基于多维随机森林的番茄灰霉病高光谱图像早期检测

高荣华[1,2]，冯 璐[1,2*]，张 月[3]，原继东[3]，吴华瑞[1,2]，顾静秋[1,2]

1. 北京市农林科学院信息技术研究中心，北京 100097
2. 国家农业信息化工程技术研究中心，北京 100097
3. 北京交通大学计算机与信息技术学院，北京 100044

摘 要 植物病害的自动早期检测对于作物精确保护至关重要。提出了一种基于多维光谱序列(multi-dimensional spectral series，MDSS)和加权随机森林(weighted random forest，WRF)的番茄灰霉病早期诊断与鉴别方法。目的是利用叶片多个观测维度的光谱曲线整体变化趋势建立作物病害检测模型，以期在肉眼明显可见叶面病斑前对作物病害实现诊断。将健康叶片接种灰霉病菌第3天作为叶片成功染病第1天。试验首先采集番茄健康叶片和染病叶片7天内每天的高光谱图像，提取感兴趣区域并计算平均光谱作为初始光谱数据，经筛选共得到(156×7)组有效样本。将样本数据按时间顺序拆分成分别包含1～7个维度的光谱数据形成多维原始光谱序列，为增加维度间差异性，相邻原始光谱序列相减构成多维关联光谱序列。分别采用符号聚合近似估计(symbolic aggregate approximation，SAX)和符号傅里叶近似估计(symbolic Fourier approximation，SFA)两种符号化方法将光谱序列离散成局部辨别性特征。基于多维光谱序列的局部辨别性特征建立加权随机森林(MDSS-SAX-SFA-WRF)分类模型，实现病害早期检测。相应地，基于单维光谱序列(single-dimensional spectral series，SDSS)的番茄灰霉病识别模型被作为基准模型与MDSS-SAX-SFA-WRF模型比较。试验结果显示，MDSS-SAX-SFA-WRF检测模型在包含2至7个光谱序列维度的56个测试样本数据中均获得90%以上识别准确率，在包含5个光谱序列维度测试集中得到最高99%的识别准确率，较SDSS-SAX-SFA-MRF检测模型在染病第5天的识别率高8.2个百分点。另外受随机干扰的影响，SDSS-SAX-SFA-MRF模型准确率在染病5～7 d出现大幅度回落至最低84%，MDSS-SAX-SFA-WRF模型识别率在肉眼可见病斑阶段依然保持超过98%的较高检测水准，未过度回落。因此，提出的基于多维光谱曲线整体变化趋势和加权随机森林(MDSS-SAX-SFA-WRF)的分类模型能够有效实现番茄灰霉病早期检测，并具有较强的鲁棒性，为染病初期的番茄灰霉病鉴别提供新思路。

关键词 早期病害识别；高光谱成像技术；番茄灰霉病；随机森林；多维时间序列

中图分类号：S41-30 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)10-3226-09

引 言

番茄灰霉病是一种低温、高湿病害，是棚室番茄栽培的常见病害。该病繁殖速度快、遗传变异大且适应性强，在患病早期无法通过肉眼发现，一旦进入发病期将扩散迅速，对保护地番茄生产威胁极大，已成为番茄设施栽培的主要限制因素。2020年5月，国务院颁布施行《农作物病虫害防治条例》，条例中指出监测预警是做好农作物病虫害防控的前提和基础[1]。虽然国内外对番茄灰霉病的防治研究已取得了一定进展，但目前生产上仍缺乏该病害早期发现的有效途径。实现对番茄灰霉病的早期检测对我国作物病害防治具有十分重要意义。

随着高光谱成像技术在作物病害识别中的深入应用，对黄瓜[2]、小麦[3]、水稻[4]、马铃薯[5]和苹果[6]等作物病害检测获得较高检出率。高光谱具有唯一性特点，不同作物中叶绿素等生物量含量不同导致其光反射率不同，因此不同物质中光谱曲线走势和波峰波谷差异较大，该种特异性使得基于光谱特征作物病害识别成为可能。秦立峰[2]等提出融合病害差异信息改进的竞争性自适应重加权算法(competitive adap-

收稿日期：2021-08-11，**修订日期**：2022-03-28

基金项目：国家自然科学基金面上项目(61771058)，北京市科技计划课题(Z191100004019007)资助

作者简介：高荣华，女，1977年生，北京市农林科学院信息技术研究中心副研究员 e-mail：gaorh@nercita.org.cn

*通讯作者 e-mail：fengl@nercita.org.cn

tive reweighted sampling，CARS)和连续投影算法(successive projections algorithms，SPA)，并运用二次降维寻优得到的特征波段建立了黄瓜霜霉病早期检测模型。针对番茄病害检测，Sun[7]等利用基于小波变换和最小二乘支持向量机回归(least square support vector machine regression，WT-LSS-VR)的方法筛选最佳波长并建立检测模型，证明了高光谱成像技术在不同镉胁迫下测定番茄叶片重金属含量的可靠性和有效性。贾方方[8]等利用去包络线法(continuum removal)筛选对叶霉病发病程度识别的敏感波段，构建基于光谱特征吸收参量的发病程度估测模型。上述研究基本思路均是首先对所获取作物叶片高光谱图像进行病斑或叶片分割并计算平均光谱曲线，后利用降维工具或相关光谱植被指数(spectral vegetation indices，SVI)对光谱特征波段筛选和提取[9]，过滤不做分类贡献的波段信息，最后基于提取的特征波段建立作物病害识别分类器。Xie[10]等对比了基于全波段和使用特征排序(feature ranking，FR)算法筛选的敏感波段分别建立的K-最近邻(K nearest neighbor，KNN)模型对 5 级患病程度番茄灰霉病叶片的分类准确率，在测试集中，基于全波段 KNN 的总体分类结果为 61.11%，FR-KNN 模型为 45.83%。由此可见，仅用个别波段光谱信息作为分类依据，忽略了光谱曲线整体变化趋势，筛选出不合格波段更易导致曲线信息片面化，尤其患病早期叶片与健康叶片光谱曲线差异不显著，容易导致识别失误，错过病害早期最佳防治时间。

