



王克晓,周蕊,李波.一种基于 SC-StackingPSO-MI 的油菜籽粒表观特征筛选方法[J].南京农业大学学报,2026,49(3):676-684.

Wang K X,Zhou R,Li B. A method of screening phenotypic features of rapeseeds based on SC-StackingPSO-MI[J]. Journal of Nanjing Agricultural University,2026,49(3):676-684.

一种基于 SC-StackingPSO-MI 的油菜籽粒 表观特征筛选方法

王克晓*,周蕊,李波

(重庆市农业科学院农业科技信息研究所,重庆 401329)

摘要:[目的]为实现油菜籽粒关键表观特征的有效筛选,服务于油菜优良种质资源快速筛选,提出了一种基于谱聚类(spectral clustering,SC)-堆叠集成学习粒子群优化(stacking particle swarm optimization,StackingPSO)-互信息(mutual information,MI)的油菜籽粒关键表观特征筛选方法 SC-StackingPSO-MI。[方法]在确定油菜籽粒最优谱聚类的基础上,结合特征与其聚类结果的 MI 值,基于逻辑回归(logistic regression,LR)算法联合 k 最近邻(k-nearest neighbor,KNN)、随机森林(random forest,RF)和支持向量机(support vector machine,SVM)3 种基础学习器设计 Stacking 集成学习模型,用于优化 PSO 中适应度函数进行油菜籽粒关键表观特征的筛选,并基于优选特征开展油菜籽粒聚类分析、不同分类器以及不同特征选择算法下的油菜籽粒分类精度比对。[结果]二分聚类条件下,SC-StackingPSO-MI 特征筛选算法能够很好地提取出油菜籽粒表观特征中统计方差较大的特征指标,优选特征谱聚类下的 Calinski-Harabasz(CH)聚类指数从 42.89 提高到 83.90,特征聚类性能得到显著改善。基于二分聚类标签对比不同分类器下 SC-StackingPSO-MI 油菜籽粒筛选特征的整体分类测试效果,除 RF 模型外的 KNN、SVM 和 LR 等分类模型的测试精度相较于全特征组均有不同程度提高,特征筛选结果具有较好的稳健性。与 LASSO、SVM-RFE 和 XGBoost 3 种特征选择方法对比,SC-StackingPSO-MI 优选特征组合下,基于二分聚类标签的 SVM 模型测试集精度准确率(ACC)和 Kappa 值分别为 97.22%和 93.88%,均高于 LASSO、SVM-RFE 和 XGBoost 优选特征组合下的测试精度,特征筛选方法具有较高的可靠性。[结论]本文提出的特征筛选方法在油菜籽粒关键表观特征选择中展现出较好的优越性,可为油菜籽粒以及其他作物优良种质资源筛选中的关键表观特征选取提供相关技术参考。

关键词:关键表观特征选择;油菜籽粒;粒子群优化;互信息;集成学习

中图分类号:TP399;S565.4

文献标志码:A

文章编号:1000-2030(2026)03-0676-09

A method of screening phenotypic features of rapeseeds based on SC-StackingPSO-MI

WANG Kexiao*,ZHOU Rui,LI Bo

(Institute of Agricultural Science and Technology Information,Chongqing Academy of
Agricultural Sciences,Chongqing 401329,China)

Abstract:[Objectives] To extract effectively the key phenologic traits of rapeseeds and serve the rapid screening of excellent rapeseed germplasm resources,based on spectral clustering(SC),stacking particle swarm optimization(StackingPSO) and mutual information(MI),a method of rapeseeds key phenotypic trait character screening SC-StackingPSO-MI was proposed. [Methods] After the determination of the optimal spectral clustering of rapeseed grains,this paper designed the Stacking ensemble learning model based on the logistic regression(LR) algorithm combining with k-nearest neighbor(KNN),random forest(RF),and support vector machine(SVM) to evaluate the fitness function in the particle swarm optimization(PSO) for the screening of key apparent trait features of rapeseed grains by combining the MI between the features and their clustering results. Furthermore,clustering analysis of rapeseed grains,the comparison of the classification accuracies of rapeseed grains under different classifiers and different feature selection algorithms based on the selected key features were carried out. [Results] Under the binary clustering condition,the SC-StackingPSO-MI feature selection algorithm could effectively extract the feature indicators with high statistical variance from the apparent features of rapeseeds. The Calinski-Harabasz(CH) indices with the optimized features increased from 42.89 to 83.90,which

收稿日期:2025-02-13

基金项目:重庆市市级财政科研项目(cqaas2023sjczhx003);重庆市自然科学基金项目(CSTB2024NSCQ-MSX0804);农业农村部农业监测预警技术重点实验室开放课题(KLAMEWT202402)

*通信作者:王克晓,硕士,助理研究员,研究方向为农业信息化,E-mail:wangkexiao_2007@126.com。

resulted in a significant improvement in the clustering performance. The overall classification test performance of rapeseeds screening features from SC-StackingPSO-MI under different classifiers was compared based on binary clustering labels. Compared with the full feature combination, three classification models such as KNN, SVM and LR all showed varying degrees of improvement in testing accuracy apart from RF model. That proved the feature selection results had good robustness. Compared to the feature selection methods of LASSO, SVM-RFE and XGBoost under binary clustering labels, the SVM model's test accuracy (ACC) and Kappa values were 97.22% and 93.88%, respectively, indicating the higher reliability. [Conclusions] The feature screening method proposed in this paper had demonstrated good superiority in the selection of key apparent trait features of rapeseed grains, and could provide relevant technical references for the selection of key apparent trait characteristics in the screening of excellent germplasm resources of rapeseed grains and other crops.

