

## 可见/近红外光谱结合 GWO-SVM 对千禧番茄的分类鉴别

张 伏<sup>1,2,3</sup>, 王新月<sup>2</sup>, 崔夏华<sup>2</sup>, 曹炜桦<sup>2</sup>, 张晓东<sup>1\*</sup>, 张亚坤<sup>2</sup>

1. 江苏大学现代农业装备与技术教育部重点实验室, 江苏 镇江 212013
2. 河南科技大学农业装备工程学院, 河南 洛阳 471003
3. 机械装备先进制造河南省协同创新中心, 河南 洛阳 471003

**摘 要** 千禧番茄营养丰富且酸甜可口, 不同千禧番茄品种的风味和营养价值均有明显差异, 尤其是番茄红素、柠檬酸、维生素 C 和氨基酸含量的差异较大。传统人工分类方式效率低、主观性强、误检率高等问题亟待解决。为筛选综合营养价值高且风味佳的千禧番茄品种, 实现千禧番茄快速准确分类, 提出了基于千禧番茄光谱特征的分类模型构建及 GWO 优化 SVM 算法研究, 以期解决千禧番茄自动化分类问题。以四个品种千禧番茄作为研究对象, 试验样本 240 个, 将其按 2:1 比例划分为训练集 160 个和测试集 80 个样本, 利用可见-近红外光谱采集系统获取 350~1 000 nm 范围内的千禧番茄反射强度, 经光谱校正得样本反射率; 为增强信噪比, 截取 481.15~800.03 nm 范围内的光谱波段作为有效波段。由于数据采集过程受无关信息干扰影响建模效果, 故将平滑点数设置为 3 进行 Savitzky-Golay(SG)平滑预处理。预处理后采用连续投影算法(SPA)提取特征波长变量, 优选得到 11 个特征波长反射率作为输入矩阵  $\mathbf{X}$ , 预设样本变量 1, 2, 3 和 4 作为输出矩阵  $\mathbf{Y}$ , 利用支持向量机(SVM)建立 SPA-SVM 千禧番茄定性分类模型, 训练集和测试集平均分类准确率分别为 59.38%和 48.75%; 在此基础上, 引入灰狼优化算法(GWO)对训练集 160 个样本训练, 寻求 SVM 最优惩罚系数( $c$ )和核函数参数( $g$ ), 根据模型训练结果对测试集 80 个样本预测, 建立 SPA-GWO-SVM 千禧番茄分类模型, 训练集和测试集平均分类准确率分别为 100%和 81.25%。研究表明: 经灰狼优化后的支持向量机模型性能明显提高, 其中训练集和测试集平均分类准确率分别提高了 40.62%和 32.50%, 灰狼优化算法可用于提高支持向量机的分类性能, 实现对千禧番茄品种的分类。本研究为千禧番茄及其他果蔬快速准确分类提供了新的思路和方法。

**关键词** 可见-近红外光谱; 千禧番茄; 分类; 灰狼优化算法; 支持向量机

**中图分类号:** TP391 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2022)10-3291-07

## 引 言

千禧番茄富含番茄红素、VC、可溶性固形物、果糖、葡萄糖、柠檬酸、苹果酸和氨基酸等丰富营养物质, 不同千禧番茄品种其营养物质含量具有明显差异<sup>[1-4]</sup>, 如何实现千禧番茄的快速准确分类, 对筛选综合营养价值高的千禧番茄品种具有重要的研究意义和价值。

随着可见-近红外光谱、高光谱、图像处理等技术在品种分类上应用, 快速和分类准确率是关键评价指标之一。

Marcos等<sup>[5]</sup>基于带描述符的 RGB 颜色模型对番茄模糊分类, 重点优化颜色空间的描述符, 分类预测结果准确率较高。Harvey等<sup>[6]</sup>利用人工蜂群算法训练的人工神经网络(artificial neural network, ANN)分类器对番茄自动分类, 分类准确率 98.19%。周云成等<sup>[7]</sup>基于深度卷积网络对番茄分类, 分类准确率为 81.64%。耿磊等<sup>[8]</sup>提出 EBn-Net 分类模型对 7 种苹果品种分类鉴别, 平均分类准确率为 96.78%。李鸿强等<sup>[9]</sup>基于高光谱建立微型种薯的分类检测方法, 通过线性判别分析、BP(back propagation)神经网络和支持向量机建立分类模型, 模型平均分类准确率为 89.75%。赵杰文等<sup>[10]</sup>基

收稿日期: 2021-08-15, 修订日期: 2021-11-11

基金项目: 现代农业装备与技术教育部重点实验室和江苏省农业装备与智能化高技术重点实验室开放基金课题(JNZ201901), 国家自然科学基金项目(61771224), 河南省科技攻关计划项目(212102110029), 河南省高等教育教学改革研究与实践项目(研究生教育)成果(2019SJGLX063Y)资助