借鉴多维时间序列分类算法(multivariate time series classification，MTSC)，以番茄灰霉病患病早期叶片光谱曲线为基础，从接病第 3 天(染病第 1 天)至完全发病连续采集图像，根据图像不同波段上反射率随时间推移而产生的变化来监测作物病害的发病情况，完整描述患病区域在不同波段下、不同发病时期光谱反射率变化趋势，建立基于多维光谱序列(multi-dimensional spectral series，MDSS)的作物病害分类器，以实现对番茄灰霉病的早期检测。

1　实验部分

1.1　高光谱图像采集

实验采用盆栽培育番茄幼苗 30 株。仪器采用四川双利合谱科技有限公司的 GaiaField-V10 内置推扫式的便携式高光谱成像仪，成像仪光谱范围为 400～1 000 nm，光谱分辨率为 4 nm，选取 360 个光谱通道，单幅拍摄速度 15 s，全幅图像像素分辨率为 960×1 101。为了确保采集的图像清晰且不失真，经预备实验，确定曝光时间为 7.5 ms，物距为 50 cm，图像采集速度为 16 mm·s^{-1}。

图像采集方式如图 1 所示，用支架固定高光谱相机，设置高光谱镜头背对太阳且斜向下 30°与叶片垂直架设，避免在采集过程中光谱仪吸收太阳光导致成像不准确。图像采集时将番茄盆栽置于纯色背景板前，以去除复杂背景的影响，同时在地面标定镜头与盆栽的相对位置以保证采集距离不变。从 30 盆番茄中选出生长状态良好、叶片平展宽大的 180 片叶片作为序列采集对象，其中 90 片作为接病实验组，90 片为正常生长健康对照组。为了探究侵染灰霉病番茄叶片光谱信息随时间推移而产生的变化，以接种灰霉病菌第 3 天(染病第 1 天)作为番茄高光谱序列图像采集的开端，对选定的叶片以固定姿势和角度连续采集 7 d。

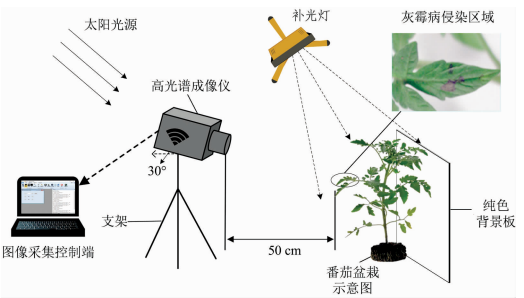

图 1　番茄叶片高光谱图像采集示意图

Fig. 1　Tomato leaves hyperspectral image acquisition diagram

由于每天需采集 180 张叶片，拍摄过程持续时间较长，为消除光照强度对高光谱图像成像影响，尽量选取上下午光照强度相同时间段采集。此外，在光照强度较弱时间段的采集过程中采用补光灯对番茄盆栽进行补光，并根据环境条件变化，对应采集全白与全黑标定图像，用作计算不同光照情况下番茄叶片的光谱反射率值。经过对 30 盆番茄盆栽连续采集，共采集了 180 片叶片 360 个光谱波段上的 7 天连续数据。经过预处理剔除模糊、过曝等不合格实例后，剩余 171 个叶片(实验组 86 个，对照组 85 个)进行光谱信息分析试验。

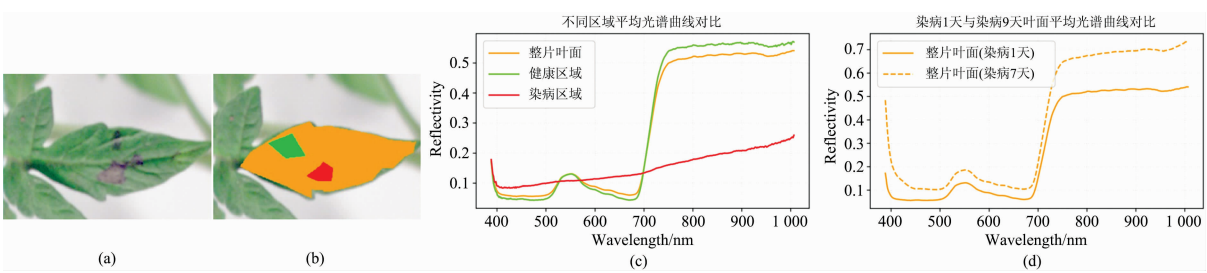

图 2　接病不同感兴趣区域平均光谱反射率比较

(a)：染病 6 天番茄叶；(b)：不同感兴趣区域；(c)：不同区域平均光谱曲线对比；(d)：染病 1 天与 9 天叶面平均光谱曲线对比

Fig. 2　Comparisons of reflectance of different ROIs after inoculation

(a)：Tomato leaf on the 6th day after infection；(b)：Different ROIs selection；

(c)：Comparisons of reflectivity of different ROIs；(d)：Comparisons of reflectivity of leaf on the 1st and 9th day after infection

1.2 高光谱图像预处理

首先对番茄叶片图像进行镜头积分球矫正，消减光照影响，再分别提取叶片病斑部分、健康部分和整片叶子感兴趣区域，获得对比光谱反射率曲线如图 2 所示。图 2(a)展示了染病第 6 天番茄叶片状态，可明显看出染病区域呈深褐色，图 2(b)表示三个不同感兴趣区域选取，橙色为整片叶片，绿色为健康区域，红色为病斑区域，对三个区域光谱反射率做平均处理得到图 2(c)。可知患病区域平均光谱与健康区域有较大差别，整片叶片光谱曲线受病斑区域影响与健康叶片曲线有轻微差异，随着时间延长，该差异逐渐增大，如图 2(d)所示，由此可从整片叶片光谱曲线随时间延长而产生的变化来判断该叶片患病情况。