Keywords: selection of key epigenetic trait characteristics; rapeseeds; particle swarm optimization; mutual information; ensemble learning

作物籽粒是下一代植株幼体生长发育的基础,其大小、形状、颜色等表观特征在一定程度上决定了作物的生长发育史,对作物生理学、栽培学以及高产育种研究等都具有重要的指导意义^[1]。目前,我国农作物籽粒的表观特征检测普遍采用有经验人员通过肉眼目测进行主观鉴定,无法满足客观、快速、准确、稳定的检测要求^[2]。

近年来,计算机图像处理和 multispectral 成像技术的发展,为作物籽粒表观特征的精准化、快速化获取提供了技术支撑,极大地提高了作物选种、育种工作效率^[3-4]。Fu 等^[5]根据燕麦的形态特征、光谱特征及其组合,分析了燕麦种子品种之间的差异,并利用线性判别分析(linear discriminant analysis, LDA)和支持向量机(support vector machine, SVM)完成了籽粒关键表观特征的选择。Zhang 等^[6]比对了不同成熟度和不同收获期年份的苜蓿种子形态特征和光谱特征差异,并利用主成分分析(principal component analysis, PCA)、LDA、SVM 和随机森林(random forest, RF)等方法进行了特征筛选,并对其活力开展了判别和预测。冯继克等^[7]采用 SelectKBest、PCA 和 LDA 等特征选择模型在小麦籽粒的颜色、形态和纹理等 28 个特征中开展了优化筛选,并对比了不同特征子集组合下的籽粒品种识别精度。黄琼等^[8]基于 4 类水稻种子图像提取出的颜色、几何和纹理特征,提出了一种基于 LDA 特征优化和贝叶斯分类算法相结合的水稻种子品种识别方法。孟惜等^[9]通过提取小麦籽粒的形态、颜色、纹理等特征,结合 BP 神经网络、主成分分析法和粒子群优化(particle swarm optimization, PSO)算法,达到了较好的品种识别效果。上述诸多的作物籽粒表观特征研究多用于作物籽粒的品种识别及其活力检测。油菜是我国重要的油料作物,与小麦、水稻等籽粒不同,其同一品种的籽粒颜色与大小等表观特征存在较大差异,致使不同品种的油菜籽粒也难以区分。因此,关于油菜籽粒品种识别的相关研究鲜有报道。另外,在作物种质资源筛选方面,优良作物籽粒筛选的关键任务之一是获取作物籽粒的关键表观特征。因此,伴随着数字化育种趋势,利用机器学习技术手段开展油菜籽粒关键表观特征筛选,可辅助相关科技人员实现特定品种油菜籽粒群体内优良资源的快速筛选。

为实现油菜籽粒关键表观特征的有效筛选,本研究提出了一种基于谱聚类的集成学习粒子群优化与互信息结合的油菜籽粒表观特征筛选方法(SC-StackingPSO-MI),实现了油菜籽粒关键表观特征的自动筛选,可为作物籽粒优良种质资源筛选过程中的关键表观特征获取提供技术参考。

1 研究方法

1.1 特征筛选方法

本研究特征筛选的技术路线如图 1 所示,主要包括油菜籽粒谱聚类、粒子群优化适应度函数确定、互信息阈值确定、堆叠集成学习模型计算粒子群迭代适应度值和随机多循环粒子群最优解搜索等过程。上述算法利用 Python 3.9.16 编程实现。

1.1.1 谱聚类 谱聚类(spectral clustering, SC)是一种基于图论的聚类方法,可以利用数据间的相互依赖程度将数据划分为多个聚类,能够有效处理非线性和高维数据^[10]。Calinski-Harabasz (CH) 指数是一种聚类的评估标准,通过簇间离散度与簇内离散度的比值,判断聚类结果集中簇内的密合程度以及各簇间的分离程度^[11]。为确定合适的油菜籽粒分类簇数,本文采用 CH 指数作为度量谱聚类效果的性能指标。CH 指数值越高,代表聚类效果越好。

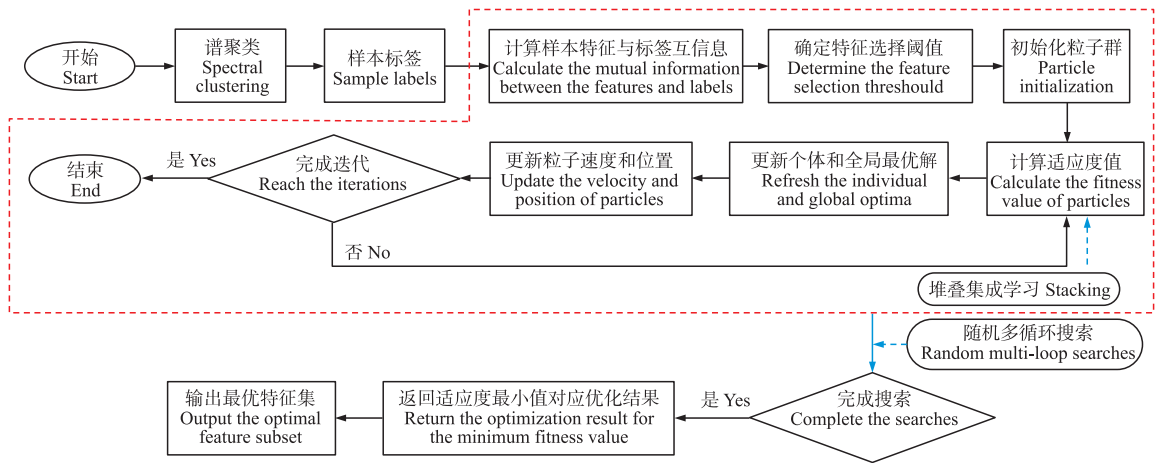


图 1 特征筛选算法流程图

Fig. 1 Flow chart of feature filtering algorithm

1.1.2 粒子群优化适应度函数确定 粒子群优化(PSO)算法是一种模拟鸟群搜索食物行为的经典群体智能优化算法,具有参数相对较少、全局搜索能力强和收敛速度快等特点^[12]。在 PSO 算法迭代优化中,粒子的每个位置代表特征子集组合的一组潜在解,其“优劣”程度可以通过适应度函数值来确定。本研究 PSO 算法适应度函数采用分类错误率(*Error*)和特征选中率(*Feature_{rate}*)加权值,其表达式如公式(1)—公式(3)所示。*Fitness* 值越小,表明粒子位置对应的特征子集越优。