作者简介: 张 伏, 1978 年生, 江苏大学现代农业装备与技术教育部重点实验室教授 e-mail: zhangfu30@126.com

\* 通讯作者 e-mail: zxd700227@126.com

于近红外光谱对不同品种苹果进行分类,以径向基函数为内核建立支持向量机分类模型,通过对比不同预处理方法的分类效果,最终确定惩罚系数  $c$  和正则化系数  $\gamma$  的取值,分类准确率为 100%。

目前,灰狼优化算法作为一种高效的群体智能优化算法,已被证明其稳定性和求解精度明显优于粒子群优化算法 (particle swarm optimization, PSO)、万有引力优化算法 (gravitational search algorithm, GSA) 等<sup>[12]</sup>,因而被广泛用于求解复杂问题。其通过模拟灰狼群体捕食行为,基于狼群群体协作的机制达到优化的目的,因具有良好的性能引起了学者的广泛关注<sup>[11]</sup>。孙俊等<sup>[12]</sup>基于迭代保留信息变量法选取特征变量,以 RBF 径向基函数作为支持向量回归 (support vector regression, SVR) 的核函数,并用灰狼算法优化 SVR 的惩罚因子  $c$  和核函数参数  $g$ ,用优化后的参数  $c$  和  $g$  建立 SVR 回归模型对番茄叶片含水率检测,结果表明经灰狼算法寻优后,模型过拟合效果得到有效改善。安娟华等<sup>[13]</sup>用高斯函数作为支持向量机的核函数,并采用灰狼算法和粒子群算法分别对支持向量机的惩罚因子  $c$  和高斯核函数参数  $\sigma$  优化,结果表明建立的 GWO-SVM 模型对小麦籽粒优劣分级的准确率和速率均明显优于 PSO-SVM。武新燕等<sup>[14]</sup>对近红外光谱的变量筛选应用灰狼算法,为玉米内部成分含量测定建立了 GWO-PLS 模型,预测均方根误差降低了 40%,提高了预测准确率。灰狼优化算法设置参数少、算法简单、鲁棒性强,且运行结果准确度高,收敛速度快,广泛用于农产品的分类,但利用灰狼优化算法对千禧番茄品种分类鲜见报道。基于此,利用可见-近红外光谱对千禧番茄进行分类研究,以 RBF 径向基函数作为支持向量机的核函数,采用灰狼算法优化支持向量机参数,以期提高支持向量机分类性能,获得较佳的分类效果,为千禧番茄的分类研究提供新方法。

## 1 实验部分

### 1.1 材料

千禧番茄样本采自河南省洛阳市孟津县的某采摘园,包括成熟且颜色相近的四个品种改良千禧番茄,分别标号为类别 1、类别 2、类别 3 和类别 4,如图 1,每个品种各有 60 个,共 240 个千禧番茄样本。

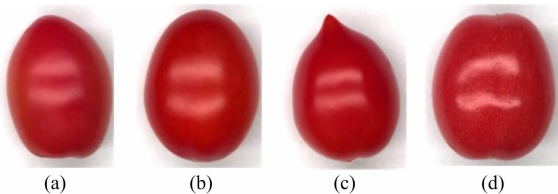


图 1 千禧番茄样本

(a): 类别 1; (b): 类别 2; (c): 类别 3; (d): 类别 4

Fig. 1 Qianxi samples

(a): Type 1; (b): Type 2; (c): Type 3; (d): Type 4

### 1.2 仪器与设备

可见-近红外光谱采集系统主要包括微型光纤光谱仪

(USB4000-VIS-NIR-ES, Ocean Optics, USA)、光学光纤 (SMA 905-VISNIR, Ocean Optics, USA)、光源 (HL-2000, Ocean Optics, USA) 和计算机等,如图 2。光谱仪的扫描范围是 350~1 000 nm,光学分辨率为 1.5~2.3 nm FWHM。利用光谱仪配套的 Spectrasuite 软件获取并处理可见-近红外光谱数据。数据分析软件为 The Unscrambler X10.4, Excel 2019, Origin 2018, Matlab 2016b。

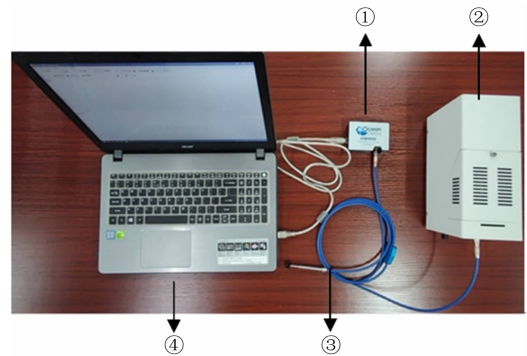


图 2 光谱采集系统

①: 光谱仪; ②: 光源; ③: Y 型光纤; ④: 计算机

Fig. 2 Spectral acquisition system

①: Spectrometer; ②: Light source; ③: Y type fiber; ④: Computer

### 1.3 光谱采集

可见-近红外光谱仪预热 30 min 后采集暗光谱和参考光谱校正,为避免杂光干扰,黑暗环境中测量;在不损坏样本情况下,光纤探头紧贴样本表面,每个果实测量最大横径处两个点的光谱反射强度(如图 3 所示),即每个果实有二组数据,取这两组数据的平均值作为该果实的光谱反射强度测试值。

