doi:10.6041/j.issn.1000-1298.2022.08.019

基于信息熵特征选择的小麦冠层叶绿素含量估测方法

苑迎春[1,2]　周　毅[1,2]　宋宇斐[3]　徐　铮[1,2]　王克俭[1,2]

(1. 河北农业大学信息科学与技术学院, 保定 071001; 2. 河北省农业大数据重点实验室, 保定 071001;
3. 石家庄学院计算机科学与工程学院, 石家庄 050035)

摘要: 为运用图像颜色特征估测作物的叶绿素含量,以自然环境下的小麦冠层图像为研究对象,提出一种基于熵权法的颜色特征选择方法,并应用机器学习方法建立小麦冠层叶绿素含量估测模型。熵权法通过信息熵来衡量颜色特征指标权重,实现冠层图像特征排序,机器学习方法选用多元线性回归(Multiple linear regression, MLR)、岭回归(Ridge regression, RR)和支持向量回归模型(Support vector regression, SVR)估测小麦冠层叶绿素含量。试验结果表明,与皮尔逊相关系数法和主成分分析法选取的特征集进行对比,熵权法得到 a^*、$R-G-B$、$R-G$、$(a^*+b^*)/L$、a^*/b^*、$(R-G)/(R+G+B)$、$(R-B)/(R+B)$、H/S、$(R-G)/(R+G)$等 9 个特征组成的特征集,可以利用较少的特征指标达到最优的预测效果。在选取相同特征指标参数的情况下,SVR 的预测能力优于其它模型,其 R^2 和 RMSE 的平均值分别为 0.80、1.89,相比于 MLR 和 RR 模型 R^2 分别提升 2.8%、1.1%,RMSE 分别下降 0.13 和 0.05。将基于熵权法建立的 SVR 模型应用到 2021 年采集的小麦冠层图像数据,结果表明模型具有很好的稳定性。

关键词: 小麦冠层; 叶绿素估测; 颜色特征选择; 信息熵

中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2022)08-0186-10　　**OSID**:

Estimation Method of Wheat Canopy Chlorophyll Based on Information Entropy Feature Selection

YUAN Yingchun[1,2]　ZHOU Yi[1,2]　SONG Yufei[3]　XU Zheng[1,2]　WANG Kejian[1,2]

(1. *College of Information Science and Technology*, *Hebei Agricultural University*, *Baoding* 071001, *China*
2. *Hebei Agricultural Data Key Laboratory*, *Baoding* 071001, *China*
3. *College of Computer Science and Engineering*, *Shijiazhuang University*, *Shijiazhuang* 050035, *China*)

Abstract: Chlorophyll is an important indicator reflecting the nitrogen nutrition status of crops, and its content is closely related to crop growth and development, photosynthesis capacity and crop yield. With the increasing maturity of image processing technology, choosing image color features to estimate the chlorophyll content of crops has become an important technical means. Taking the wheat canopy image in the natural environment as the research object, a color feature selection method was proposed based on the entropy weight method, and machine learning methods were applied to establish a wheat canopy chlorophyll content estimation model. The entropy method used information entropy to measure the weight of color feature indicators to achieve the canopy image feature ranking. The machine learning method used multiple linear regression (MLR), ridge regression (RR) and support vector regression models (SVR) to estimate the chlorophyll content of wheat canopy. The experimental results showed that compared with the feature set selected by the Pearson correlation coefficient method and principal component analysis, the entropy weight method obtained a^*, $R-B-G$, $R-G$, $(a^*+b^*)/L$, a^*/b^*, $(R-G)/(R+G+B)$, $(R-B)/(R+B)$, H/S, $(R-G)/(R+G)$ and other nine features. The feature sets can use fewer feature indicators to achieve the best prediction effect. In the case of selecting the same characteristic index parameters, the predictive ability of SVR was better than that of other models, and the average values of R^2 and RMSE were 0.80 and 1.89, compared with MLR and RR models, its $R^2$was improved by 2.8% and 1.1%, RMSE was decreased by 0.13 and 0.05, respectively. The SVR model based on the entropy weight method was applied to the wheat canopy image data collected

收稿日期: 2021-08-31　修回日期: 2021-11-17

基金项目: 河北省重点研发计划项目(41130100862301002)和河北省高等学校科学技术研究项目(QN2021409)

作者简介: 苑迎春(1970—),女,教授,博士生导师,主要从事农业信息化和大数据技术研究,E-mail: nd_hd_yyc@163.com

in 2021, and the results showed that the model had good stability. The above research results showed that image processing technology and machine learning methods had very good application value in the estimation of chlorophyll content of crops, providing an important theoretical basis for image-based estimation of chlorophyll content of field crops.

Key words: wheat canopy; chlorophyll estimation; color feature selection; information entropy

0　引言

小麦是我国华北地区主要种植的谷物之一,其长势、产量的准确预测对农业生产和区域经济的发展具有重要意义[1]。叶绿素是反映作物氮素营养状况的重要指标[2],其含量与作物的生长发育、光合作用能力、作物产量密切相关,准确、快速地估测小麦叶绿素含量具有重要的应用价值[3-4]。随着图像处理技术的日益成熟,运用图像特征估测作物的叶绿素含量成为重要的技术手段之一。图像特征指标的选择是建立叶绿素含量预测模型的基础,其选择方法影响着叶绿素估测模型的准确性和稳定性,有效的图像特征指标可以降低数据集的维度,提高估测模型的预测精度和运行效率。因此,有效提取图像特征、构建有效的预测模型来保证叶绿素含量的预测效果是需要解决的关键问题。

利用数字图像技术进行叶绿素含量的估测,国内外已有许多学者对此开展了研究。模型中选用的颜色特征指标大多来自于 RGB 颜色空间,除了红(R)、绿(G)、蓝(B)3 个基本颜色特征以及一阶矩(均值)和二阶矩(方差)特征外,大量基于这 3 个特征构造的组合特征也表现出与叶绿素含量有很强的相关性[5-7]。近年来,构造与叶绿素相关性更强的复杂特征成为一个研究热点[8-9],RGB 颜色空间受光照强弱影响较大,消除光照影响的颜色特征也被相关学者研究[10],除 RGB 颜色空间外,颜色特征提取还扩展到 HIS 和 La^*b^* 2 个颜色空间上[11-12]。颜色特征的构造研究有效提升了叶绿素预测模型的准确性,但也为颜色特征选取和模型的深入研究提出了挑战。

在基于多特征的叶绿素估测建模研究中,现有文献基本采用随机挑选[13]、皮尔逊相关系数法[14]或主成分分析[15]选取与叶绿素相关性高的特征进行建模。在其它研究领域,已有学者运用信息熵的方法进行特征选择,均取得较好的估测效果[16-17]。对于叶绿素估测模型的研究,基于统计回归的模型受到大多数研究者的关注,多元线性回归[18]、岭回归[19]等是比较常用的方法。随着机器学习方法的应用,支持向量回归模型也被提出,支持向量回归模型具有很好的泛化性能,在处理小样本和非线性问题上具有很好的效果[20-22]。