$$Fitness = \alpha \times Error + (1 - \alpha) \times Feature_{rate} \quad (1)$$

$$Error = 1 - Accuracy \quad (2)$$

$$Feature_{rate} = SF/B \quad (3)$$

式中:*Accuracy* 为机器学习模型的预测准确率;*SF* 表示选择出的特征数量;*B* 表示特征总数; α 为分类错误率 *Error* 的权重因子,文中取值 0.95。

1.1.3 堆叠集成学习构建 堆叠(Stacking)集成学习算法是一种含有基学习器和元学习器的多层集成学习框架,能够将各个基学习器的输出作为特征变量输入来训练元学习器,从而降低单一分类模型预测结果的不确定性,有效提高学习模型整体的精度和泛化能力^[13]。本研究选用 *k* 最近邻(*k*-nearest neighbor, KNN)、RF 和 SVM 3 种学习算法作为基学习器,并以逻辑回归(logistic regression, LR)算法为元学习器构建集成学习模型进行模型训练、预测,计算不同特征组合迭代时的分类错误率(*Error*),并结合特征数量(*SF*)进一步确定每次粒子迭代的适应度值。

1.1.4 互信息阈值确定 互信息(mutual information, MI)是信息论中一种用于衡量 2 个随机变量之间关联程度的指标,已经广泛地应用于机器学习中特征之间相关性度量^[14]。MI 值越大,表示 2 个随机变量之间的关联度越高,反之越低。在特征筛选中,某个特征是否被选中主要依赖于其与类标签间阈值(∂)的大小。为了更好地选择阈值,本研究引入互信息阈值自适应策略^[15]来衡量各个特征与类标签之间的相关性。阈值定义如下:

$$\partial_b = 1 - \lambda \frac{I(f_b, C)}{\max(I(f, C))}, b \in \{1, 2, \dots, B\} \quad (4)$$

式中: $I(f_b, C)$ 表示特征 f_b 与类标签 C 之间的互信息; B 为特征数量; λ 为削减因子。过小或过大的 λ 值会导致特征入选的数量过少或过多,从而影响 Stacking 模型下 PSO 算法的性能。

超体积(hyper volume, HV)是一种在多目标优化领域中用于评估解集质量的指标。HV 值越大,对应特征选择算法的综合性越好^[16]。本研究通过设定不同的削减因子(λ),比较 Stacking 学习模型下 PSO 特征选择算法的整体性来确定 λ 的取值。

1.1.5 多循环粒子群最优解搜索 在 PSO 算法中,粒子位置、速度和部分参数初始化过程的随机性会导致每次算法优化结果的不稳定。如果粒子在搜索最优值时过早收敛,也会使算法的寻优停滞在局部最小值,从而无法找到全局最优解^[17]。为有效获取粒子群最优解,本研究在增加粒子群搜索迭代次数的同时,采用多次随机循环粒子群优化(recurrent particle swarm optimization, RPSO)搜索策略,在捕获的多个

局部最小值结果中筛选适应度函数值最小的粒子搜索结果,确定特征入选数量和最优特征组合。针对可能存在的多个特征子集对应同一最小适应度函数值的问题,以多个最优子集的并集作为最优特征集。

1.2 特征筛选组合评价方法

为评判上述特征筛选算法所确定的关键特征子集质量,本研究基于样本谱聚类标签,分别利用 KNN、RF、SVM、LR 4 种分类器在训练集样本上进行模型训练,并在测试集样本上进行验证,以分类准确率 (accuracy, ACC) 和 Kappa 值为指标开展油菜籽表观特征筛选前后的分类精度对比分析。同时,基于套索算法 (least absolute shrinkage and selection operator, LASSO)、支持向量机递归特征消除 (support vector machine-recursive feature elimination, SVM-RFE) 和极端梯度上升 (extreme gradient boosting, XGBoost) 3 种特征选择算法进行特征筛选,并利用 SVM 分类器进行测试集分类精度对比。

2 试验数据

试验材料选取 120 粒完整、未发芽、无霉变的‘庆油 8 号’(QY-8)油菜籽粒。采用 Videometer Lab 多光谱测量仪 (Videometer A/S, Herlev, 丹麦) 进行光谱成像 (图 2), 其光谱波段为 405、435、450、470、505、525、570、590、630、645、660、700、780、850、870、890、910、940、970 nm 19 个波长,能够在合成高分辨率多光谱图像的同时,快速准确地获取成像目标的颜色、形状、大小等表观特征指标。本研究在获取油菜籽粒的多光谱图像后,用典型判别分析进行图像降维,再利用目标籽粒与背景的光谱差异进行阈值分割,使油菜籽粒成为二值化图像中的 Blob 区域;通过 Blob 分析提取各籽粒的形态特征和 CIE 颜色空间分量数据,包括面积 (area)、长度 (length)、宽度 (width)、长宽比 (aspect ratio, AR)、圆紧密度 (compactness circle, CC)、椭圆紧密度 (compactness ellipse, CE)、形状参数 a (shape_a)、形状参数 b (shape_b)、垂直偏斜 (vertical skewness, VS)、垂直方向 (vertical orientation, VO) 10 个形态参数和 CIE[0]—CIE[9] 10 个颜色分量,总计 20 个油菜籽粒表观特征。

