of the Chinese Society for Agricultural Machinery,2015,46(9):319 - 324. (in Chinese)