通过对现有研究方法分析,发现数字图像处理技术能够比较快速、准确地构建叶绿素预测模型,但对于目前提出的众多颜色图像特征指标,并没有建立很好的特征选择模型,基于皮尔逊相关系数选取的特征集充分考虑自变量和因变量之间的相关性,主成分分析方法则对考虑的评价特征指标集进行降维,为基于多特征的叶绿素预测建模提供了很好的支持。然而,这些特征之间存在的冗余问题仍会使模型的准确度、稳定性和效率受到影响。在模型构建的研究中,相比最小二乘、多元线性回归等方法,逐步回归、岭回归等方法在一定程度上解决了特征输入的共线性问题,但模型的准确度和稳定性有待于进一步提高。

本文以小麦冠层图像为研究对象,通过挖掘和分析多个常用颜色特征自身所含的信息量差异,并利用信息领域中的信息熵概念进行形式化定义和描述,从而拟提出一种颜色特征筛选方法,在该方法选取的特征集上,再进一步开展小麦冠层叶绿素含量的估测模型研究,以期提升预测模型精度和稳定性,实现小麦冠层图像快速、准确的叶绿素估测。

1　材料与方法

1.1　研究区及 SPAD 数据获取

数据获取地点位于河北省保定市清苑区石桥乡黄陀村河北农业大学示范田基地(38°46′24.90″N,115°32′33.23″E)。试验田划分为 15 个样区,设定 5 组不同施氮水平,分别为:不施氮肥(N0)、氮肥施用量 100 kg/hm^2(N100)、氮肥施用量 180 kg/hm^2(N180)、氮肥施用量 255 kg/hm^2(N255)、氮肥施用量 330 kg/hm^2(N330)。基地种植的小麦品种为济麦 22,该品种具有抗逆性强、产量高等优点。

数据采样在小麦拔节期进行,拔节期是小麦生长发育的重要时期,此时小麦生长迅速,适合监测叶绿素含量。采集时间为 2019 年 4 月 10 日和 2021 年 4 月 10 日 12:00 左右,天气多云无风。小麦叶绿素获取采用 SPAD-502PLUS 型便携式叶绿素仪。每个样区拍摄 6 幅长势均匀的小麦冠层图像,在每幅小麦冠层图像的拍摄区域内选取 5 株对整个样区长势具有代表性的小麦进行叶绿素含量测量,每株

小麦上选取最上面的 3 片叶片的叶绿素含量,每个叶片分别测取叶尖、叶中和叶基部 3 个部位的叶绿素(SPAD 值),将 45 次测量平均值作为该冠层图像的 SPAD 值。样本区共采集 90 幅小麦冠层图像,对采样进行分析,得出小麦 SAPD 值在 32.4 ~ 52.1 范围内,平均值为 43.8,分布差异明显。SPAD 值的分布情况如图 1 所示。

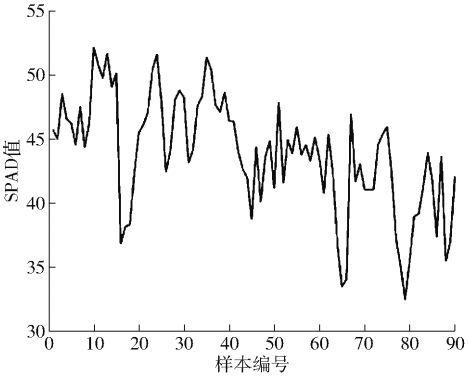


图 1 小麦冠层 SPAD 值分布情况

Fig. 1 Distribution of wheat canopy SPAD value

1.2 小麦冠层图像获取及预处理

小麦冠层图像获取设备采用索尼 FDR - AXP35 4K 型高清摄录一体机,图像采集时将相机固定在三脚架上,镜头距离小麦冠层 1 m,垂直拍摄,图像分辨率为 4 288 像素 ×2 408 像素,部分图像如图 2 所示。

图 2 不同施氮水平小麦冠层图像

Fig. 2 Wheat canopy images of different nitrogen levels

由于大田环境下小麦冠层图像具有背景复杂、光照不均(如土壤、干草、叶片遮挡形成的阴影)等特点,为了提高冠层图像颜色特征值的精确度,需要对采集的图像进行分割处理,以便把小麦冠层图像信息提取出来。

首先提取超绿特征[23](ExG),提高绿色通道的权重,增加绿色小麦冠层与背景(土壤、秸秆、杂草等)的对比度。结合超绿特征设计了基于阈值的小麦冠层分割方法,它按照图像的灰度特性,将图像分成背景和目标两部分。为进一步提升小麦冠层图像颜色特征值的精确度,又将图像中黄色叶片区域去除,经反复测试,阈值设定为 $R>150$,$G>150$,$B<80$,最终得到分割后的小麦冠层图像如图 3 所示。

1.3 特征选取

如前所述,尽管目前提出了很多与叶绿素相关的图像颜色特征指标,但通常都采用随机挑选或依据皮尔逊相关系数选择几个特征来构建预测模型。实际上,颜色特征的选择是模型构建的重要环节,它决定着预测模型的精度。因此,本文首先根据现有文献挑选一组具有代表性的候选颜色特征集,然后通过挖掘分析这些特征中所含信息量的差异,最终选出一组特征用于构建预测模型。

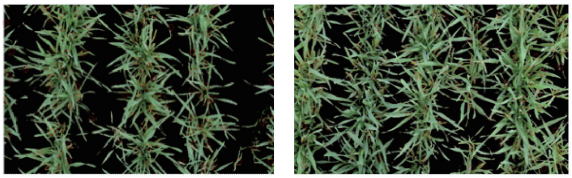


图 3 分割后的小麦冠层图像

Fig. 3 Wheat canopy images after segmentation

1.3.1 颜色特征集

利用文献统计方法,选取的 22 个有代表性的颜色图像特征指标分别为 H[24]、S[13]、a^*[9]、b^*[5]、$G-R$[14]、$R-G-B$[9]、R/B[12]、G/R[25]、$B/(R+G+B)$[26]、$G/(R+G+B)$[26]、$R/(R+G+B)$[26]、$(R-B)/(R+B)$[6]、$(R-G)/(R+G)$[25]、$(G+B-R)/(2G)$[27]、$(G+B-R)/(2R)$[27]、$(R-G)/(R+G+B)$[28]、$(G-B)/(R+G+B)$[6]、$R/\sqrt{R^2+G^2+B^2}$[29]、$(R-G-B)/(R+B)$[27]、B/L[12]、H/S[28]、DGCI[11]。其中 H 为色调均值,S 为饱和度均值,a^* 为红绿色值,b^* 为黄蓝色值,L 为亮度,DGCI 为深绿色指数。

经过统计发现,选取的特征指标集中多数分布在 RGB 空间和 HSI 空间,La^*b^* 空间中指标较少。为了比较全面地评估 3 个颜色空间的图像特征,参照 RGB 特征组合方法,组合构造了 La^*b^* 空间的 2 个指标 a^*/b^* 和 $(a^*+b^*)/L$。这样,候选颜色特征集总共包含 24 个。