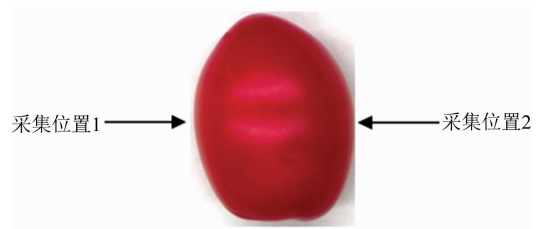


图 3 采集位置示意图

Fig. 3 Schematic diagram of acquisition location

采集后通过 Spectrasuite 软件将光谱数据导入计算机进行分析,处理后得到 345.89~1 040.49 nm 波长范围的平均光谱反射强度曲线,通过式(1)得到千禧番茄反射率,如图 4。为提高光谱测量精度,增强光谱信噪比,剔除 345.89~480.94 和 800.21~1 040.49 nm 两段噪声较大的光谱,即实际有效波长范围是 481.15~800.03 nm。

$$R = \frac{I - I_{AN}}{I_{CK} - I_{AN}} \quad (1)$$

式(1)中,  $R$  为果实反射率,  $I$  为果实反射强度,  $I_{AN}$  为暗光谱,  $I_{CK}$  为参考光谱。

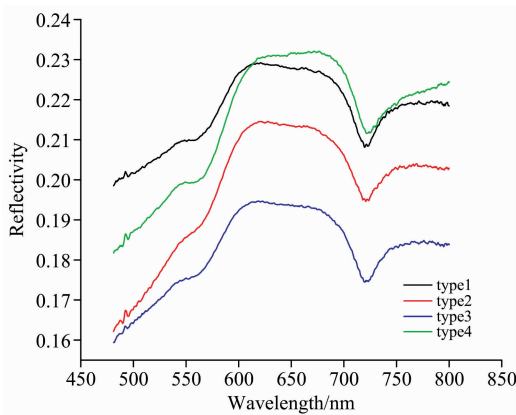


图 4 千禧番茄样本平均反射率曲线图

Fig. 4 Average reflectivity curves of Qianxi tomato samples

测量过程中,每隔 10 min 采集一次暗光谱和参考光谱重新校正,以确保测量准确性。另外,在试验操作过程中,操作人员穿深色衣服,有助于最大限度地减少操作员衣服反射光线干扰。

#### 1.4 建模方法和模型评价标准

连续投影算法(successive projections algorithm, SPA)是一种提取特征变量方法,可实现对原始数据降维处理;支持向量机(support vector machine, SVM)用于定性分类和预测,通过对训练集样本训练,调整参数所得 SVM 模型对测试集数据进行预测,采用投票方式实现样本分类<sup>[15]</sup>,采用性能较稳定的 RBF 径向基函数作为 SVM 核函数,其中,惩罚因子  $c$  和核函数参数  $g$  为两个重要的参数;灰狼优化算法(grey wolf optimization, GWO)模拟自然界中灰狼群体的社会等级制度与狩猎行为的新型群体智能优化算法,主要通过灰狼群体搜索、包围、追踪和攻击猎物等过程以达到优化搜索的目的<sup>[16]</sup>,灰狼算法通过搜索最优参数  $c$  和  $g$  优化支持向量机模型。

光谱数据预处理后,将去噪后的 481.15~800.03 nm 波段范围内的数据提取特征波长作为模型输入变量,四种不同的千禧番茄类别作为模型输出变量建立 SVM 定性分类模型,发现分类效果较差,于是采用 GWO 来寻找 SVM 的最优参数  $c$  和  $g$ ,设置狼群数量、最大迭代次数和参数范围后,初始化  $\alpha$ ,  $\beta$  和  $\delta$  狼的位置及目标函数值,按照  $\alpha$ ,  $\beta$  和  $\delta$  的等级排序后,对训练集样本进行训练,建立不同品种千禧番茄的预测分类模型,通过分类准确率(accuracy)评价模型的性能。

## 2 结果与讨论

### 2.1 训练集与测试集的划分

综合分析了本次采样千禧番茄品种的代表性,同时又考虑到构建模型的稳定性和普适性,将采样得到的 240 个样本按照 2:1 的比例划分为训练集和测试集,如表 1,以此计算分析训练集和测试集平均分类准确率(accuracy)。

### 2.2 光谱数据预处理

由于光谱探测器获得的光谱数据除含被测样本待测成分

表 1 样本划分结果

Table 1 Sample partition results

数据集	训练集	测试集	总计
类别 1	40	20	60
类别 2	40	20	60
类别 3	40	20	60
类别 4	40	20	60
总计	160	80	240