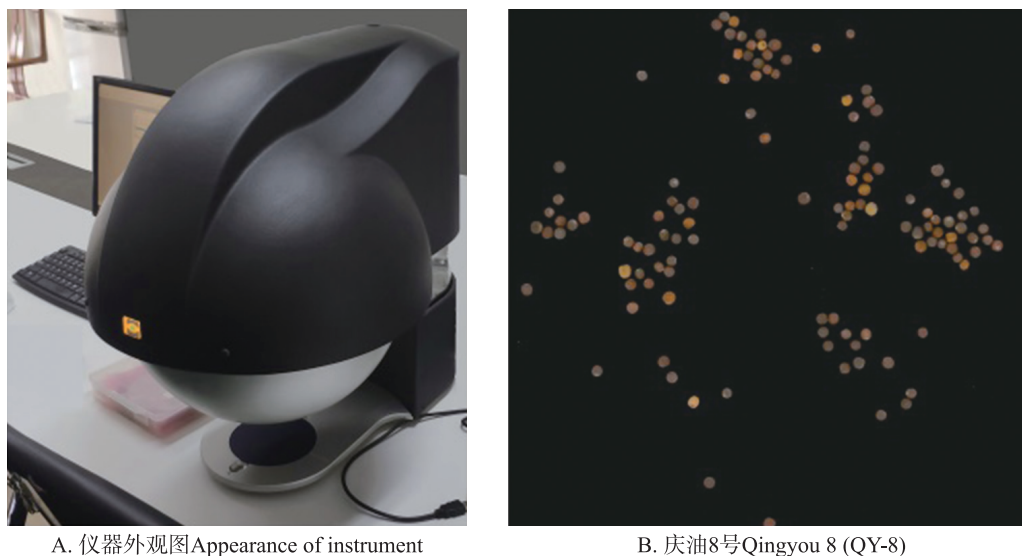


图 2 多光谱成像系统与样品 RGB 图像

Fig. 2 Multispectral imaging system and sample RGB image

油菜籽粒样本各个参数指标的最小值、最大值及方差统计如表 1 所示。对于近圆形的油菜籽粒而言,各样本间形态参数整体方差分布相对较小,大部分在 0.03 以内,差异性最大的为籽粒面积 Area,方差为 0.214;颜色参数中,差异性比较大的有 CIE[0]、CIE[2]、CIE[4]、CIE[6] 和 CIE[8] 5 个颜色分量,方差分别为 20.104、24.610、40.729、58.800 和 46.586。为消除不同表观参数之间的量纲差异,先用 min-max 方法对上述参数进行标准化处理,将各参数的值域范围变换至 0~1。本研究随机选取 70% 的籽粒作为训练样本用于模型的构建,剩下 30% 的油菜籽粒作为测试样本用于模型的预测精度评价。

表 1 ‘庆油 8 号’油菜籽粒样本表观特征指标统计

Table 1 Statistics of apparent characteristic indicators of ‘Qingyou 8’ (QY-8) rapeseed seed samples

类型 Type	指标 Index	最小值 Min	最大值 Max	方差 Variance	类型 Type	指标 Index	最小值 Min	最大值 Max	方差 Variance
形态 Morphology	面积/mm ² Area	1.815	4.284	0.214	颜色 Color	CIE[0]	16.897	47.565	20.104
	长度/mm Length	1.543	2.414	0.027		CIE[1]	1.530	8.746	2.364
	宽度/mm Width	1.425	2.256	0.027		CIE[2]	2.803	22.980	24.610
	AR	0.837	1.024	0.001		CIE[3]	1.157	5.659	0.917
	CC	0.835	0.995	0.001		CIE[4]	5.984	33.993	40.729
	CE	0.989	0.999	0.001		CIE[5]	1.780	8.701	2.491
	形状参数 a Shape_a	1.366	1.884	0.007		CIE[6]	6.772	38.357	58.800
	形状参数 b Shape_b	1.279	1.659	0.006		CIE[7]	1.661	9.993	3.015
	VS	-0.111	-0.001	0.001		CIE[8]	43.984	78.146	46.586
	VO	-0.368	0.336	0.011		CIE[9]	2.573	16.396	9.530

注:AR:长宽比 Aspect ratio;CC:圆紧密度 Compactness circle;CE:椭圆紧密度 Compactness ellipse;VS:垂直偏斜 Vertical skewness;VO:垂直方向 Vertical orientation.下同 The same as follows.

3 结果与分析

3.1 基于 SC 的油菜籽粒聚类划分

采用 SC 对样本进行聚类时,设置不同的簇数 k 会产生不同的聚类结果。本文利用 SC 算法并依据油菜籽粒表观特征进行聚类划分,并通过 tSNE(t-distributed stochastic neighbor embedding) 技术进行高维数据降维和可视化显示。图 3 展示了 QY-8 油菜籽粒聚类簇数 k 为 2、3、4、5 时样本的聚类效果,可以看出,随着聚类簇数的增加,CH 指数逐渐减小;当 k 为 2 时,CH 指数值最高为 42.89,油菜籽粒聚类效果最好。