小麦冠层图像经过分割处理后,只留下正常小麦叶片的色彩像素。对照每个特征,先计算图像样本中各个像素的颜色值,然后再求平均值作为每幅图像的特征值。

1.3.2 基于熵权法的特征选择

小麦冠层图像中的颜色特征尽管与叶绿素含量都有较强的相关性,但这些信息对预测叶绿素含量的贡献程度并不相同。本文旨在通过信息熵描述这些特征中含有的信息量差异,以此得到每个特征的重要程度。

信息熵[30-31]最早由香农提出,常被用作一个系统信息含量的量化指标,作为系统函数优化的目标或者参数选择的判断依据,广泛应用于通信和计算机等领域。香农定义信息熵 $H(x)$ 公式为

$$H(x) = -c\sum_{i=1}^{n} p(x)\ln p(x) \tag{1}$$

式中　c——常数

$p(x)$——随机事件 x 发生的概率

n——样本总数

香农的信息熵表示信息的不确定程度，它与事件的概率分布情况有关，概率分布越平均，信息熵就越大[32]。信息熵作为一种客观赋权的方法，可以避免人为因素带来的误差[33]。基于香农的信息熵权重赋值思想，反观小麦冠层图像的颜色特征，如果一个颜色特征在所有样本中的信息熵越大，则表明该特征在所有样本中的分布越均衡，也就是说，该特征在所有样本中的差异不明显，则表明它对叶绿素评估的贡献程度低。由此引入颜色特征信息熵（Color feature information entropy，CFIE）的定义。

定义 1：颜色特征信息熵定义为该特征指标在所有样本集上概率分布的数学期望值。

假设 $A_j(1\leqslant j\leqslant 24)$ 表示本文研究的任意一个颜色特征指标，它的颜色特征信息熵 $\mathrm{Info}(A_j)$ 计算式为

$$\mathrm{Info}(A_j) = -c\sum_{i=1}^{n} P_i(A_j)\ln P_i(A_j)$$
$$(i=1,2,\cdots,n;j=1,2,\cdots,m) \tag{2}$$

式中　$P_i(A_j)$——第 i 幅图像的颜色特征指标值为 A_j 的概率

m——颜色特征指标总数

为了便于计算，常数 c 记为 $1/\ln n$，这里 n 为图像样本总数。需要说明的是，$P_i(A_j)$ 是颜色特征信息熵的重要参数，其计算方法是特征指标 A_j 在所有样本图像中出现的概率，计算式为

$$P_i(A_j) = \frac{x_i(A_j)}{\sum_{i=1}^{n} x_i(A_j)} \tag{3}$$

式中　$x_i(A_j)$——特征 A_j 在第 i 幅图像上的特征数值

为确保每个指标提取信息的合理性与准确性，对每个颜色指标进行了标准化处理，使所有颜色特征指标值分布在(0,1)区间。

根据定义 1 和熵权法思想，可以定义颜色特征信息权重系数（Information utility weight coefficient，IUWC），用来反映特征指标的区分能力。

定义 2：颜色特征信息效用权重系数：给定任意一个颜色特征 A_j，其信息效用权重系数定义为其信息效用值占全部颜色特征的信息效用值的百分比，计算式为

$$W(A_j) = \frac{1-\mathrm{Info}(A_j)}{\sum_{j=1}^{m}(1-\mathrm{Info}(A_j))}\times 100\% \tag{4}$$

式中，$1-\mathrm{Info}(A_j)$ 为颜色特征信息效用值，目的是使其与颜色特征指标的利用价值成正比。也就是说，一个颜色特征的信息效用权重系数越大，它的区分能力就越强，对于叶绿素估测的贡献度就越大。

综上，基于信息熵的特征选取方法流程用伪代码表示为：

输入：特征指标数据集 D。

输出：每个颜色特征值指标 IUWC 值并排序。

初始化特征指标个数为 m、图像样本个数为 n。

for $j=1$ to m do

　将 j 指标归一化为(0,1)区间。

　for $i=1$ to n do

　　根据式(2)、(3)计算 j 指标的颜色特征信息熵。

　End for

　根据式(4)计算 j 指标的信息效用权重系数。

End for

按照信息效用权重系数降序排序特征。

输出排序结果。

1.4　模型建立

采用多元线性回归、岭回归、支持向量回归 3 种算法建立小麦冠层叶绿素含量估测模型。多元线性回归作为基础的回归预测方法，广泛应用于数字图像估测模型中，岭回归能在一定程度上解决图像特征之间的多重共线性问题，支持向量回归可解决特征指标与叶绿素含量间存在的非线性问题，故选用 3 种模型。

多元线性回归（Multiple linear regression，MLR）是一种常用的回归方法，其通过多个自变量表示因变量来获取一条最佳拟合直线，常用的求解方法为最小二乘法。

岭回归[34]（Ridge regression，RR）是一种用于解决数据共线性问题的有偏估计方法，它在最小二乘估计的基础上增加了一个 L2 正则化项 $\lambda\|\beta\|_2^2$，通过放弃最小二乘估计的无偏性，以损失部分信息为代价来保证回归系数相对稳定，估测效果更加准确。

支持向量机（Support vector regression，SVR）模型具有优异的全局优化性能，在维数较高且具备复杂非线性特点的小样本回归预测应用中展现出了较好的泛化能力。

1.5　模型验证

选用决定系数（Coefficient of determination，R^2）、均方根误差（Root mean square error，RMSE）进行模型精度的检验。模型的 R^2 越接近 1，说明其预测能力越好，相对应的 RMSE 越小，说明其可靠程度越高。

2 结果分析

2.1 特征指标权重赋值对比

运用熵权法、皮尔逊法以及主成分分析法对 24 种特征的权重赋值结果进行对比,皮尔逊法是通过分析每个特征自变量与叶绿素的相关性得到一个相关系数。相关系数代表每个特征与目标因变量的相关程度,相关系数越大,表明该特征对估测叶绿素的重要程度越大。一般选取相关系数大于 0.6,说明该特征与目标变量相关性较大。主成分分析(PCA)是通过降维技术将多个变量转换为少数几个主成分(综合变量)的一种多元统计方法。它可以在信息丢失最小的前提下,实现多元数据的特征融合,在提取出主要特征的同时去除多元变量间的线性相关。通过提取最终的主成分特征变量完成对目标变量的预测。表 1 给出了 24 个特征的信息熵权重系数、SPAD 的皮尔逊相关系数和主成分累计贡献率的计算结果,按照信息权重系数降序显示。

从表 1 可以看出,特征 a^*、$R-G-B$、$R-G$、$(a^*+b^*)/L$、a^*/b^* 信息权重系数较高,达到 10% 以上。按相关程度来看,所有特征参数与 SPAD 相关系数都在 0.6 以上,并且都达到极显著相关水平,其中 $(R-G-B)/(R+B)$ 与 SPAD 的相关性最高,相关系数达到了 0.801。按主成分分析来看,3 个主成分的累计贡献率便达到 97.73%,选取 12 个主成分时累计贡献率达到 100%。对本文提出的 a^*/b^* 和 $(a^*+b^*)/L$ 2 个特征指标,信息权重系数和 SPAD 的相关系数均取得了较好效果。