信息外,还容易受到杂散光、基线漂移、噪声、样本背景等无关信息的干扰,从而影响建模效果<sup>[17]</sup>,故在建立模型前对光谱数据预处理,以此降低各种非目标因素对检测信息的影响,以期提高模型精度,建立更稳定可靠的数学模型。

选择 S-G 平滑法(Savitzky-Golay smoothing)对光谱数据预处理,平滑点数为 3,所使用的数据处理软件是 The Unscrambler X 10.4,经过 S-G 平滑法预处理后的光谱平均反射率曲线图,如图 5。

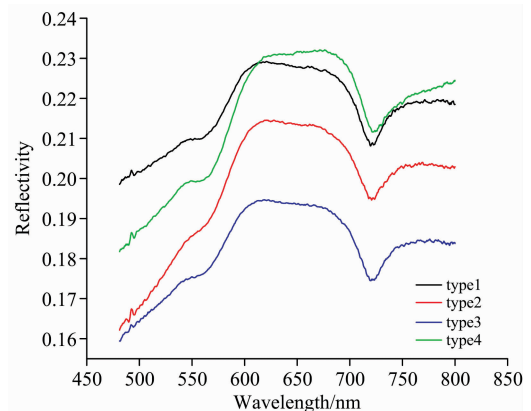


图 5 预处理后的光谱平均反射率曲线图

Fig. 5 Spectral average reflectivity curves after preprocessing

### 2.3 模型分析

#### 2.3.1 连续投影算法

预处理后的光谱数据从 481.15~800.03 nm 波段范围内共 1 621 个波长变量,数据维度过高,冗余信息过多,运行时间过长,直接建立模型会影响分类效果<sup>[18]</sup>,所以需对经过 S-G 平滑预处理后的数据提取特征进行降维;因连续投影算法筛选出的特征波长数目较少,且优选出的波长建立模型预测效果较好<sup>[19]</sup>,故采用连续投影算法提取特征波长,建立千禧番茄分类模型。根据均方根误差(RMSE)最小原则,共优选出 11 个波长变量,如图 6(a)所示,分别是第 56, 198, 427, 951, 1 129, 1 195, 1 228, 1 264, 1 350, 1 559 和 1 618 个变量,即对应 492.54, 521.78, 568.42, 672.55, 707.07, 719.75, 726.06, 732.94, 749.28, 788.52 和 799.47 nm 处的波长,如图 6(b)所示。

#### 2.3.2 支持向量机

使用支持向量机分类器将 SPA 算法提取的 11 个特征波长作为输入变量建立 SPA-SVM 模型,训练集和测试集每个千禧番茄品种的预测正确数目分别如表 2 和表 3,预测分类

结果分别如图 7 和图 8。

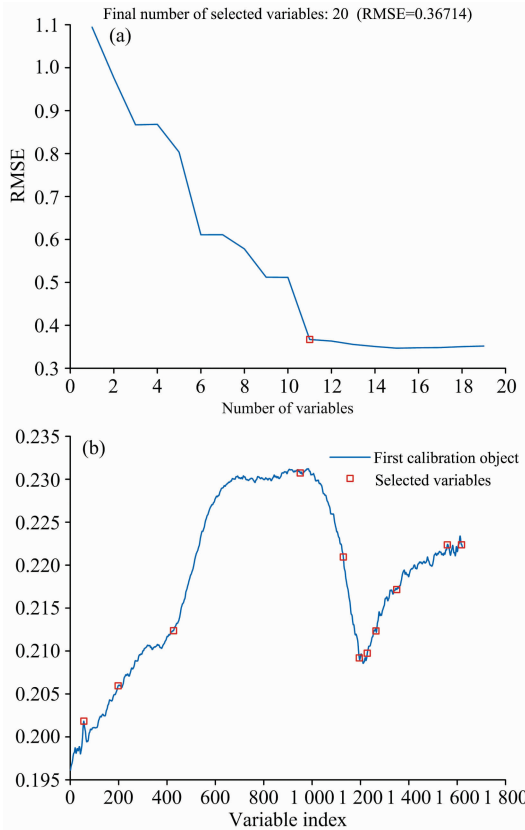


图 6 连续投影算法提取特征波长

(a): 特征波长数目; (b): 对应的特征波长

Fig. 6 Characteristic wavelengths extracted by successive projections algorithm

(a): Number of characteristic wavelengths;  
(b): Corresponding characteristic wavelength

表 2 训练集分类结果

Table 2 Classification results of training set

样本类别	样本数目	预测正确数目	准确率/%
类别 1	40	34	85.0
类别 2	40	20	50.0
类别 3	40	26	65.0
类别 4	40	15	37.5