作物在染病初期，如图 3 所示，由于无法用肉眼识别患病区域，对 171 片叶片进行整片光谱信息提取。其中部分叶片在不同观察日存在叶片卷曲、脱落等问题，为保证多维序列样本完整性，删除数据维度小于 7 d 的样本，共得到 156 个可供实验样本数据，其中(83×7)组患病实验组叶片，(73×7)组对照组叶片，全部接病叶片与健康叶片在连续 7 天的观测下得到的光谱信息如图 4 所示。可以看出，随着接病天数的增加，叶片的平稳反射率在波长 750～1 000 nm 区间内越来越接近 1，该变化为病叶判断提供可能。以番茄叶片高光谱图像的 360 个通道波长作为序列长度，将 156 组试验样本按照近似 65∶35 比例划分为 100 组训练数据和 56 组测试数据进行光谱图像分析。

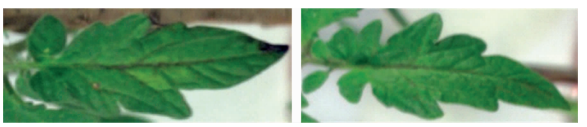

图 3 番茄灰霉病染病第 1 天叶片的 RGB 图像

Fig. 3 RGB image of leaves

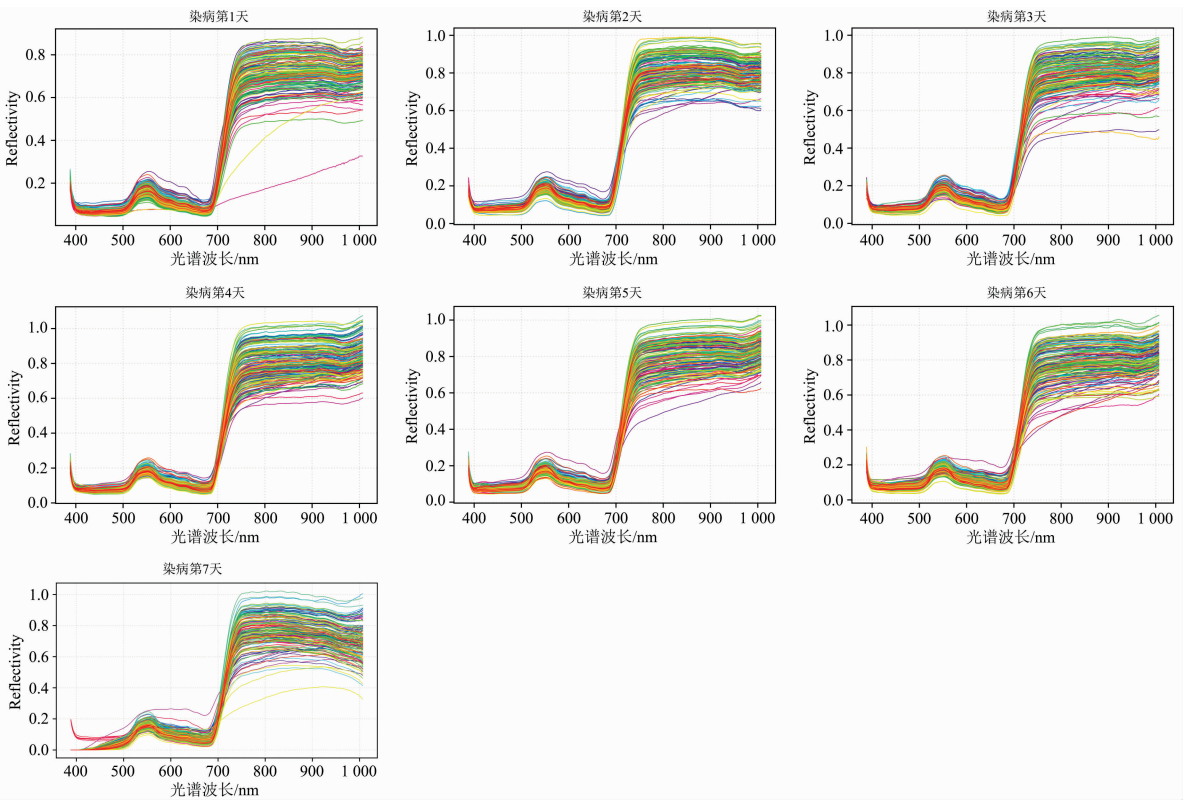

图 4 连续 7 天观测下的全部样本光谱反射率变化情况

Fig. 4 Changes of spectral reflectance of samples observed for 7 consecutive data

1.3 多维光谱序列定义

在探究作物病害的高光谱数据中，光谱序列指一个目标区域在相同波长间隔上连续取反射率值生成的一组数据。以一维光谱序列为基础，多维光谱序列涵盖多个观测时间维度，在执行分类任务时，不仅需要考虑每个维度内部的时序特征，也需要考虑多个维度之间的关联特征。多维光谱序列与病害分类的数学形式表达如下

1.3.1 一维光谱序列

由 n 个有序的观测值 $r_i \mid 1 \leqslant i \leqslant n$ 组成的序列 $R=\{r_1, r_2, \cdots, r_n\}$ 被称为长度为 n 的一维光谱序列，$n=360$。

1.3.2 多维光谱序列

由 d 个一维光谱序列组成的序列集合 $M=\{R_1, R_2, \cdots, R_d\}$ 被称为维度为 d 的多维光谱序列，每个维度序列的长度均为 n，$d \in \{2, 3, 4, 5, 6, 7\}$。多维光谱序列 M 中第 i 维序列的第 j 个观测值用 $r_{i,j}$ 表示。

1.3.3 基于多维光谱序列病害分类

给定包含 m 个实例的训练集 $S=\{M_1, M_2, \cdots, M_m\}$，其中每个实例 M 都是一个长度为 n 维度为 d 的多维光谱序

列，$m=100$。训练集中的实例都属于 C 个类别之一，$C\in$｛患病，健康｝。多维光谱序列分类的目标是在训练集 S 中学习多维光谱序列观测值到所属类别的映射，即叶片是否患病。

1.3.4　光谱子序列

给定一个长度为 n 的光谱序列 T，从第 i 个观测值起，截取 w 个连续观测值组成的新序列$\{t_i, t_{i+1}, \cdots, t_{i+w-1}\}$被称为光谱序列 T 的子序列，其中 $1\leqslant i\leqslant n-w+1$。