表 1 小麦冠层图像特征与叶绿素含量信息权重、相关系数及成分累计贡献率分析

Tab.1 Analysis of weight, correlation coefficient and cumulative contribution rate of wheat canopy image features and chlorophyll content information

编号	指标	信息熵权重系数/%	皮尔逊相关系数	主成分累计贡献率/%	编号	指标	信息熵权重系数/%	皮尔逊相关系数	主成分累计贡献率/%
1	a^*	17.98	−0.619**	62.18	13	b^*	0.92	0.700**	100
2	$R-G-B$	15.12	−0.662**	94.83	14	$(G+B-R)/(2R)$	0.64	0.777**	100
3	$R-G$	14.09	−0.754**	97.73	15	$B/(R+G+B)$	0.59	0.703**	100
4	$(a^*+b^*)/L$	12.13	−0.782**	99.09	16	$(G-B)/(R+G+B)$	0.50	−0.611**	100
5	a^*/b^*	11.60	−0.762**	99.60	17	H	0.47	0.783**	100
6	$(R-G)/(R+G+B)$	8.70	−0.785**	99.85	18	R/B	0.46	−0.762**	100
7	$(R-B)/(R+B)$	8.37	−0.753**	99.89	19	$(G+B-R)/(2G)$	0.29	0.764**	100
8	H/S	2.82	0.714**	99.93	20	B/L	0.28	0.622**	100
9	$(R-G)/(R+G)$	1.39	0.788**	99.96	21	$G/(R+G+B)$	0.15	0.752**	100
10	S	1.34	−0.721**	99.98	22	G/R	0.08	0.783**	100
11	$(R-G-B)/(R+B)$	0.96	−0.801**	99.99	23	$R/(R+G+B)$	0.08	−0.785**	100
12	DGCI	0.94	0.606**	100	24	$R/\sqrt{R^2+G^2+B^2}$	0.07	−0.787**	100

注:** 表示极显著相关($p<0.01$)。

通过分析发现,按照熵权法得到信息权重系数最高的特征,其相关系数并不是最大的,而得到信息权重系数较低的特征,如 $G/(R+G+B)$、G/R、$R/(R+G+B)$ 其相关系数并不是最小的,说明两种“赋权”方法从不同角度对每个特征刻画其“重要”程度。熵权法是通过分析特征集本身的数据分布得出的,一个特征如果在所有样本上分布均匀,它的权重就越低,说明它对目标 SPAD 的贡献程度不明显。反之,一个特征如果在所有样本上分布差异大,则表明它对于预测目标 SPAD 值特性贡献程度有差异,其权重系数就越大。主成分分析降低了特征变量间的相关关系,由该方法降维后形成的主成分变量对 SPAD 值进行建模估测。而相关系数是直接刻画特征自变量和因变量 SPAD 值的相关程度,相关系数越高,说明它对目标因变量 SPAD 值估测越重要。3 种方法本质的区别在于:信息熵权重是分析自变量空间中特征变量之间重要程度的差异性,相关系数则是考虑自变量和因变量之间的相关程度,主成分分析最大程度降低了特征变量间的相关性。

2.2 颜色特征筛选结果对比

颜色特征是模型构建的输入参数,如何从一组特征中筛选出满足模型预测精度,同时又能达到数据降维是研究小麦营养诊断模型的关键问题,采用逐步回归模型对 24 种特征进行筛选试验,特征以熵权法权重降序排序。逐步回归首先选择一个信息权重系数最高的指标作为输入,构建线性模型,决定系数 R^2 作为模型评价依据;然后在原有输入参数基础上再添加一个信息权重系数次高的指标作为模型输入参数,再次构建逐步回归模型。依次类推,直至全部指标均进入模型。同样,皮尔逊法也是按照相关

系数大小依次建立逐步回归模型，作为特征选择方法进行对比分析。

表 2 给出了基于熵权法和皮尔逊特征选择建立逐步回归模型得到的预测精度对比结果，其中两种特征集中的数字代表选择的特征指标编号，是与特征编号一一对应的。从表 2 中明显看出，在熵权法特征选择中，基于 a^*、$R-G-B$、$R-G$、$(a^*+b^*)/L$、a^*/b^*、$(R-G)/(R+G+B)$、$(R-B)/(R+B)$、H/S、$(R-G)/(R+G)$ 特征集输入参数构建的回归预测模型精度最高，最优结果 $R^2=0.772$，指标数占全部指标的 37.5%。在皮尔逊特征选择集中，基于 $(R-G-B)/(R+B)$、$(R-G)/(R+G)$、$R/\sqrt{R^2+G^2+B^2}$、$(R-G)/(R+G+B)$、$R/(R+G+B)$、G/R、H、$(a^*+b^*)/L$、$(G+B-R)/(2R)$ 特征集构建的模型精度 R^2 要低 2.2%，尽管在 $(R-G-B)/(R+B)$、$(R-G)/(R+G)$、$R/\sqrt{R^2+G^2+B^2}$、$(R-G)/(R+G+B)$、$R/(R+G+B)$、G/R、H、$(a^*+b^*)/L$、$(G+B-R)/(2R)$、$(G+B-R)/(2G)$、a^*/b^*、R/B 特征集输入参数构建的回归预测达到了最高模型精度，但其结果为 $R^2=0.770$，精度不仅略低于熵权法特征选择方法，并且也比熵权法多了 3 个特征参数，指标数占全部指标的 50%。这可能是由于皮尔逊相关系数法选取的特征集多集中在 RGB 空间，特征之间的冗余信息多，多重共线性问题对预测结果造成一定干扰，而基于信息熵选取的特征指标分布在 3 个颜色空间内，各个特征包含的信息效用值最大，在一定程度上避免了共线性问题。由于信息熵权法的特征选择利用每个特征所包含的信息效用值来选择特征，它能够找出最重要的一组特征构建模型，增强了特征指标自变量之间的线性无关性，所以能在较少的特征集上建立预测精度高的估测模型。

表 2　小麦冠层叶绿素含量诊断模型

Tab. 2　Diagnostic model of chlorophyll content in wheat canopy

模型编号	熵权法特征选择集	熵权法决定系数 R^2	皮尔逊特征选择集	皮尔逊法决定系数 R^2
1	1	0.607	11	0.656
2	1,2	0.653	11,9	0.671
3	1,2,3	0.671	11,9,24	0.685
4	1,2,3,4	0.723	11,9,24,23	0.694
5	1,2,3,4,5	0.727	11,9,24,23,6	0.716
6	1,2,3,4,5,6	0.739	11,9,24,23,6,17	0.725
7	1,2,3,4,5,6,7	0.753	11,9,24,23,6,17,22	0.748
8	1,2,3,4,5,6,7,8	0.769	11,9,24,23,6,17,22,4	0.749
9	1,2,3,4,5,6,7,8,9	0.772	11,9,24,23,6,17,22,4,14	0.750
10	1,2,3,4,5,6,7,8,9,10	0.769	11,9,24,23,6,17,22,4,14,19	0.763
11	1,2,3,4,5,6,7,8,9,10,11	0.770	11,9,24,23,6,17,22,4,14,19,18	0.766
12	1,2,3,4,5,6,7,8,9,10,11,12	0.768	11,9,24,23,6,17,22,4,14,19,18,5	0.770
13	1,2,3,4,5,6,7,8,9,10,11,12,13	0.770	11,9,24,23,6,17,22,4,14,19,18,5,3	0.767
⋮	⋮	⋮	⋮	⋮
24	所有特征组成的特征集合	0.770	所有特征组成的特征集合	0.770