根据表 2 和图 7, 训练集样本每个类别有 40 个共 160 个, 其中类别 1 预测正确数目为 34, 预测准确率为 85%; 类别 2 预测正确数目为 20, 预测准确率为 50%; 类别 3 预测正确数目为 26, 预测准确率为 65%; 类别 4 预测正确数目为 15, 预测准确率为 37.5%。由表 3 和图 8, 测试集样本每个类别有 20 个共 80 个, 其中类别 1 预测正确数目为 17, 预测准确率为 85%; 类别 2 预测正确数目为 14, 预测准确率为 70%; 类别 3 预测正确数目为 8, 预测准确率为 40%; 类别 4 预测正确数目为 5, 预测准确率为 25%。

综上, 训练集的平均分类准确率为 59.38%, 测试集的平均分类准确率为 48.75%, 结果表明支持向量机模型对干

禧番茄品种的分类效果差。

表 3 测试集分类结果

Table 3 Classification results of test set

样本类别	样本数目	预测正确数目	准确率/%
类别 1	20	17	85
类别 2	20	9	45
类别 3	20	8	40
类别 4	20	5	25

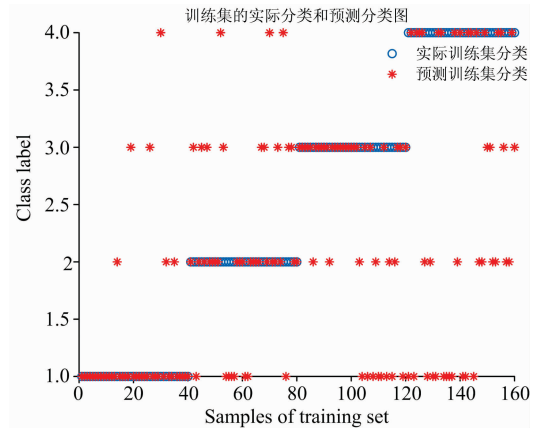


图 7 训练集分类结果

Fig. 7 Classification results of training set

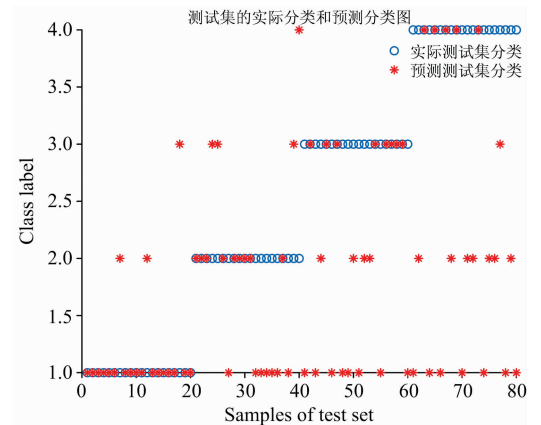


图 8 测试集分类结果

Fig. 8 Classification results of test set

2.3.3 GWO 优化 SVM

采用灰狼优化算法优化上述支持向量机的参数  $c$  和  $g$ , 建立 SPA-GWO-SVM 模型对干禧番茄分类预测, 灰狼优化算法所用公式如式(2)一式(10)。

灰狼个体与猎物的距离

$$D = |CP(n) - W(n)| \tag{2}$$

灰狼个体的位置更新

$$W(n+1) = P(n) - AD \tag{3}$$

$\alpha$  狼与其他灰狼个体的距离

$$D\alpha = |C_1W\alpha(n) - W(n)| \tag{4}$$

$\beta$  狼与其他灰狼个体的距离

$$D\beta = |C_2 W\beta(n) - W(n)| \quad (5)$$

$\delta$  狼与其他灰狼个体的距离

$$D\delta = |C_3 W\delta(n) - W(n)| \quad (6)$$

$\omega$  狼向  $\alpha$  狼移动

$$W_1 = W_\alpha(n) - A_1 D\alpha \quad (7)$$

$\omega$  狼向  $\beta$  狼移动

$$W_2 = W_\beta(n) - A_2 D\beta \quad (8)$$

$\omega$  狼向  $\delta$  狼移动

$$W_3 = W_\delta(n) - A_3 D\delta \quad (9)$$

$\omega$  狼的最终位置

$$W(n+1) = (W_1 + W_2 + W_3)/3 \quad (10)$$

其中,  $W$  表示当前灰狼的位置;  $P$  表示猎物的位置;  $n$  表示当前迭代次数;  $N$  表示最大迭代次数;  $D$  表示灰狼个体与猎物的距离;  $A$  和  $C$  是系数向量,  $A=2ab_1-a$ ,  $C=2b_2$ ;  $a$  是收敛因子,  $a=2-2n/N$ ;  $b_1$  和  $b_2$  是  $[0, 1]$  内的随机数向量。

灰狼算法调优步骤如下:

Step 1 初始化灰狼  $\alpha$ ,  $\beta$ ,  $\delta$  位置;

Step 2 计算灰狼个体适应度值;

Step 3 比较灰狼个体适应度值, 找到最优解、次优解、

第三优解并分别保存位置到  $W_\alpha$ ,  $W_\beta$  和  $W_\delta$ ;

Step 4 计算  $a$ ,  $A$ ,  $C$  值;