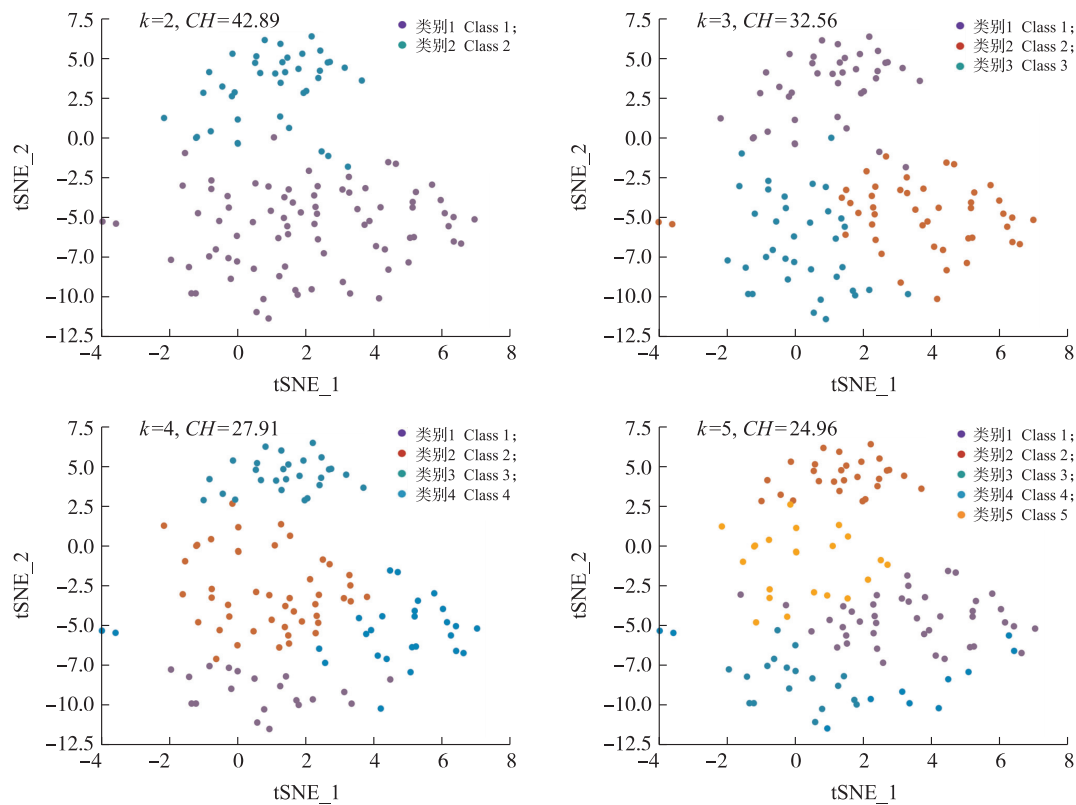


图 3 QY-8 油菜籽粒在不同簇数下的聚类分布

Fig. 3 Cluster distribution of QY-8 rapeseed seeds under different cluster numbers

3.2 籽粒关键特征筛选结果

设定 PSO 算法中粒子数量(N)为 20,惯性权重(ω)为 0.9,认知参数(c_1)和社会参数(c_2)均为 2,粒子最大飞行速度为 0.5,迭代次数为 200 次。由 KNN、RF 和 SVM 为基学习器,LR 算法为元学习器构建集成学习模型计算粒子群迭代搜索适应度值。各学习器的参数优化采用格网搜索与样本 5 折交叉验证,其中

KNN 的优化参数为最近邻样本数 $n_neighbors$, RF 模型优化参数为决策树数量 $n_estimators$ 和最大深度 max_depth , SVM 模型采用默认 rbf 核函数, 优化参数为惩罚因子 C 与核系数 $gamma$, LR 采用默认的 L2 正则化项, 搜索参数为正则化强度倒数 C 。为确定削减因子 λ 的取值, 比较不同 λ 值时以 Stacking 为学习模型的 PSO 特征选择算法的 HV 变化。如图 4 所示, 伴随着 λ 值的增大, HV 值整体上呈先增大后减小的趋势。当 $\lambda=0.6$ 时, HV 值达到最大(1.965)。因此, 本文确定削减因子 λ 的取值为 0.6。

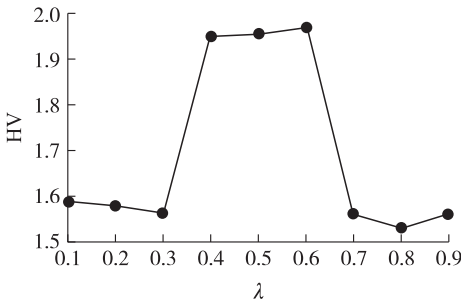


图 4 不同 λ 值对超体积(HV)的影响

Fig. 4 Influence of different λ values on hyper volume(HV)

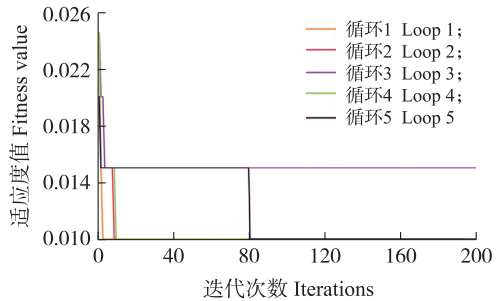


图 5 循环搜索收敛曲线

Fig. 5 Convergence curve of loop searches

基于分类簇数 $k=2$ 时的油菜籽粒训练样本的粒子适应度函数的收敛曲线如图 5 所示。随着迭代次数的增加, 粒子适应度值逐渐趋于稳定。通过粒子群随机 5 次循环搜索收敛曲线的比较, 4 次搜索结果的适应度函数值能够快速趋于一致, 达到相同的最小值 0.010, 说明在 200 次迭代范围内这 4 次独立的粒子群随机搜索结果具有较高的收敛一致性。基于循环粒子群进行优化搜索的特征选择结果如表 2 所示, 联合 4 次迭代搜索特征选择结果获取最优关键特征子集, 包括面积 $Area$ 、 $CIE[0]$ 、 $CIE[2]$ 、 $CIE[4]$ 、 $CIE[6]$ 和 $CIE[8]$ 这 6 个表观特征分量。对比表 1 可见, 基于 StackingPSO-MI 算法筛选的油菜籽粒关键特征子集主要包含了样本方差较大的颜色分量特征, 同时在油菜籽粒形态特征中优选出了具有相对较大方差的面积 $Area$, 显示了本文特征筛选结果的合理性。