为进一步分析 2 种特征选择过程中的预测精度变化趋势，图 4 给出了 2 种特征选择方法构建逐步回归模型得到的每一步 R^2 的折线图。从图 4 看出，随着输入变量的增加，两种特征选择方法所构建的模型精度有大致相同的变化趋势。前期，随着输入特征数的增加，决定系数 R^2 均呈快速上升趋势，达到了最高预测精度，随后，随着输入变量的增加，2 个模型趋于稳定，精度几乎不再变化，基本与全特征输入参数建立的模型相同。

从熵权法的逐步回归模型预测精度曲线看，尽管初期建模精度很低，但随着输入变量的增加，其 R^2 显著提升，当输入变量增加至 9 个时，R^2 达到最大值，随后几乎不再变化。而皮尔逊法的逐步回归模型预测精度曲线初始预测精度较高，但随着输入特

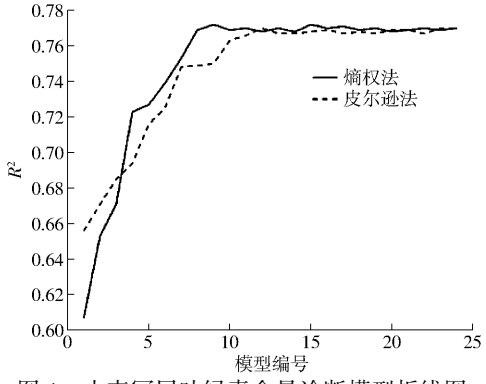


图 4　小麦冠层叶绿素含量诊断模型折线图

Fig. 4　Diagnostic model line chart of chlorophyll content in wheat canopy

征变量数目的增加,其模型精度增长速度比较缓慢,最后达到最好预测精度时,特征参数为 12 个,比熵权法输入的特征数量多 3 个。

2.3 不同特征选择方法的估测模型对比分析

本文以 2.2 节中熵权法选取的 9 个特征和皮尔逊方法选取的 12 个特征以及主成分分析的 3 个特征为自变量,分别使用多元线性回归(MLR)、岭回归(RR)、支持向量回归(SVR)构建小麦叶绿素含量估测模型。RR 模型中,根据岭迹分析,将岭正则化参数 α 设置为 0.1,来保证岭回归系数基本稳定。SVR 模型中,核函数类型为高斯核函数,惩罚系数 C 和核函数参数 σ 的取值范围分别为[0,100]、[0,10],分别以步长 1 和 0.1 为变化单位,每次试验选取效果最佳的 C、σ。使用十折交叉验证方式来评估算法的泛化能力,即将样本随机分成 10 份,9 份为训练集,1 份为测试集,用交叉验证遍历全部 10 份样本。最终通过测试集的决定系数 R^2 和均方根误差 RMSE 来衡量回归模型的预测能力,如图 5 所示。

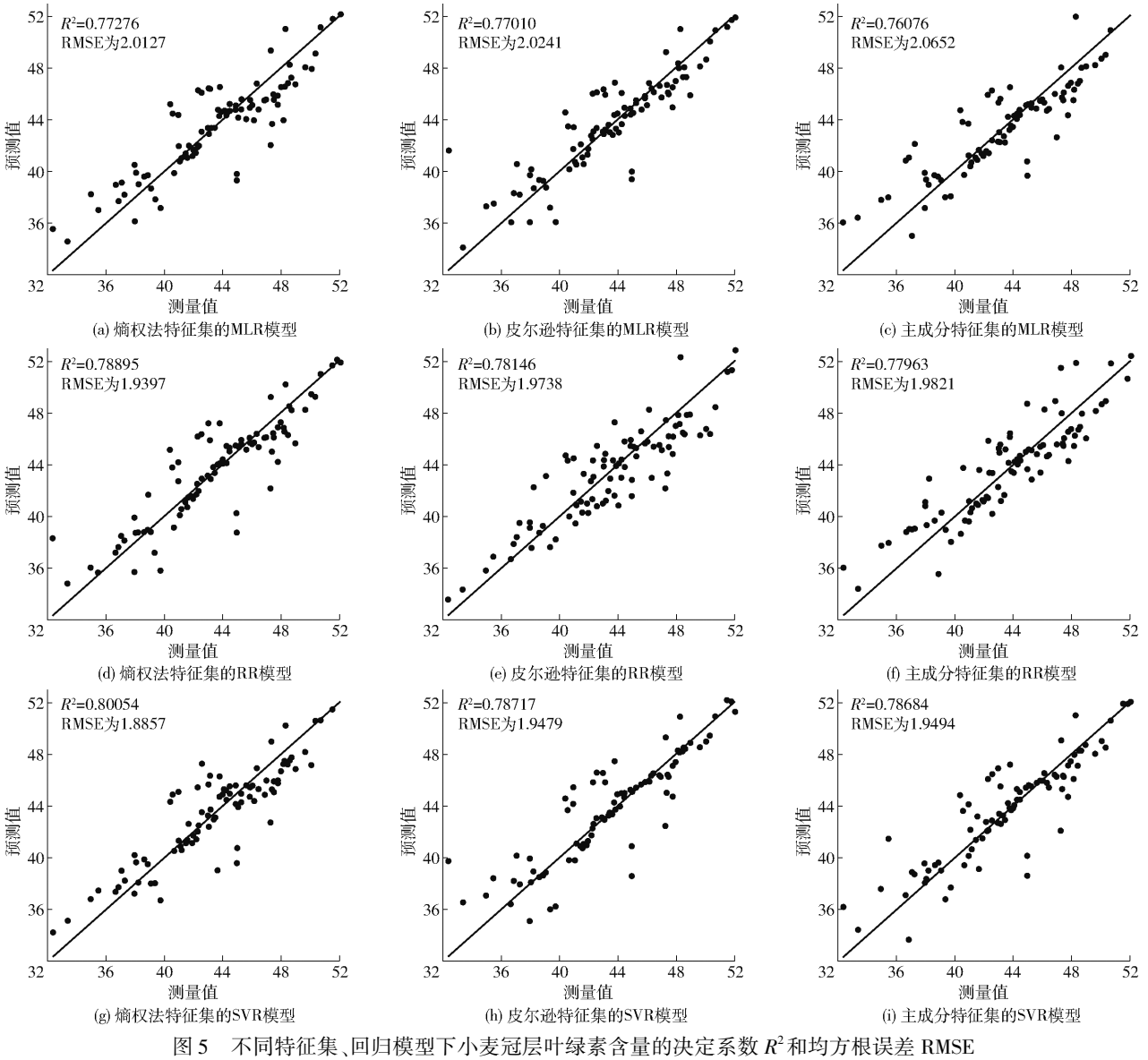


图 5 不同特征集、回归模型下小麦冠层叶绿素含量的决定系数 R^2 和均方根误差 RMSE

Fig. 5 Determination coefficient and root mean square error of wheat canopy chlorophyll content under different regression models and feature sets