Step 5 根据式(1)~式(9)更新灰狼位置;

Step 6 判断是否达到最大迭代次数, 若是则输出  $\alpha$  位置即为所优化参数的最优解; 若否则转到 Step 2。

经灰狼算法优化后的支持向量机参数如表 4 所示。

表 4 经灰狼算法寻优的支持向量机参数

Table 4 Optimized SVM parameters

参数	训练集
$c$	56.888 3
$g$	36.971 1

训练集和测试集每个千禧番茄品种的预测正确数目分别见表 5 和表 6, 预测分类结果分别见图 9 和图 10。

表 5 训练集分类结果

Table 5 Classification results of training set

样本类别	样本数目	预测正确数目	准确率/%
类别 1	40	40	100
类别 2	40	40	100
类别 3	40	40	100
类别 4	40	40	100

根据表 5 和图 9, 训练集样本每个类别有 40 个共 160 个, 四个品种全部预测正确。由表 6 和图 10, 测试集样本每个类别有 20 个共 80 个, 其中类别 1 预测正确数目为 19, 预测准确率为 95%; 类别 2 预测正确数目为 17, 预测准确率为 85%; 类别 3 预测正确数目为 12, 预测准确率为 60%; 类别 4 预测正确数目为 17, 预测准确率为 85%。

表 6 测试集分类结果

Table 6 Classification results of test set

样本类别	样本数目	预测正确数目	准确率/%
类别 1	20	19	95
类别 2	20	17	85
类别 3	20	12	60
类别 4	20	17	85

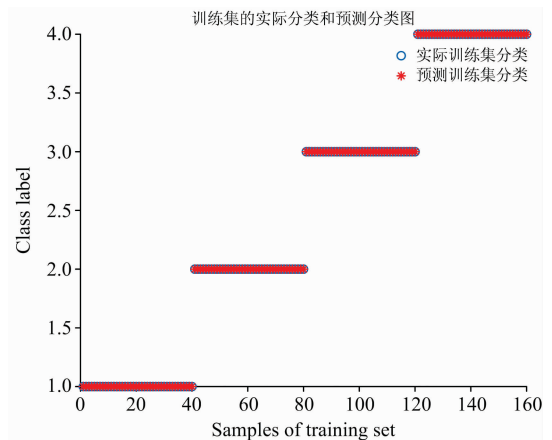


图 9 训练集分类结果

Fig. 9 Training set classification results

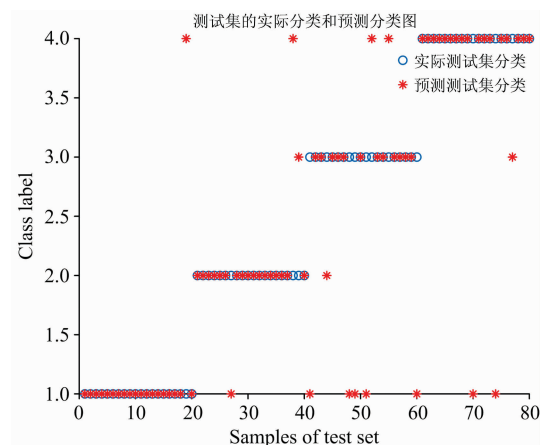


图 10 测试集分类结果

Fig. 10 Classification results of test set

综上, 训练集平均分类准确率均为 100%, 较支持向量机模型准确率提高了 40.62%, 测试集的平均分类准确率为 81.25%, 较支持向量机模型准确率提高了 32.50%, 结果表明经过灰狼算法优化的支持向量机模型较未优化支持向量机模型准确率有明显提高, 对千禧番茄品种具有较好的分类效果。

使用 SPA 算法对 S-G 平滑处理后的光谱数据提取 11 个特征波长 (492.54, 521.78, 568.42, 672.55, 707.07, 719.75, 726.06, 732.94, 749.28, 788.52 和 799.47 nm) 作为模型输入变量分别建立 SPA-SVM 模型和 SPA-GWO-SVM 模型对四个品种千禧番茄分类, 训练集和测试集分类准确率对比图分别如图 11 和图 12。分析可得, 不论是训练集还是测试集, 每个品种优化后的 SPA-GWO-SVM 模型分类