本试验数据共有 7 个特征维度，分别对应番茄叶片染病灰霉病菌的第 1 天至第 7 天，为了实现番茄灰霉病病害的早期检测，分别制作包含 2～7 个特征维度的光谱序列数据。例如特征维度为 3 时，156 组实验样本只包含实验叶片染病第 1 天至染病第 3 天和相同观测日期的对照叶片数据，其他维度同理。

1.3.5　多维关联光谱序列

维度随机选择，即在构建每棵决策树时随机选择 D 维光谱序列中的 d 个维度，$D=7$，生成初始光谱序列，从而降低计算复杂度，同时增加决策树之间的差异性。

为了提取初始光谱序列中各个维度之间的关联特征，计算初始序列中任意两条序列之间对应属性点的差值，生成与初始序列长度相同的关联光谱序列，初始序列 T_A 与 T_B 之间的关联序列 T_I 的计算公式如式(1)

$$T_I[i]=T_A[i]-T_B[i],\ 0\leqslant i<n \tag{1}$$

初始序列维度为 d 时，转换后的关联序列维度为 $d'=d(d-1)/2$。图 5(a,b)分别展示了 $d=3$ 时初始序列转换为关联序列的过程。建模时，在平衡样本维度与计算开销情况下，序列维度 d 取值为 4，即样本维度≥4 时，随机选取其中 4 维参与决策树的构建，样本维度<4 时则用原维度。

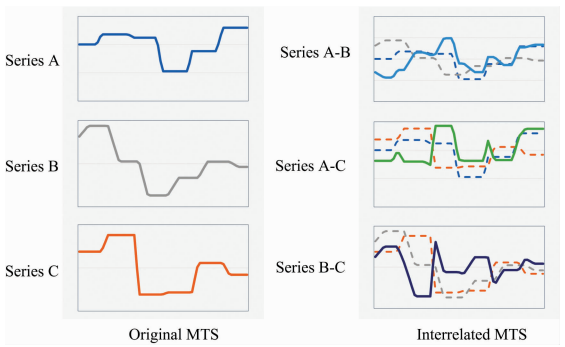

图 5　初始序列(a)与关联序列(b)

Fig. 5　Original series (a) and interrelated series (b)

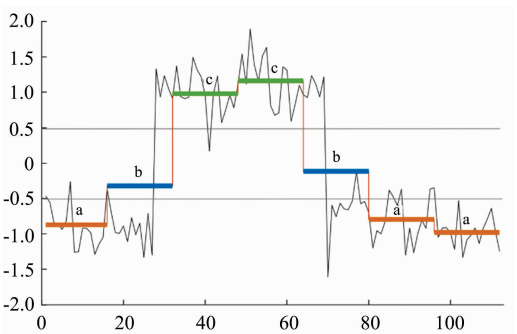

图 6　符号聚合近似估计(SAX)方法

Fig. 6　Symbolic aggregate approximation (SAX) method

1.4　光谱序列符号化表达

建立分类器时，直接采用光谱序列整体的相似性进行分类的方法时间复杂度较高，因此为减小模型复杂度，提取代表光谱序列局部辨别性的特征进行分类以缩减模型运算规模。SAX-VSM 算法(the symbolic aggregate approximation-vector space model)采用符号聚合近似估计(symbolic aggregate approximation，SAX)技术将时间序列转换到时域空间，生成单词特征；BOSS(the bag of SFA symbols)算法采用符号傅里叶近似(symbolic Fourier approximation，SFA)技术将时间序列转换为频域空间的单词特征集合。该两种符号化方法允许在离散后的符号表示上定义距离，运行机器学习算法，同时产生与对原始数据进行操作的算法相同的结果。该性质使得该两种方法在降维的同时还能够保留原始序列的大体形状，因此在各类时间序列任务中被广泛应用。

1.4.1　光谱序列符号化

给定一个长度为 n 的光谱序列 T，利用光谱序列符号化技术将序列 T 或其子序列转换为离散符号组成的字符串，这个过程被称为光谱序列符号化。

1.4.2　SAX 方法符号化光谱序列

给定长度为 w 的光谱序列或其子序列 T、字母表大小 a、单词长度 l，将序列分为等长的 l 个子序列，分别计算每个子序列中各个点的平均值，基于高斯分布将平均值划分为 a 个区域，每个区域对应一个字符，将平均值序列离散化为对应字符，得到序列 T 对应的字符串，这个过程就是 SAX 方法。图 6 展示了采用 SAX 方法将长度为 112 的序列转换为长度为 7 的字符串的过程。

1.4.3　SFA 方法符号化光谱序列

SFA 方法与 SAX 方法类似，区别在于 SFA 方法不计算序列的平均值，而是对序列进行傅里叶变换后将傅里叶系数离散处理，为每个系数计算合适的离散化分割点。图 7 展示了采用 SFA 方法将长度为 64 的序列转换为长度为 4 的字符串的过程。

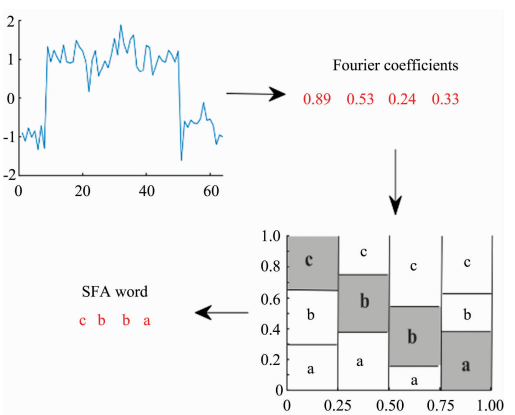

图 7　符号傅里叶近似(SFA)方法

Fig. 7　Symbolic Fourier approximation (SFA) method

1.5　加权随机森林模型构建

随机森林能较好地学习多个特征维度之间的潜在关系，

且复杂度较低、抗干扰能力强，因此常被用来处理高维度数据。例如时间序列森林(time series forest，TSF)算法采用时间序列的间隔特征作为决策树的节点，构建随机森林进行分类。但这类算法只能提取到一条时间序列上的辨别性特征，未考虑到多个维度之间相关性。因此结合作物光谱曲线连续性，本工作选择使用一种用于多维时间序列分类的随机森林算法作为识别作物病害的分类器。在构建随机森林中的每棵决策树时，首先随机选择多维光谱序列中的 d 条维度作为初始序列，以增加决策树之间的差异性并降低时间复杂度；然后将初始序列转换为与原序列长度相等的关联光谱序列，从而提取光谱序列中多个维度之间的关联特征。通过 SAX 方法和 SFA 方法，提取初始序列与关联序列的时域与频域特征。将转换后的光谱序列特征集合作为决策树的输入。在每个结点为每个类别选择代表特征，以待预测光谱序列特征与代表特征之间的相似性作为决策树的分支依据。在独立构造每棵树之后，计算每棵树的权重以进行加权分类，实现番茄灰霉病识别的目标。