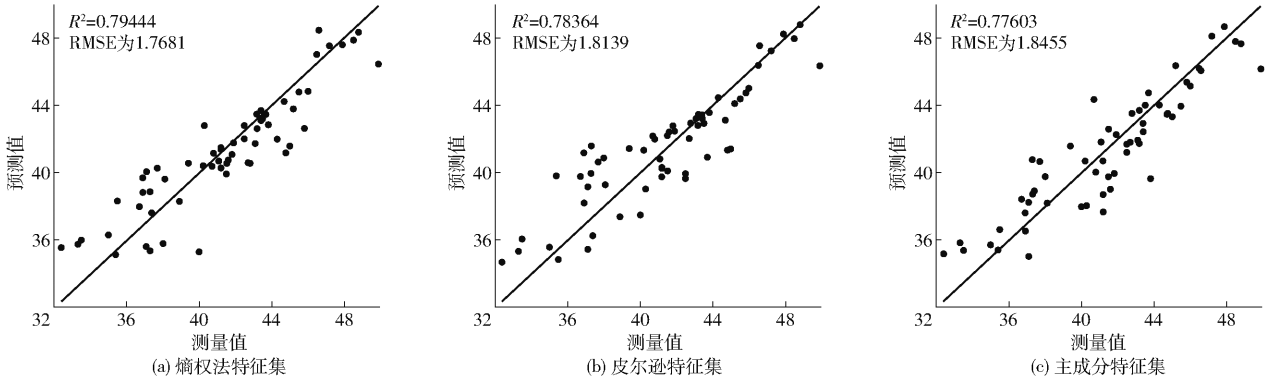
由图 5 可知,从模型角度分析,3 个特征集建立的 SVR 模型的 R^2 均大于 RR 和 MLR,RMSE 小于 RR 和 MLR,其中熵权法特征集 SVR 模型的 R^2 和 RMSE 分别约为 0.80、1.89,相较于 MLR 和 RR 模型,R^2 分别提升了约 2.8% 和 1.1%,RMSE 分别下降了约 0.13 和 0.05,故 3 种回归模型中 SVR 的预测能力最优。从特征集选取的角度分析,3 个特征集在不同估测模型中 R^2 相差不大,熵权法特征集的 RMSE 略小于皮尔逊特征集和主成分特征集,这说明熵权法特征集的泛化能力较优,模型预测精度较高。皮尔逊特征集可能仍存在特征变量数量过多的情况,特征间的冗余信息影响了试验结果。而主成分分析的特征集由

24 个特征变量降维到 3 个，降维后的主成分特征在最大程度上降低了变量间的相关关系，但是一些有利于 SPAD 值预测的特征信息不能完全体现出来。主成分特征集的特征指标个数最少，在降低模型复杂度的同时有效提高了模型的运行效率。

进一步分析，由于 RR 在 MLR 的基础上添加了偏差因子，其预测能力相较于 MLR 有所提高，但仍不能有效避免共线性问题。而 SVM 最终决策函数只由支持向量所确定，计算复杂性取决于支持向量的数目，而不是样本中所有特征，这在某种意义上对特征进行了降维处理，在一定程度上减少了特征间的共线性，从而提高了模型预测能力。

2.4　模型检验

为检验特征选择方法的有效性和模型的可靠性，2021 年 4 月 10 日，在同一试验田的各施肥小区，按照同样方法随机采集了 60 幅冠层图像。将所有 60 个样本采用十折交叉验证的方式用于验证熵权法特征集、皮尔逊特征集和主成分特征集建立的 SVR 模型。将 24 个特征指标通过逐步线性回归的方法确定出熵权法最优特征集为 $(a^*+b^*)/L$、$R-B-G$、$R-G$、$(R-G-B)/(R+B)$、$(R-G)/(R+G)$、a^*、a^*/b^*、$(R-G)/(R+G+B)$、DGCI、$(R-B)/(R+B)$，皮尔逊最优特征集为 R/B、DGCI、$(R-B)/(R+B)$、$(G+B-R)/(2G)$、$(a^*+b^*)/L$、$B/(R+G+B)$、$(G+B-R)/(2R)$、b^*、$R/\sqrt{R^2+G^2+B^2}$、a^*/b^*、$(R-G-B)/(R+B)$，主成分特征集选取 3 个主成分，累计贡献率为 99.5%。由图 6 可以看出，熵权法特征集结果 $R^2=0.794\,44$，RMSE 为 1.768 1，皮尔逊特征集结果 $R^2=0.783\,64$，RMSE 为 1.813 9，主成分特征集结果 $R^2=0.776\,03$ 和 RMSE 为 1.845 5，表明熵权法仍然能够使用较少的特征集获得较好的效果，而且 SVR 模型在 2021 年数据集上表现较好。



(a) 熵权法特征集　(b) 皮尔逊特征集　(c) 主成分特征集

图 6　支持向量回归模型验证数据集效果

Fig. 6　Validation data set effects of SVR

3　讨论

在利用图像处理技术对大田作物进行叶绿素含量估测的研究中，对于颜色特征的定义和构造已经有许多成果，从目前已有的文献研究成果中，发现构建的颜色特征参数至少有 50 多种。有不少学者利用皮尔逊相关系数和主成分分析进行特征选择，然而借助信息熵进行特征选择的相关研究还比较少。本文利用信息熵、皮尔逊和主成分分析特征选择方法分别确定了最优特征集对小麦 SPAD 值进行预测，其中信息熵、皮尔逊特征集都包含的特征指标有 4 个，分别是 $(a^*+b^*)/L$、a^*/b^*、$(R-G)/(R+G+B)$、$(R-G)/(R+G)$，说明 2 个特征集有交集，而主成分特征集大幅度减少了变量维数，将 3 种特征选择方法的结合使用还需要进一步探究。另外，本文构造的 $(a^*+b^*)/L$、a^*/b^* 2 个特征对 SPAD 值预测起了积极促进的作用，接下来可以在 La^*b^* 空间上探索预测能力更好的特征。

从 3 种方法选取的最佳特征集看，熵权法选取的特征集在 3 个颜色空间中都有所分布，9 个特征时得到了最优预测模型，皮尔逊特征集中的特征指标大都分布在 RGB 颜色空间，12 个特征时出现了最优预测模型。这可能是由于不同颜色空间特征指标的相关性较小，来自多个空间的特征可以在一定程度上减少这些共线性，因此信息熵特征集所用指标个数较少便达到了最佳预测效果。可以进一步研究该特征选择方法是否适用于小麦单叶片、其它大田作物 SPAD 值的预测。