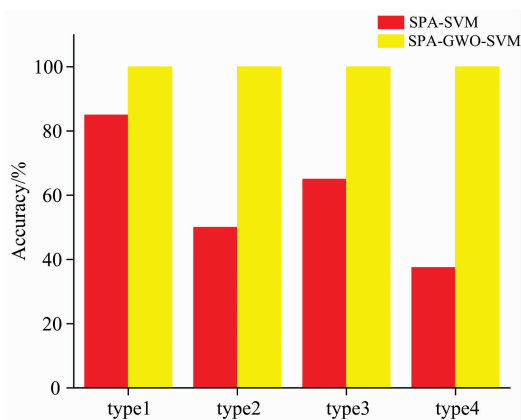


图 11 训练集准确率对比

Fig. 11 Accuracy comparison of training set

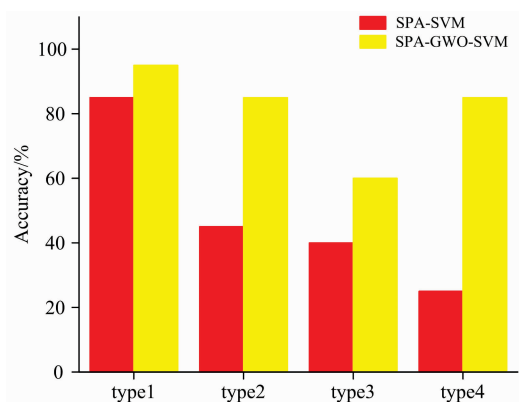


图 12 测试集准确率对比

Fig. 12 Accuracy comparison of test set

准确率均明显提高, 其中类别 2 和类别 4 分类准确率差异最为显著, 这说明灰狼优化算法对支持向量机参数寻优效果较好, 可实现对千禧番茄品种的分类。

### 3 结 论

为筛选风味佳且综合营养价值高的千禧番茄品种, 实现千禧番茄的快速准确分类, 对 4 种不同品种千禧番茄的分类开展了试验和分析。

(1)利用 S-G 平滑法对所采集的光谱数据作平滑去噪预处理, 确定 481.15 ~ 800.03 nm 波段的光谱数据为有效光谱;

(2)采用连续投影算法对预处理后的样本降维, 得到 11 个特征变量, 分别是 492.54, 521.78, 568.42, 672.55, 707.07, 719.75, 726.06, 732.94, 749.28, 788.52 和 799.47 nm;

(3)SPA 提取的特征波长作为输入变量建立支持向量机模型, 训练集共 160 个样本, 预测正确数目为 95, 预测平均分类准确率为 59.38%; 测试集共 80 个样本, 预测正确数目为 39, 预测平均分类准确率为 48.75%;

(4)用灰狼算法寻找最优支持向量机参数  $c$  和  $g$ , 并用优化的支持向量机分类模型对训练集样本训练, 训练集的最优参数为  $c=56.8883$  和  $g=36.9711$ , 四种样本的平均分类准确率为 100%, 较支持向量机模型准确率提高了 40.62%; 测试集平均分类准确率为 81.25%, 较支持向量机模型准确率提高了 32.50%。

研究结果表明经灰狼算法优化可明显提高支持向量机的分类性能, 实现千禧番茄快速准确分类, 为千禧番茄及其他果蔬准确分类提供了新思路和方法。

### References

- [1] CHANG Pei-pei, LIANG Yan, ZHANG Jing, et al(常培培, 梁燕, 张静, 等). Food Science(食品科学), 2014, 35(22): 215.
- [2] ZHOU Yan-chao, XUE Kun, GE Hai-yan, et al(周艳超, 薛坤, 葛海燕, 等). Acta Agriculturae Zhejiangensis(浙江农业学报), 2021, 33(12): 2320.
- [3] CHENG Yuan, WAN Hong-jian, LIU Chao-chao, et al(程远, 万红建, 刘超超, 等). Journal of Zhejiang Agricultural Sciences(浙江农业学报), 2018, 30(11): 1859.
- [4] CHENG Yuan, WAN Hong-jian, YAO Zhu-ping, et al(程远, 万红建, 姚祝平, 等). Journal of Nuclear Agricultural Sciences(核农学报), 2019, 33(11): 2177.
- [5] Villaseñor-Aguilar M J, Botello-álvarez J E, Pérez-Pinal F J, et al. Journal of Sensors, 2019, 2019: 1.
- [6] Opeña H J G, Yusiong J P T. Malaysian Journal of Computer Science, 2017, 30(1): 12.
- [7] ZHOU Yun-cheng, XU Tong-yu, ZHENG Wei, et al(周云成, 许童羽, 郑伟, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2017, 33(15): 219.
- [8] GENG Lei, HUANG Ya-long, GUO Yong-min(耿磊, 黄亚龙, 郭永敏). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2022, 53(6): 304, 369.
- [9] LI Hong-qiang, SUN Hong, LI Min-zan(李鸿强, 孙红, 李民赞). Journal of Instrumental Analysis(分析测试学报), 2020, 39(11): 1421.
- [10] ZHAO Jie-wen, HU Huai-ping, ZOU Xiao-bo(赵杰文, 呼怀平, 邹小波). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2007, 23(4): 149.
- [11] ZHANG Xiao-feng, WANG Xiu-ying(张晓凤, 王秀英). Computer Science(计算机科学), 2019, 46(3): 30.
- [12] SUN Jun, MO Yun-nan, DAI Chun-xia, et al(孙俊, 莫云南, 戴春霞, 等). Transactions of the Chinese Society of Agricultural Engi-