表 2 基于循环粒子群优化搜索的特征选择

Table 2 Feature selection based on cyclic particle swarm optimization search

循环次数 Number of cycles	最小适应度值 Minimum fitness value	对应特征 Corresponding features	联合特征 Joint features
1	0.010	$Area, CIE[0], CIE[6]$	$Area, CIE[0], CIE[2], CIE[4], CIE[6], CIE[8]$
2	0.010	$CIE[4], CIE[8]$	
3	0.015	$CIE[2], CIE[4], CIE[7]$	
4	0.010	$CIE[2], CIE[6], CIE[8]$	
5	0.010	$Area, CIE[4], CIE[6]$	

3.3 基于优选特征的油菜籽粒聚类分析

利用油菜籽粒表观特征优化特征集, 再次采用谱聚类 SC 对样本进行簇数 $k=2$ 的聚类划分, 经过 tSNE 技术进行数据降维后可可视化显示, 并计算相应 CH 指数。由图 6 可以看出, 在优选特征组合下, 聚类数 $k=2$ 时的 CH 指数为 83.90, 较优化前的全特征聚类 CH 指数有显著提高, 表明本文算法能够有效捕捉油菜籽粒表观特征中的敏感特征, 改善油菜籽粒聚类性能。

3.4 基于不同分类器的特征筛选结果分析

为验证上述优选特征结果的有效性, 基于油菜籽粒二分类类别标签, 利用 KNN、RF、SVM 和 LR 4 种分类器在随机训练样本上对优选特征子集组合和原始全特征集组合开展籽粒类别模型构建与识别, 其中模型的参数优化采用格网搜索与样本 5 折交叉验证法, 并利用测试样本进行验证。由图 7 可以看出, 对 20 个油菜籽粒表观特征进行筛选后, 除了 RF 模型的准确率(ACC)和 Kappa 值有所降低外, 其他 3 个模型的精度指标相对于原始特征集均有不同程度提高, 表明基于优选特征的分类

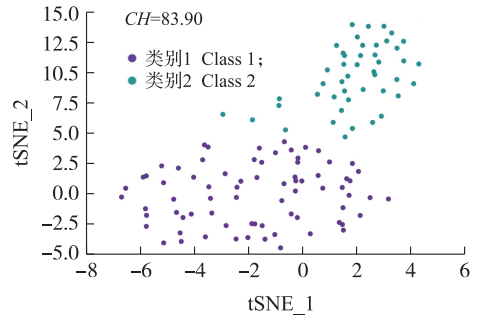


图 6 特征优化后油菜籽粒聚类分布

Fig. 6 Cluster distribution of rapeseeds after feature optimization

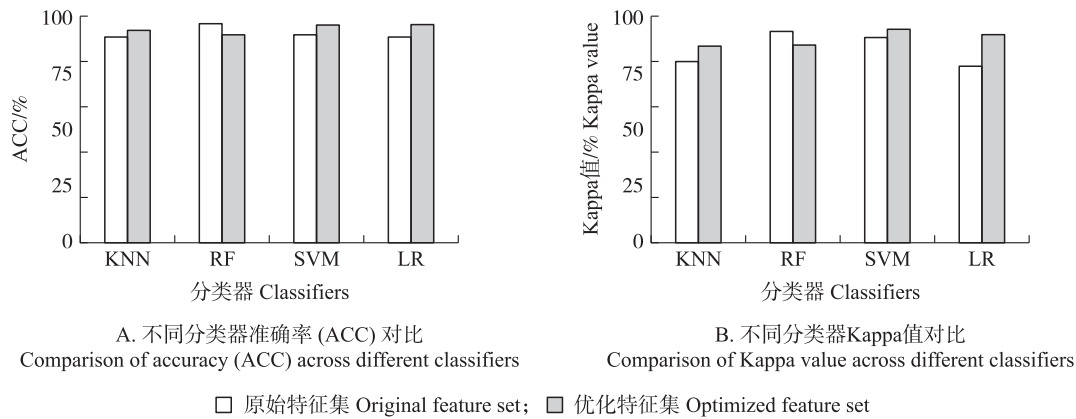


图 7 特征筛选前、后不同分类器分类精度对比

Fig. 7 Comparison of classification accuracy of different classifiers before and after feature screening

结果具有较好的稳健性。RF 模型对特征冗余有较强的鲁棒性,有助于提高模型的泛化能力,较容易获得较高的分类精度。本文原始特征集下的 RF 模型特征重要性排序如图 8 所示,排序前 6 位的特征分别为 CIE [4]、CIE [6]、CIE [2]、CIE [7]、CIE [9]和 CIE [5]。对比表 2 中所选特征结果,Area、CIE [0]和 CIE [8]的特征重要性明显较弱。因此,基于特征筛选结果进行的 RF 模型分类精度降低的原因极大可能是 CIE [7]、CIE [9]和 CIE [5]等 RF 模型中相对较为重要的特征被筛选过滤导致。

3.5 基于不同特征选择算法的精度对比

LASSO 算法和 SVM-RFE 算法是 2 种常用的特征选择方法。XGBoost 是一种基于梯度提升框架 (gradient boosting) 的决策树集成学习方法,拥有内置的特征重要性计算方法。本文分别采用 LASSO、SVM-RFE、XGBoost 和 StackingPSO-MI 这 4 种特征选择方法在随机训练样本上基于格网搜索优化 SVM 分类器进行训练,并在测试样本上进行精度比对。其中,XGBoost 特征选择采用默认的 gain 方法获取特征的重要性得分,保留非零得分的特征作为选择结果。如表 3 所示,StackingPSO-MI 算法优选特征子集支持下的 SVM 模型测试集精度 ACC 和 Kappa 值分别为 97.22%和 93.88%,略优于 LASSO、SVM-RFE 和 XGBoost,说明 StackingPSO-MI 算法能够得到较好的特征子集组合,并在参数优化的 SVM 分类器下取得较好的分类精度,其筛选结果有较好的可靠性。