小麦冠层图像是在自然环境下采集的，受到光照、天气的影响，对颜色特征参数的提取会造成一定的误差，在图像的分割算法上还需要进一步优化。建模方法的选择对预测精度具有一定的影响，近年来，国内外学者开始利用机器学习算法进行大田作物 SPAD 值的预测，与应用较广泛的 MLR 模型相

比,机器学习回归模型具有更佳的稳健性。本文建立的 SVR 相较于 MLR 和 RR,拟合能力最优,但是其惩罚因子 C 和核函数参数 σ 的选择还需要进一步探索。通过研究建模算法对预测效果的影响,来提高模型预测能力。

4 结论

(1)引用信息熵思想提出了熵权法的特征选择方法,确定了 a^*、$R-G-B$、$R-G$、$(a^*+b^*)/L$、a^*/b^*、$(R-G)/(R+G+B)$、$(R-B)/(R+B)$、H/S、$(R-G)/(R+G)$一组特征集,通过对比皮尔逊特征选择方法,表明熵权法在选取较少的特征指标下建模便能达到较好的预测效果。

(2)构造了 La^*b^* 空间下$(a^*+b^*)/L$、a^*/b^* 2 个图像特征指标,熵权法和皮尔逊特征选择方法特征集均包含这 2 个特征指标,说明其对 SPAD 值预测起了积极促进作用。

(3)运用 3 个特征集分别建立了多元线性回归、岭回归、支持向量回归的小麦冠层叶绿素含量估测模型,在不同的模型上,熵权法选取的特征集均表现了较好的预测效果,从模型角度分析,支持向量回归的预测能力和泛化能力最优,R^2 和 RMSE 分别约为 0.80、1.89,相比于 MLR 和 RR 模型 R^2 分别提升约 2.8%、1.1%,RMSE 分别下降了约 0.13 和 0.05。

参 考 文 献

[1] LIU Chuang, LIU Yi, LU Yanhong, et al. Use of a leaf chlorophyll content index to improve the prediction of above-ground biomass and productivity[J]. Peer J, 2019, 6: 6240.

[2] 李长春,施锦锦,马春艳,等. 基于小波变换和分数阶微分的冬小麦叶绿素含量估算[J]. 农业机械学报,2021,52(8):172-182.
LI Changchun, SHI Jinjin, MA Chunyan, et al. Estimation of chlorophyll content in winter wheat based on wavelet transform and fractional differential[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(8): 172-182. (in Chinese)

[3] 王丽爱,马昌,周旭东,等. 基于随机森林回归算法的小麦叶片 SPAD 值遥感估算[J]. 农业机械学报,2015,46(1):259-265.
WANG Liai, MA Chang, ZHOU Xudong, et al. Remote sensing estimation of SPAD value of wheat leaves based on random forest regression algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 259-265. (in Chinese)

[4] JOSE F R, CHRISTIAN C, JAVIER Z. Reliability of different color spaces to estimate nitrogen SPAD values in maize[J]. Computers and Electronics in Agriculture, 2017, 143: 14-22.

[5] 李红军, 张立周, 陈曦鸣,等. 应用数字图像进行小麦氮素营养诊断中图像分析方法的研究[J]. 中国生态农业学报, 2011, 19(1): 155-159.
LI Hongjun, ZHANG Lizhou, CHEN Ximing, et al. Research on the image analysis method in the diagnosis of wheat nitrogen nutrition using digital images[J]. Chinese Journal of Eco-Agriculture, 2011, 19(1): 155-159. (in Chinese)

[6] 张立周, 侯晓宇, 张玉铭, 等. 数字图像诊断技术在冬小麦氮素营养诊断中的应用[J]. 中国生态农业学报, 2011, 19(5): 1168-1174.
ZHANG Lizhou, HOU Xiaoyu, ZHANG Yuming, et al. Application of digital image diagnosis technology in winter wheat nitrogen nutrition diagnosis[J]. Chinese Journal of Eco-Agriculture, 2011, 19(5): 1168-1174. (in Chinese)

[7] 程立真, 朱西存, 高璐璐, 等. 基于 RGB 模型的苹果叶片叶绿素含量估测[J]. 园艺学报, 2017, 44(2): 381-390.
CHENG Lizhen, ZHU Xicun, GAO Lulu, et al. Estimation of chlorophyll content in apple leaves based on RGB model[J]. Acta Horticulture, 2017, 44(2): 381-390. (in Chinese)

[8] 陈佳悦. 小麦冠层图像处理及氮素图像评价指标研究[D]. 南京:南京农业大学, 2017.
CHEN Jiayue. Research on wheat canopy image processing and nitrogen image evaluation index[D]. Nanjing: Nanjing Agricultural University, 2017. (in Chinese)

[9] 宋宇斐. 基于数字图像的小麦营养诊断研究[D]. 保定:河北农业大学, 2020.
SONG Yufei. Research on wheat nutrition diagnosis based on digital image[D]. Baoding: Hebei Agricultural University, 2020. (in Chinese)

[10] 周利亚, 苑迎春, 宋宇斐, 等. 基于图像处理的小麦叶绿素估测模型研究[J]. 河北农业大学学报, 2018, 41(2): 105-109.
ZHOU Liya, YUAN Yingchun, SONG Yufei, et al. Research on wheat chlorophyll estimation model based on image processing [J]. Journal of Hebei Agricultural University, 2018, 41(2): 105-109. (in Chinese)

[11] LINDSEY A J, STEINKE K, RUTAN J, et al. Relationship of DGCI and SPAD values to corn grain yield in the eastern corn belt[J]. Crop, Forage & Turfgrass Management, 2016, 2(1): 1-9.

[12] GUPTA S D, PATTANAYAK A K. Intelligent image analysis (IIA) using artificial neural network (ANN) for non-invasive estimation of chlorophyll content in micropropagated plants of potato[J]. In Vitro Cellular & Developmental Biology, 2017, 53(6): 520-526.

[13] 王方永, 王克如, 李少昆, 等. 利用数码相机和成像光谱仪估测棉花叶片叶绿素和氮素含量[J]. 作物学报, 2010, 36(11): 1981-1989.
WANG Fangyong, WANG Keru, LI Shaokun, et al. Estimating the chlorophyll and nitrogen content of cotton leaves using digital cameras and imaging spectrometers[J]. Acta Agronomica Sinica, 2010, 36(11): 1981-1989. (in Chinese)

[14] 柴阿丽, 李宝聚, 王倩, 等. 基于计算机视觉技术的番茄叶片叶绿素含量的检测[J]. 园艺学报, 2009, 36(1): 45-52.
CHAI Ali, LI Baoju, WANG Qian, et al. Detection of chlorophyll content of tomato leaves based on computer vision technology[J]. Acta Horticulture, 2009, 36(1): 45-52. (in Chinese)

[15] 钱争, 冯绍元, 庄旭东, 等. 基于主成分分析的土壤盐碱化特征研究[J]. 中国农村水利水电,2022(5):119-124.