1.5.1 随机森林模型构建

借鉴邻近森林[11]的思想，按照待分类实例与每个结点的代表实例之间的相似度进行结点分割。首先在决策树的每个结点中随机选择多个代表不同类别的光谱序列字典特征实例，然后计算当前结点的数据集中各个实例与各个代表实例的余弦相似度，将与某一代表实例最相似的实例划分到该代表实例所属的结点分支。不断重复这个划分过程，直到某一结点中的全部实例都属于同一个类标，则将这一结点作为叶子结点。算法 1 展现了决策树的具体构建过程，重复这一过程构建多颗决策树生成随机森林。

算法 1 决策树构建算法 buildMTSTree(S)

输入：符号化的多维光谱序列数据集 S

输出：决策树根结点 T

01. IF S 中的实例都属于同一个类
02. 将该结点设为叶子结点，结点类标为 S 中任一实例的类标
03. END IF
04. 创建结点 T 与实例集合数组 $E[]$
05. FOR $0 \leqslant i < S.$ numClasses
06. $E[i]$=Random(S_i)，S_i 为数据集 S 中类标为 i 的实例集合
07. END FOR
08. FOR $e \in E$
09. 将与实例 e 距离最近的实例放入集合 S_e 中
10. t=buildMTSTree(S_e)
11. $S_e=\{s \in S \mid \text{argmindist}(s, e)\}$ 在当前树结点 T 中添加分支(e, t)
12. END FOR
13. RETURN T

1.5.2 加权分类过程

传统的随机森林算法采用多数投票法决定分类结果，这种方法操作简单，但在一定程度上忽略了决策树之间的差异。所采用的加权方式，根据每棵决策树的结构，计算其权重，可以反映出当前决策树随机选择的维度、特征提取等是否适合当前数据集，从而提高分类准确率。

决策树的原理可以理解为通过某些特征选择方式，将训练数据集划分为几个部分，使得划分后的数据子集相比与划分之前，具有更高的纯度，即划分后的数据子集中各示例所属的类别不确定性更小。因此在度量决策树的权重时，可以通过度量决策树中每个结点划分前后数据集的纯度来实现。本文选择了计算方式较为简单的基尼指数。基尼指数也被称为基尼不纯度，表示在实例集合中任意一个实例在分类时被分到错误类标的概率，基尼指数的数值越小说明实例集合中实例分类错误的概率越小，即实例集合的纯度越高。其计算公式如式(2)

$$\text{Gini}(S) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2}$$

式(2)中，p_k 表示任意一个实例的类标为 k 的概率，K 为实例集合中所有存在的类标数。

每个决策树的结点包含多个分支，因此计算结点权重时需要分别计算每个分支的基尼指数。然后将每个分支的基尼指数乘以每个分支的数据子集占结点总数据集的比例[式(3)]，累加起来得到结点的权重[式(4)]。将决策树中所有结点的权重加起来，作为整棵决策树的权重。这时决策树的权重数值越小，说明这棵决策树的总体划分效果越好，在随机森林中占的投票权重应该越高，因此采用归一化指数函数[式(5)]处理得到的权重值。权重的计算公式如式(3)—式(5)

$$W_{gj} = \sum_{j=0}^{n} \text{Gini}(S_j) \frac{\text{Size}(S_j)}{\text{Size}(S_{\text{all}})} \tag{3}$$

$$W_t = \sum_{i=0}^{p} W_{gi} \tag{4}$$

$$w_t = \text{softmax}(W_t) = \frac{e^{-W_t}}{\sum_{k=0}^{K} e^{-W_k}} \tag{5}$$

式中，w_t 为最终得到的决策树权重，p 为决策树的结点数，n 为当前结点的分支数。

在分类阶段，对于单棵决策树，待分类实例首先按照当前决策树的符号化方式将多维光谱序列转换为局部辨别性字典特征，然后从决策树的根结点开始，选择与待分类实例最接近的代表实例所属的分支，重复此过程直到到达叶子结点，将叶子结点类别作为待分类实例的类别。最后，通过加权投票得到最后的分类结果，实现基于光谱序列的番茄灰霉病识别。

2 结果与讨论

2.1 番茄灰霉病病程发展

未接种番茄灰霉病毒的植株在整个试验期间保持健康，接种后的植株没有症状的情况下，经过一段潜伏期后出现典型症状。灰霉病的发病症状随着接种后病菌培育时间延长逐渐显著，染病 4 d 后，叶片染病从叶尖或叶缘开始，发生不定形的湿润状、灰褐色病斑。染病 6 d 后，小病斑逐渐发展

成湿腐，并长出一层鼠灰色茸毛状的霉层，此为病菌的分生孢子梗和分生孢子。按照可见光条件下的病程发展，染病 1～4 d 划分为肉眼不可见，染病 5～7 d 分别划分为肉眼可见颜色变化、肉眼可见小病斑、肉眼可见明显病斑。

对比健康叶片与染病叶片的光谱曲线(图 8)，染病叶片的光谱曲线在 550～700 和 800～1 000 nm 两个区间内均与健康叶片呈现不同特点与差异，该差异不仅体现在单一波段上，更体现在曲线整体的变化趋势上。在 550～700 nm 波段区间内，染病叶片的光谱反射率之间不密集[如图 8(a,b)]，呈现分散的特点，而健康叶片的较为紧凑，如图 8(c,d)。在 800～1 000 nm 波段区间内，染病叶片光谱曲线趋势形成带有正斜率向上延伸的曲线，健康叶片呈水平波动或略有抬升的走向特点。该差异特点为基于光谱曲线整体变化趋势特征的早期灰霉病识别提供了基本条件。