表 3 基于 SVM 的不同特征选择方法的分类精度

Table 3 Classification accuracy of different feature selection methods based on SVM					
方法 Method	参数 Parameters	特征 Features	数量 Number	ACC/%	Kappa 值/% Kappa value
LASSO	Alpha=0.000 7, max_iter=10 000, C=1,gamma=0.001	Width,AR,CC,CE,Shape_a,VS,VO,CIE[1],CIE[5], CIE[7],CIE[8],CIE[9]	12	94.44	88.03
SVM-RFE	Step=1,CV=10, C=1,gamma=0.001	Width,AR,CC,VO, CIE[0]—CIE[7],CIE[9]	13	97.18	93.60
XGBoost	CV=10,C=10,gamma=0.001	AR,CC,CE,VO,CIE[4],CIE[5],CIE[6]	7	97.05	92.30
StackingPSO-MI	CV=10, C=1,gamma=0.001	Area,CIE[0],CIE[2],CIE[4],CIE[6],CIE[8]	6	97.22	93.88

4 讨论与结论

油菜籽粒较小,多呈球形或近似球形,籽粒颜色主要分为黑色、褐色、黄色 3 大类。不同油菜品种的形

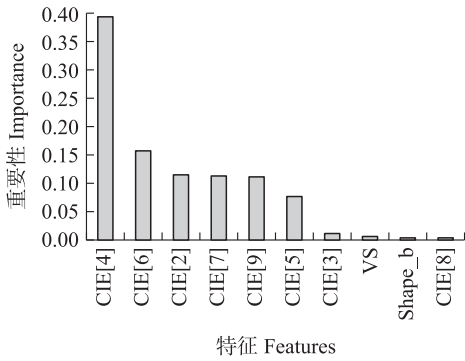


图 8 原始特征集下 RF 模型特征重要性排序 (前 10 位)

Fig. 8 Ranking of feature importance of the RF model under the original feature set (top10)

态与颜色特征可能存在差异,即便同一品种的油菜籽粒,其大小、形态和颜色等表观特征也存在明显的个体差异。薛汉军等^[18]在放大镜下辅助下,按照颜色的深浅程度将籽粒颜色分为8级;陈浩^[19]利用图像处理技术,基于 HSV 颜色空间以 S+V-H 作为评价油菜籽粒颜色深浅的指标将油菜籽粒进行了5级划分。上述研究主要集中在油菜籽粒的颜色分级,并未顾及其大小与形态特征。本文基于 QY-8 油菜籽粒的试验数据构建了无监督分类场景下的二分类油菜籽粒表观特征筛选方法,优选出了具有显著差异的5个 CIE 颜色分量和能够表征籽粒大小的面积特征,可用于优良油菜籽粒等级的快速划分,有效提高油菜籽粒分类效率。另外,研究已经表明特征选择算法能够在机器学习特征筛选方面表现出较好的应用潜力,但采用的算法、参数的不同,均可能导致特征选择结果存在较大差异^[20]。相比于 LASSO、SVM-RFE 和 XGBoost 特征选择算法,本文提出的油菜籽粒关键表观特征筛选方法,能够较精准地捕捉其关键特征,实现了所选关键特征数目少且分类精度高的双重目标。最后,PSO 特征选择中可采用多种高性能的机器学习模型(如 XGBoost、LightGBM)用于评价不同特征组合迭代分类错误率,进而影响最终的特征选择结果,本文未对此进行对比,有待后续进一步分析。

本研究基于油菜籽粒样本的大小、形状、颜色等表观特征,基于 LR 算法联合 KNN、RF 和 SVM 3 种基础学习器设计 Stacking 集成学习算法,提出了 SC-StackingPSO-MI 的油菜籽粒关键表观特征筛选方法,主要结论如下:1)在二分类条件下,基于本方法筛选的油菜籽粒表观特征包括形态特征中方差相对较大的面积 Area 和方差较大的颜色特征 CIE[0]、CIE[2]、CIE[4]、CIE[6]和 CIE[8]6 个表观特征分量,其谱聚类的 CH 聚类指数从 42.89 提高到 83.90,特征聚类性能得到显著改善,证明本文 SC-StackingPSO-MI 特征筛选方法的有效性;2)基于油菜籽粒二分聚类标签,优选特征组合与全特征集组合构建的 KNN、RF、SVM 和 LR 等模型测试精度相比,除 RF 模型外,其他 3 个模型的精度指标相对于原始特征集均有不同程度提高,说明利用本特征筛选方法获取的特征组合结果具有较好的稳健性;3)在 StackingPSO-MI 优选特征子集下,基于油菜籽粒二分谱聚类标签的 SVM 模型测试集精度 ACC 和 Kappa 值分别为 97.22%和 93.88%,优于 LASSO、SVM-RFE 和 XGBoost 算法,表明本文特征筛选方法具有较高的可靠性。

参考文献 References:

- [1] 张智猛,慈敦伟,丁红,等.花生种子大小和形状对出苗和幼苗建成的影响[J].花生学报,2014,43(1):16-23.
Zhang Z M,Ci D W,Ding H,et al. Effects of peanut seed size and shape on seedling emergence and establishment[J]. Journal of Peanut Science,2014,43(1):16-23(in Chinese with English abstract).
- [2] 杨蜀秦.农作物籽粒的图像处理和识别方法研究[D].杨凌:西北农林科技大学,2012.
Yang S Q. The research of image processing and identification method for crop's kernels[D]. Yangling: Northwest A&F University, 2012 (in Chinese with English abstract).
- [3] 刘伟,刘长虹,郑磊.基于支持向量机的多光谱成像稻谷品种鉴别[J].农业工程学报,2014,30(10):145-151.
Liu W,Liu C H,Zheng L. Discrimination in varieties of rice seeds with multispectral imaging using support vector machine[J]. Transactions of the Chinese Society of Agricultural Engineering,2014,30(10):145-151(in Chinese with English abstract).
- [4] 倪浩然,王润,胡昊,等.基于多光谱成像技术鉴别不同紫花苜蓿品种的种子[J].草地学报,2023,31(10):2985-2991.
Ni H R,Wang R,Hu H,et al. Seed identification of alfalfa varieties based on multispectral imaging[J]. Acta Agrestia Sinica,2023,31(10): 2985-2991(in Chinese with English abstract).
- [5] Fu X Z,Bai M J,Xu Y W,et al. Cultivars identification of oat(*Avena sativa* L.) seed via multispectral imaging analysis[J]. Frontiers in Plant Science,2023,14:1113535.
- [6] Zhang S H,Zeng H G,ji W,et al. Non-destructive testing of alfalfa seed vigor based on multispectral imaging technology[J]. Sensors,2022,22(7):2760.
- [7] 冯继克,郑颖,李平,等.基于特征选择的小麦籽粒品种识别研究[J].中国农机化学报,2022,43(7):116-123.
Feng J K,Zheng Y,Li P,et al. Research on the identification of wheat grain varieties based on feature selection[J]. Journal of Chinese Agricultural Mechanization,2022,43(7):116-123(in Chinese with English abstract).
- [8] 黄琼,杨红云,万颖.基于特征数据的水稻种子分类识别方法[J].江苏农业学报,2021,37(1):8-15.
Huang Q,Yang H Y,Wan Y. Classification and recognition method of rice seeds based on feature data[J]. Jiangsu Journal of Agricultural Sciences,2021,37(1):8-15(in Chinese with English abstract).
- [9] 孟惜,王克俭,韩宪忠.基于改进 BP 网络的小麦品种识别[J].贵州农业科学,2017,45(10):156-160.
Meng X,Wang K J,Han X Z. Classification of wheat varieties by improved BP neural network[J]. Guizhou Agricultural Sciences,2017, 45(10):156-160(in Chinese with English abstract).
- [10] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.

- [11] 李帅,常锦才,李吕牧之,等. 基于差分隐私保护的 Stacking 集成聚类算法研究[J]. 计算机工程与科学,2022,44(8):1402-1408.
Li S,Chang J C,Lili M Z,et al. A stacking ensemble clustering algorithm based on differential privacy protection[J]. Computer Engineering and Science,2022,44(8):1402-1408(in Chinese with English abstract).
- [12] Kennedy J,Eberhart R. Particle swarm optimization[C]//Proceedings of ICNN'95-International Conference on Neural Networks. November 27-December 1,1995,Perth,WA,Australia. IEEE,1995:1942-1948.
- [13] Yao H,Fu B L,Zhang Y,et al. Combination of hyperspectral and quad-polarization SAR images to classify marsh vegetation using stacking ensemble learning algorithm[J]. Remote Sensing,2022,14(21):5478.
- [14] 张婧,曹峰,董毓莹,等. 基于互信息和遗传算法的特征选择算法[J]. 山西大学学报(自然科学版),2024,47(1):1-8.
Zhang J,Cao F,Dong Y Y,et al. Feature selection algorithm based on mutual information and genetic algorithm[J]. Journal of Shanxi University (Natural Science Edition),2024,47(1):1-8(in Chinese with English abstract).
- [15] 王天一. 基于互信息和多目标粒子群优化的特征选择方法研究[D]. 镇江:江苏大学,2022.
Wang T Y. A study of feature selection method based on mutual information and multi-objective particle swarm optimization[D]. Zhenjiang: Jiangsu University,2022(in Chinese with English abstract).
- [16] 王学武,魏建斌,周昕,等. 一种基于超体积指标的多目标进化算法[J]. 华东理工大学学报(自然科学版),2020,46(6):780-791.
Wang X W,Wei J B,Zhou X,et al. Hypervolume-based multi-objective evolutionary algorithm[J]. Journal of East China University of Science and Technology,2020,46(6):780-791(in Chinese with English abstract).
- [17] 康岩松,臧顺来. 基于多种策略的改进粒子群优化算法[J]. 东北大学学报(自然科学版),2023,44(8):1089-1097.
Kang Y S,Zang S L. Improved particle swarm optimization algorithm based on multiple strategies[J]. Journal of Northeastern University (Natural Science),2023,44(8):1089-1097(in Chinese with English abstract).
- [18] 薛汉军,孙利军,梁升平,等. 甘蓝型油菜粒色与含油量的研究[J]. 西北农业学报,2010,19(1):81-85.
Xue H J,Sun L J,Liang S P,et al. Research of the color and oil content of the seed in *Brassica napus* L. [J]. Acta Agriculturae Boreali-Occidentalis Sinica,2010,19(1):81-85(in Chinese with English abstract).
- [19] 陈浩. 不同地域气候条件对油菜籽粒颜色及萌发活力的影响研究[D]. 武汉:华中农业大学,2022.
Chen H. The influence of different regional climatic conditions on grain color and germination vigor of rapeseed[D]. Wuhan:Huazhong Agricultural University,2022(in Chinese with English abstract).
- [20] 马元花,汪乐印,张祯鑫,等. 融合光谱和纹理特征的玉米产量预测研究[J]. 西北农业学报,2024,33(10):1827-1838.
Ma Y H,Wang L Y,Zhang Z X,et al. Remote sensing estimation for maize yield integrating spectral and texture features[J]. Acta Agriculturae Boreali-Occidentalis Sinica,2024,33(10):1827-1838(in Chinese with English abstract).

责任编辑:范雪梅