QIAN Zheng, FENG Shaoyuan, ZHUANG Xudong, et al. Research on the characteristics of soil salinization based on principal component analysis[J]. China Rural Water and Hydropower, 2022(5):119 - 124. (in Chinese)
[16] 齐振,程广涛,张友奎, 等. 基于信息熵的水声目标识别模型评估方法[J]. 舰船科学技术,2021,43(11):134 - 137.
QI Zhen, CHENG Guangtao, ZHANG Youkui, et al. Evaluation method of underwater acoustic target recognition model based on information entropy[J]. Ship Science and Technology,2021,43(11):134 - 137. (in Chinese)
[17] 唐胜雨,哈丽旦木·托呼提买提,张鹏, 等. 基于信息熵隶属度的油田集输管道风险评价方法[J]. 油气储运,2021,40(7):761 - 767.
TANG Shengyu, HALI Danmu · Tohutimaiti, ZHANG Peng, et al. Risk assessment method of oilfield gathering and transportation pipeline based on membership degree of information entropy[J]. Oil & Gas Storage and Transportation,2021,40(7):761 - 767. (in Chinese)
[18] ROOSJEN P P, BREDE B, SUOMALAINEN J M, et al. Improved estimation of leaf area index and leaf chlorophyll content of a potato crop using multi-angle spectral data-potential of unmanned aerial vehicle imagery[J]. International Journal of Applied Earth Observations and Geoinformation,2018,66:14 - 26.
[19] SONG Y F, TENG G F, YUAN Y C, et al. Assessment of wheat chlorophyll content by the multiple linear regression of leaf image features[J]. Information Processing in Agriculture, 2021,8(2):12 - 23.
[20] 刘斌, 毕小熊, 党军朋, 等. 基于支持向量回归的变电站蓄电池退化趋势预测[J]. 电源学报, 2020, 18(6): 207 - 214.
LIU Bin, BI Xiaoxiong, DANG Junpeng, et al. Prediction of substation battery degradation trend based on support vector regression[J]. Journal of Power Sources, 2020, 18(6): 207 - 214. (in Chinese)
[21] 黄芬, 高帅, 姚霞, 等. 基于机器学习和多颜色空间的冬小麦叶片氮含量估算方法研究[J]. 南京农业大学学报, 2020, 43(2): 364 - 371.
HUANG Fen, GAO Shuai, YAO Xia, et al. Research on estimation method of winter wheat leaf nitrogen content based on machine learning and multi-color space[J]. Journal of Nanjing Agricultural University, 2020, 43(2): 364 - 371. (in Chinese)
[22] 梁亮, 杨敏华, 张连蓬, 等. 基于 SVR 算法的小麦冠层叶绿素含量高光谱反演[J]. 农业工程学报, 2012, 28(20): 162 - 171.
LIANG Liang, YANG Minhua, ZHANG Lianpeng, et al. Hyperspectral inversion of chlorophyll content in wheat canopy based on SVR algorithm[J]. Transactions of the CSAE, 2012, 28(20): 162 - 171. (in Chinese)
[23] 田杰,韩冬,胡秋霞,等. 基于 PCA 和高斯混合模型的小麦病害彩色图像分割[J]. 农业机械学报,2014,45(7):267 - 271.
TIAN Jie, HAN Dong, HU Qiuxia, et al. Segmentation of wheat rust lesion image using PCA and Gaussian mix model[J]. Transactions of the Chinese Society for Agricultural Machinery,2014,45(7):267 - 271. (in Chinese)
[24] 张彦娥, 李民赞, 张喜杰, 等. 基于计算机视觉技术的温室黄瓜叶片营养信息检测[J]. 农业工程学报, 2005,21(8): 102 - 105.
ZHANG Yan'e, LI Minzan, ZHANG Xijie, et al. Detection of nutrition information of greenhouse cucumber leaves based on computer vision technology[J]. Transactions of the CSAE, 2005,21(8): 102 - 105. (in Chinese)
[25] AGARWAL A, GUPTA S D. Assessment of spinach seedling health status and chlorophyll content by multivariate data analysis and multiple linear regression of leaf image features[J]. Computers and Electronics in Agriculture, 2018, 152: 281 - 289.
[26] 王克如, 李少昆, 王崇桃, 等. 用机器视觉技术获取棉花叶片叶绿素浓度[J]. 作物学报, 2006,32(1): 34 - 40.
WANG Keru, LI Shaokun, WANG Chongtao, et al. Obtaining the chlorophyll concentration of cotton leaves with machine vision technology[J]. Acta Agronomica Sinica, 2006,32(1): 34 - 40. (in Chinese)
[27] 张沛健, 尚秀华, 吴志华. 基于图像处理技术的 5 种红树林叶片形态特征及叶绿素相对含量的估测[J]. 热带作物学报, 2020, 41(3): 496 - 503.
ZHANG Peijian, SHANG Xiuhua, WU Zhihua. Estimation of leaf morphology and relative chlorophyll content of five mangrove species based on image processing technology[J]. Chinese Journal of Tropical Crops, 2020, 41(3): 496 - 503. (in Chinese)
[28] 江朝晖, 杨春合, 周琼, 等. 基于图像特征的越冬期冬小麦冠层含水率检测[J]. 农业机械学报, 2015, 46(12): 260 - 267.
JIANG Zhaohui, YANG Chunhe, ZHOU Qiong, et al. Detection of winter wheat canopy moisture content in winter based on image features[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(12): 260 - 267. (in Chinese)
[29] 时雷, 庞晓丹, 钱诚, 等. 基于图像处理技术的小麦群体叶绿素状况估计研究[J]. 太原理工大学学报, 2016, 47(2): 223 - 227.
SHI Lei, PANG Xiaodan, QIAN Cheng, et al. Study on the estimation of wheat population chlorophyll status based on image processing technology[J]. Journal of Taiyuan University of Technology, 2016, 47(2): 223 - 227. (in Chinese)
[30] SOLTANGHARAEI V, AI L, ANAY R, et al. Implementation of information entropy, b-value, and regression analyses for temporal evaluation of acoustic emission data recorded during ASR cracking[J]. Practice Periodical on Structural Design and Construction, 2021, 26(1):191 - 199.
[31] GEORGII M T, PILGE S, SCHNEIDER G,et al. State entropy and burst suppression ratio can show contradictory information: a retrospective study[J]. European Journal of Anaesthesiology, 2020,37(1):353 - 360.
[32] SON J, LEE J, LARSEN K R, et al. Understanding the uncertainty of disaster tweets and its effect on retweeting: the perspectives of uncertainty reduction theory and information entropy[J]. Journal of the Association for Information Science and Technology, 2020, 71(10):121 - 129.
[33] YANG B, GAN D Y, TANG Y C, et al. Incomplete information management using an improved belief entropy in dempster-shafer evidence theory[J]. Entropy, 2020, 22(9):993 - 1001.
[34] 王俊迪, 许蕴山, 彭芳, 等. 基于岭回归的红外协同定位优化算法[J]. 北京航空航天大学学报, 2020, 46(3): 563 - 570.
WANG Jundi, XU Yunshan, PENG Fang, et al. Infrared co-location optimization algorithm based on ridge regression[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(3): 563 - 570. (in Chinese)