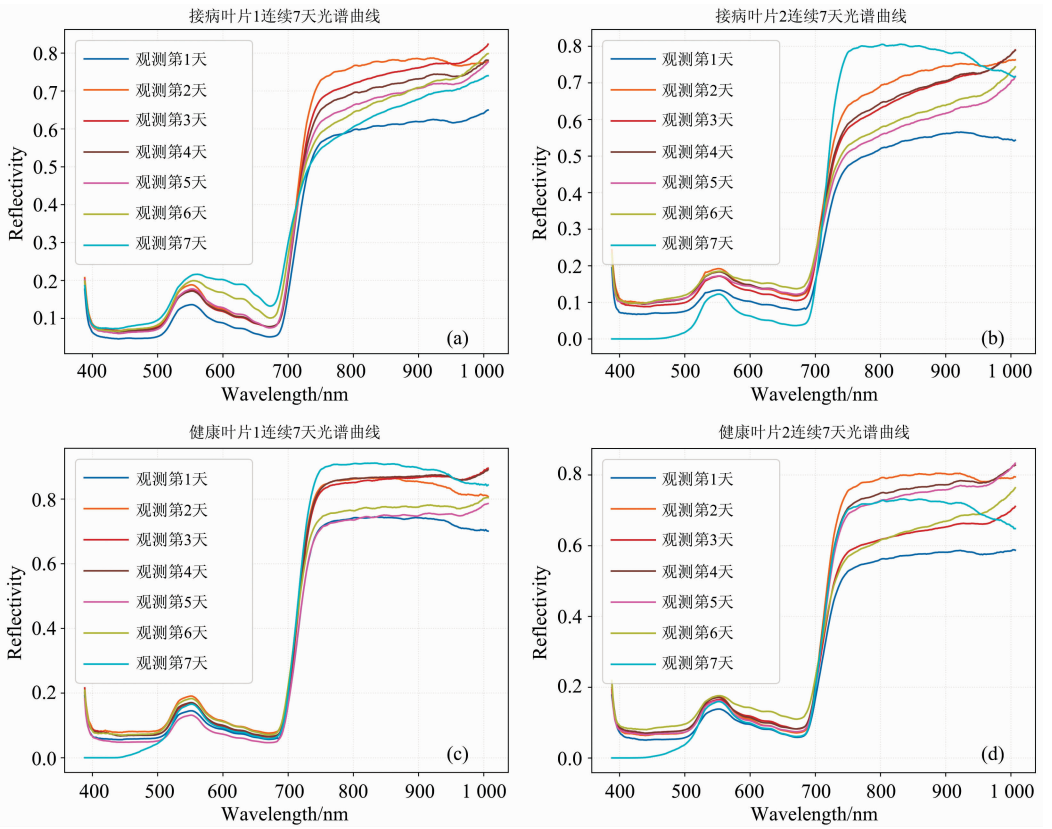

图 8　染病叶片与健康叶片连续 7 天观测下光谱曲线变化

(a)：接病叶片 1 连续 7 天光谱曲线；(b)：接病叶片 2 连续 7 天光谱曲线；
(c)：健康叶片 1 连续 7 天光谱曲线；(d)：健康叶片 2 连续 7 天光谱曲线

Fig. 8　Hyperspectral curve of diseased and healthy leaves for 7 consecutive observations

(a)：Consecutive 7-day reflectivity of infected leaf 1；(b)：Consecutive 7-day reflectivity of infected leaf 2；
(c)：Consecutive 7-day reflectivity of healthy leaf 1；(b)：Consecutive 7-day reflectivity of healthy leaf 2

表 1　符号化方法与随机森林模型构建的参数

Table 1　Parameters of symbolic methods and weighted random forest model

模型参数	参数选择(程序随机)
决策树数量	50
符号化方法	{SAX，SFA，SAX+SFA}
字母表大小 a	{3，4，5}
单词长度 l	{3，4，5，6}
滑动窗口 w	{20%，30%，40%，50%，60%}

2.2　基于单维光谱原始序列的早期检测模型(SDSS-SAX-SFA-WRF)

为验证模型从样本中识别出染病叶片的能力，将健康叶片样本作为正例，7 个不同染病阶段的样本叶片分别作为反例，建立基于单维光谱原始序列的早期番茄灰霉病检测模型。模型的参数选择如表 1 所示。根据预实验结果，在符号化算法 SAX 和 SFA 的字典特征提取中，字母表大小 a 的取值范围为 3～5，单词长度 l 的取值范围为 3～6，滑动窗口大小 w 的取值范围为光谱序列长度的 20%～60%，且以减少随机波动干扰增强鲁棒性为目的，以上 3 个参数的数值随机选择参与序列符号化的构建。

为验证模型稳定性，基于 SFA，SAX 与 SAX+SFA 三种符号化方法建模，重复实验 20 次，将 20 次实验的识别率求平均得到结果如图 9 所示，其中识别准确率 = $\frac{\text{模型正确分类的样本数}}{\text{总样本数}}\times 100\%$。

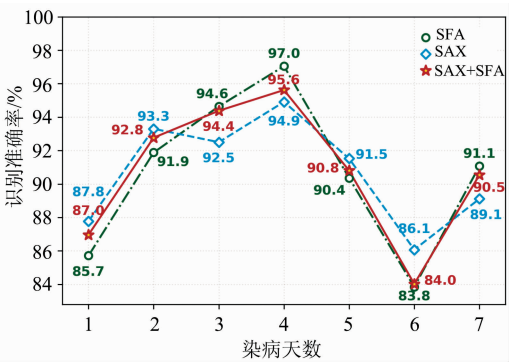

图 9 基于单维光谱序列模型的识别结果

Fig. 9 Recognition results of SDSS-SAX-SFA-MRF model

基于单维原始光谱序列的番茄灰霉病检测模型的分类准确率随着疾病严重程度的增加先增加后降低。在肉眼可见染病叶片明显颜色变化前(染病 1～4 d)，分类准确率在三种符号化方法中均先呈上升趋势，并且在染病第 4 天达到顶峰，识别准确率最高 97%。然而随着染病程度加深，基于单维原始光谱序列的模型识别效果出现回落，染病第 5 天和第 7 天落回到 90%附近，染病第 6 天的识别率降到最低，仅有 83.8%。分析此现象的原因是染病 5～7 d 的叶片光谱曲线产生随机波动[1]，模型识别准确率受到影响，因此基于单维光谱序列的分类器无法作为番茄灰霉病害的早期检测模型。