- neering(农业工程学报), 2018, 34(14): 188.
- [13] AN Juan-hua, DONG Xin, WANG Ke-jian, et al(安娟华, 董 鑫, 王克俭, 等). Crop Magazine(作物杂志), 2021, 37(1): 200.
- [14] WU Xin-yan, BIAN Xi-hui, YANG Sheng, et al(武新燕, 卞希慧, 杨 盛, 等). Journal of Instrumental Analysis(分析测试学报), 2020, 39(10): 1288.
- [15] CHU Song-feng, ZHAO Feng-xia, FANG Shuang, et al(楚松峰, 赵凤霞, 方 双, 等). Food & Machinery(食品与机械), 2021, 37(1): 156.
- [16] LONG Wen, CAI Shao-hong, JIAO Jian-jun, et al(龙 文, 蔡绍洪, 焦建军). Acta Electronica Sinica(电子学报), 2019, 47(1): 169.
- [17] PENG Yao, BIAN Xi-hui, WANG Zi-fang, et al(鹏 瑶, 卞希慧, 王姿方, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(9): 2800.
- [18] PENG Yan-kun, ZHAO Fang, LI Long, et al(彭彦昆, 赵 芳, 李 龙, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(5): 159.
- [19] LUO Xia, HONG Tian-sheng, LUO Kuo, et al(罗 霞, 洪添胜, 罗 阔, 等). Laser & Optoelectronics Progress(激光与光电子学进展), 2015, 52(8): 315.

## Classification of Qianxi Tomatoes by Visible/Near Infrared Spectroscopy Combined With GWO-SVM

ZHANG Fu<sup>1, 2, 3</sup>, WANG Xin-yue<sup>2</sup>, CUI Xia-hua<sup>2</sup>, CAO Wei-hua<sup>2</sup>, ZHANG Xiao-dong<sup>1\*</sup>, ZHANG Ya-kun<sup>2</sup>

1. Key Laboratory of Modern Agricultural Equipment and Technology of Ministry of Education, Jiangsu University, Zhenjiang 212013, China
2. College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang 471003, China
3. Collaborative Innovation Center of Advanced Manufacturing of Machinery and Equipment of Henan Province, Luoyang 471003, China

**Abstract** Qianxi tomatoes are rich in nutrition, tasting sweet, sour and delicious, different varieties of qianxi tomato's flavor and nutritional value is obviously different, especially lycopene, citric acid, vitamin C and amino acid content varies greatly and the traditional artificial classification method of low efficiency, strong subjectivity, high rate of error detection and other issues are pressing to be solved. Therefore, in order to screen the high comprehensive nutritional value and good flavor of the qianxi tomatoes to achieve the rapid and accurate classification of the qianxi tomatoes, a classification model based on qianxi tomatoes spectral features and a GWO optimized SVM algorithm was proposed to solve the problem of automated qianxi tomatoes classification. In this study, a total of 240 qianxi tomatoes of four varieties were taken as the research objects, divided into 160 training sets and 80 test sets according to the ratio of 2 : 1. The qianxi tomatoes fruit reflective intensity in the range of 350 to 1 000 nm was obtained by using a visible/near-infrared spectral acquisition system, and the sample reflectance by spectrally corrected was obtained and analyzed. The effective information of the qianxi tomatoes spectrum in the range of 481.15 to 800.03 nm was intercepted to enhance the signal-to-noise ratio. Since the modeling effect is affected by the interference of irrelevant information in the data acquisition process, Savitzky-Golay (SG) smoothing pretreatment was performed with the smoothing point to 3. After SG smoothing pretreatment, the characteristic wavelength variables are extracted by successive projections algorithm (SPA), the reflectance of the optimal selected 11 characteristic wavelength variables as the input matrix  $\mathbf{X}$ , preset sample variables 1, 2, 3, and 4 as output matrix  $\mathbf{Y}$ , the SPA-SVM qualitative classification model of qianxi tomatoes was established. The average classification accuracy of the training set is 59.38%, the test set is 48.75%. On this basis, the gray wolf optimization (GWO) algorithm was introduced to train 160 samples training set, seeking the optimal penalty coefficient ( $c$ ) and the nuclear function parameter ( $g$ ) of the SVM. Based on the training results of the model, the classification results of 80 test set samples were predicted to establish the SPA-GWO-SVM qualitative classification model of qianxi tomatoes and the average classification accuracy of the training set is 100%, the test set is 81.25%. The research results show that the performance of the support vector machine model optimized by the grey wolf algorithm has been improved significantly. The average classification accuracy of the training set is improved by 40.62%, and the average classification accuracy of the test set is improved by 32.50%, which shows that the gray wolf optimization algorithm can be used to improve the performance of the support vector machine classification model and realize the classification of qianxi tomatoes. This study provides a new idea and method for the rapid and accurate classification of qianxi tomatoes and other fruits and vegetables.

**Keywords** Visible/near-infrared spectroscopy; Qianxi tomato; Classification; Grey wolf optimization; Support vector machine

\* Corresponding author

(Received Aug. 15, 2021; accepted Nov. 11, 2021)