2.3 基于多维光谱序列的早期检测模型(MDSS-SAX-SFA-WRF)

为解决染病严重时模型识别准确率偏低的问题，从染病第 1 天开始，将单维光谱序列依照染病进程顺序，分别组合成 1～7 个特征序列的多维度数据，离散成局部辨别特征后，基于该特征建立分类模型，以此实现番茄灰霉病早期检测，模型参数与表 1 设置相同。结果如图 10(a,b,c)所示，基于多维光谱序列模型在测试集上的识别准确率逐步上升，对染病 2 天到染病 7 天后发病严重等 6 个阶段的识别均超过 90%，相较于基于单维光谱原始序列的模型均有提高。在 5 维序列时达到最高识别率 99%，并且在维度 6 和 7 中未发生过度回落，准确率依然保持在 98%左右，表明基于多维光谱序列特征能有效避免光谱干扰性波动造成的识别效果不稳定的情况。

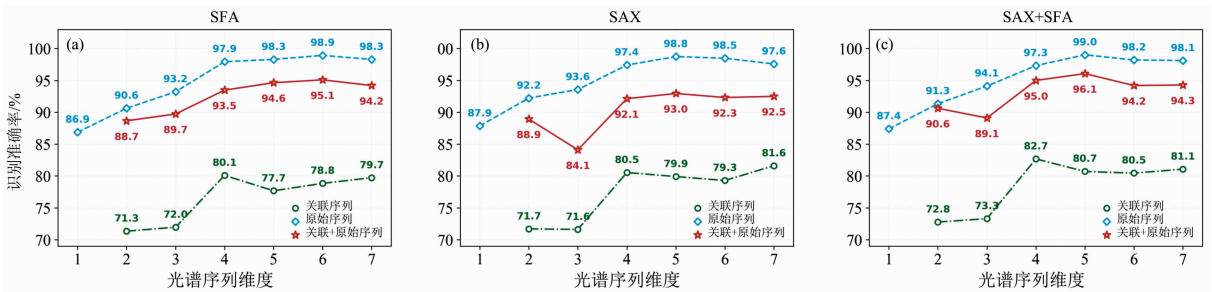

图 10 基于多维光谱序列模型的识别结果

(a)：SFA 多维光谱；(b)：SAX 多维光谱；(c)：SAX+SFA 多维光谱

Fig. 10 Recognition results of MDSS-SAX-SFA-MRF model

(a)：SFA multidimensional spectrum；(b)：SAX multidimensional spectrum；(c)：SAX+SFA multidimensional spectrum

对模型的三种符号化方法进一步分析，基于原始光谱序列的分类器检出率整体优于基于关联序列和组合序列的分类器；基于关联光谱序列的模型检测效果在三种序列表达中的表现最差，整体准确率不高于 80%；基于组合光谱序列的模型识别准确率介于其他两种序列表达方法之间，分布于 86.5%～94.9%。随着序列维度的增加，基于关联序列的识别模型识别效果随时间维度增加提升较大，最大增幅 37.7%，并均在维度为 7 时达到最高点。说明维度特征对关联序列表达尤为重要，叶片光谱特征信息越丰富，关联光谱信息越能表达番茄灰霉病的健康信息与患病信息。然而，在低维度光谱序列中，原始序列则能最大程度保留番茄灰霉病的光谱特征并有效表达，在稳定性和识别准确率中均优于其他两种序列表达方式。

2.4 MDSS-SAX-SFA-WRF 与 SDSS-SAX-SFA-WRF 模型检测效果比较

对比基于多维光谱序列和单维光谱序列在最优符号化方式下模型的准确率(图 11)，在光谱维度为 2～3 时，SDSS-SAX-SFA-WRF 模型检测效果略优于 MDSS-SAX-SFA-WRF

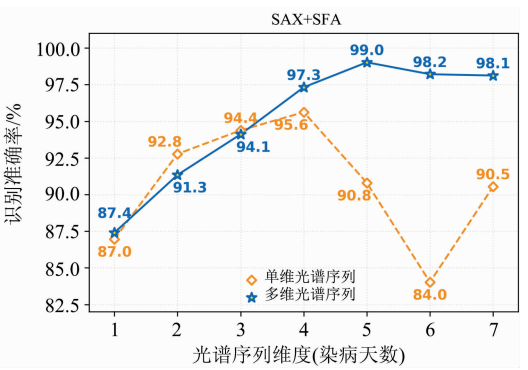

图 11 基于多维与单维原始光谱序列识别准确率对比

Fig. 11 Comparison results of MDSS-SAX-SFA-MRF model and SDSS-SAX-SFA-MRF model

模型，差异基本控制在 1.5%以内。然而，在光谱维度为 4(染病第 4 天)后，MDSS-SAX-SFA-WRF 番茄灰霉病早期检测模型则相较 SDSS-SAX-SFA-WRF 模型一直保持优势，维持不低于 98%的识别准确率，走势较平稳。即使在染病第 1

天，MDSS-SAX-SFA-WRF 模型同样达到 87.4%的检测准确率，将作物病害的检测时间大大提前。随着观测维度增加，MDSS-SAX-SFA-WRF 模型的识别准确率每日增幅 3%，直至染病第 5 天(肉眼可见颜色变化)达到最高。然而即使在该阶段和维度 6～7 d 时，植保人员依然无法仅凭肉眼通过病叶表面颜色的变化对植株所患疾病确诊，仍需待病情进一步发展或从植株中摘取叶片进行化学鉴定，因此在该阶段，作物病害检测模型的准确识别依然发挥重要作用。

3　结　论

借鉴多维时间序列的思想，将多个观测日的光谱曲线累积形成多维光谱序列，为减小模型计算量，采用符号化方法提取光谱序列的局部辨别特征，并且依据该累积局部辨别特征建立加权随机森林模型 MDSS-SAX-SFA-WRF，学习健康叶片与染病叶片在不同观测维度间的差异信息表达，实现番茄灰霉病的早期检测，同时在数据维度逐渐增加情况下，能够在吸收新维度特征的同时最大化保存之前观测维度的辨别特征，并将新旧特征有效结合形成累积辨别特征，将所观测到多维度的光谱曲线特征融合确保模型识别的精准度。

References

[1] LIU Jie, WANG Fu-xiang, ZENG Juan, et al(刘　杰，王福祥，曾　娟，等). China Plant Protection(中国植保导刊), 2020, 40(7): 5.

[2] QIN Li-feng, ZHANG Xi, ZHANG Xiao-qian(秦立峰，张　熹，张晓茜). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2020, 51(11): 212.

[3] LEI Yu, HAN De-jun, ZENG Qing-dong, et al(雷　雨，韩德俊，曾庆东，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2018, 49(5): 226.

[4] YUAN Pei-sen, CAO Yi-fei, MA Qian-li, et al(袁培森，曹益飞，马千里，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2021, 52(1): 139.

[5] JIN Xiu, QI Hai-jun, LI Shao-wen(金　秀，齐海军，李绍稳). Food Science(食品科学), 2018, 39(19): 233.

[6] LIU Si-jia, TIAN You-wen, ZHANG Fang, et al(刘思伽，田有文，张　芳，等). Food Science(食品科学), 2017, 38(8): 277.

[7] Sun Jun, Zhou Xin, Wu Xiaohong, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 212: 215.

[8] JIA Fang-fang, HONG Quan-chun, SONG Wei-yi(贾方方，洪权春，宋唯一). Chinese Journal of Eco-Agriculture(中国生态农业学报), 2017, 25(6): 805.

[9] Zhang Ning, Yang Guijun, Pan Y C, et al. Remote Sensing, 2020, 12(19): 3188.

[10] Xie C, Yang C, He Y. Computers and Electronics in Agriculture, 2017, 135: 154.

[11] Lucas B, Shifaz A, Pelletier C, et al. Data Mining and Knowledge Discovery, 2019, 33(3): 607.

Early Detection of Tomato Gray Mold Disease With Multi-Dimensional Random Forest Based on Hyperspectral Image

GAO Rong-hua[1,2], FENG Lu[1,2*], ZHANG Yue[3], YUAN Ji-dong[3], WU Hua-rui[1,2], GU Jing-qiu[1,2]

1. Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing　100097, China

2. National Engineering Research Center for Information Technology in Agriculture, Beijing　100097, China

3. School of Computer and Information Technology, Beijing Jiaotong University, Beijing　100044, China

Abstract　Automatic early detection of plant diseases is essential for precision crop protection. This paper proposes an early diagnosis and detection method for tomato gray mold based on multi-dimensional spectral series (MDSS) and weighted random forest (WRF) algorithm. The aim was to establish a crop disease detection model by utilizing the overall trend of the spectral curve among multiple observation dimensions of the target leaves to realize the diagnosis before the leaf spot is visible. Generally, the third day after healthy leaves were inoculated with the Botrytis cinerea was treated as the first day that was successfully infected. Therefore, hyperspectral images were recorded from both healthy and infected leaves for 7 days after infection respectively. Then extracted, the region of interest and calculated the average spectrum to form the original spectral samples, whilst (156×7) groups were obtained in total after selection. The group samples were split into multi-dimensional spectral series with 1～7 dimensions per the course of the disease to make up multi-dimensional original spectral series. In order to increase the difference between dimensions, the adjacent original spectral series were subtracted to generate multi-dimensional related spectral series. Afterwards, two symbolic methods, symbolic aggregate approximation (SAX) and symbolic Fourier

approximation (SFA), were employed to discretize each spectral series into local discriminant features. Finally, a weighted random forest classification model (MDSS-SAX-SFA-WRF) based on the local discriminant features of multi-dimensional spectral series is established to realize early disease detection. Accordingly, the model based on single-dimensional spectral series (SDSS) is also built as the benchmark to compare with the MDSS-SAX-SFA-WRF model. The experiment results indicate that the MDSS-SAX-SFA-WRF detection model achieves detection accuracies of more than 90% in 56 testing samples containing 2 to 7 spectral series dimensions, and the highest accuracy up to 99% reached in the 5-dimensional sample data, which is 8.2 percentage higher than that of SDSS-SAX-SFA-MRF detection model on the 5th day of infection. Different from the SDSS-SAX-SFA-MRF model detection performance dropped significantly to the lowest 84% in the 5th～7th days of infection due to random interference. While the discrimination accuracy of the MDSS-SAX-SFA-WRF model still retained a high level of more than 98% in the visible stage of infection without excessive decline. Therefore, the classification model based on the overall change trend of the multi-dimensional spectral curve and weighted random forest (MDSS-SAX-SFA-WRF) proposed in this paper can effectively realize the early detection of tomato gray mold with the strong robustness, which provides a new idea for the early differentiation of crop disease.

Keywords Early disease detection; Hyperspectral imaging; Tomato gray mold; Random forest; Multi-dimensional times series

(Received Aug. 11, 2021; accepted Mar. 28, 2022)

* Corresponding author

关于《光谱学与光谱分析》调整审稿费收费标准的通知

尊敬的《光谱学与光谱分析》广大作者、读者：我刊自 2018 年 7 月 1 日以后登记的稿件向投稿作者收取审稿费 200 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。

2. 没有国家级基金资助的稿件，请您不要投稿。

3. 不是光谱专业的稿件，请您不要投稿。

4. 与其他文章重合率超过 10%的稿件，请您不要投稿。

所投稿件经初审通过后，作者会收到缴纳审稿费的通知。请作者及时从我刊网站(http://www.gpxygpfx.com)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，我刊不再受理。交费后我刊开据增值税电子普通发票，并传至作者提供的电子邮箱，作者可自行打印。

联系电话：010-62181070，62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2018 年 6 月 30 日