

油菜基因预测表型的优化方法研究

汪静¹, 刘昌华¹, 卢晓莲¹, 管文杰²

(1. 武汉轻工大学 数学与计算机学院; 2. 中国农业科学院油料作物研究所, 湖北 武汉 430048)

摘要: 植物育种是指改善植物所需表型性状的过程, 基因组预测可以帮助减少广泛表型评估的时间和成本。在油菜育种过程中, 通常利用单核苷酸多态性来预测油菜表型, 从而提高育种效率。然而, 传统的统计分析方法由于维度过大、模型单一等诸多因素的影响, 在某些情况下预测效果不尽如人意。为此, 提出一种基于优化支持向量回归与多层感知器的堆叠模型, 采用融合递归特征消除和随机森林特征选择方法, 通过对油菜基因数据和6个表型进行分析预测, 得出与基因型相关性最高的表型。实验结果表明, 含油量与基因数据的关联性最高, 皮尔逊相关系数达到0.83, 相较单独使用多层感知器的准确率提高了20%。优化的堆叠模型在基因预测表型中表现良好, 显著提升了预测准确性和鲁棒性。

关键词: 基因预测; 甘蓝型油菜; 堆叠模型; 特征选择

DOI: 10.11907/rjdk.251046

中图分类号: TP181

文献标识码: A

文章编号: 1672-7800(2026)002-0091-06

扫描二维码阅读全文:



Research on Optimization Methods for Predicting Phenotypes by Rapeseed Genes

WANG Jing¹, LIU Changhua¹, LU Xiaolian¹, GUAN Wenjie²

(1. School of Mathematics and Computer Science, Wuhan Polytechnic University;

2. Oil Crops Research Institute, Chinese of Agricultural Sciences, Wuhan 430048, China)

Abstract: Plant breeding refers to the process of improving the required phenotypic traits of plants, and genome prediction can help reduce the time and cost of extensive phenotypic evaluation. In the process of rapeseed breeding, single nucleotide polymorphisms are usually used to predict rapeseed phenotypes, thereby improving breeding efficiency. However, traditional statistical analysis methods may not achieve satisfactory prediction results in some cases due to factors such as large dimensions and single models. To this end, a stacked model based on optimized support vector regression and multilayer perceptron is proposed, which combines recursive feature elimination and random forest feature selection methods. By analyzing and predicting rapeseed gene data and six phenotypes, the phenotype with the highest correlation with genotype is obtained. The experimental results showed that the correlation between oil content and genetic data was the highest, with a Pearson correlation coefficient of 0.83, which increased the accuracy by 20% compared to using multi-layer perceptron alone. The optimized stacking model performs well in using gene to predict phenotype, significantly improving prediction accuracy and robustness.

Key Words: genomic prediction; Brassica napus; stacked model; feature selection

0 引言

油菜是世界上重要的油料作物之一, 在中国农业生产中占有重要地位^[1]。然而, 国产油菜的产油量和品质仍然面临诸多挑战^[2]。据统计, 中国油菜籽产油量为38%~45%, 低于世界平均水平的46%~50%^[3]。随着生活水平的提高, 人们对食用油的需求持续增加, 而中国食用油自给

率不足40%, 对外依存度较高^[4]。因此, 提高国内油菜的产油量有助于减少对进口食用油的依赖, 提升国内食用油的自给能力。

单株角果数(Silique Number per plant, SN)、每角果粒数(Seeds per Pod, SP)和千粒重(Thousand-Kernel Weight, TKW)是与油菜产量相关的重要农艺性状。此外, 油菜产油量还受到油菜株高(Plant Height, PH)、含油量(Seed Oil Content, SOC)和单株产量(Yield per Plant, YP)等因素的影

收稿日期: 2025-02-09

基金项目: 湖北省高等学校省级教学研究项目(2022343)

作者简介: 汪静(1999-), 女, 武汉轻工大学数学与计算机学院硕士研究生, 研究方向为智能识别与检测; 刘昌华(1963-), 男, 武汉轻工大学数学与计算机学院教授、硕士生导师, 研究方向为智能检测与食品安全、网络安全。本文通讯作者: 刘昌华。

响^[5]。因此,分析研究这些油菜表型具有重要意义。基因组预测(Genome Prediction, GP)能够显著减少育种周期中的表型评估时间和成本,加速关键性状的遗传改进。Meuwissen等^[6]最早提出利用GP来实施基因组辅助育种的方法,依赖单核苷酸多态性(Single Nucleotide Polymorphism, SNP)标记预测表型,从而提高育种效率。初期主要使用基于混合线性模型(Mixed Linear Model, MLM)与最小二乘法(Least Squares, LS)的方法,通过假设基因效应的线性组合来进行表型预测,但是难以捕捉基因型与表型之间的复杂非线性关系^[7]。因此,研究者们开始探索将机器学习和深度学习应用于GP中。例如,Liu等^[8]通过全基因组关联分析(Genome-Wide Association Study, GWAS)识别了与TKW相关的数量性状位点(Quantitative Trait Loci, QTL),为油菜育种提供了重要依据。Liu等^[9]研究了基因型与环境($G \times E$)互作对TKW的影响,发现环境因素对TKW表现有显著作用。Zhao等^[10]结合支持向量机(Support Vector Machine, SVM)与随机森林模型成功预测了油菜TKW。Xie等^[11]提出基于分子标记辅助选择的方法,提高了油菜的TKW。Yang等^[12]通过互信息法对小麦基因型数据进行特征选择,成功识别出与抗病性相关的关键基因位点,提高了基因型数据的预测精度。Xie等^[13]在使用递归特征消除(Recursive Feature Elimination, RFE)方法分析油菜基因型数据时,发现其能够有效筛选出与产量相关的关键特征,并显著提升了后续模型的表现。其他特征选择方法,如主成分分析和自动编码器亦被广泛应用于基因组数据的降维与特征提取。

尽管上述方法在GP中取得了一定成果,但单一模型的预测能力有限,难以充分挖掘基因型与表型之间的复杂关系。因此,本文提出一种基于优化支持向量回归(Support Vector Regression, SVR)与多层感知器(Multilayer Perceptron, MLP)堆叠模型的方法,结合融合后的RFE和随机森林特征选择方法,以提升表型预测的准确性和鲁棒性。通过对油菜基因型以及6个重要表型性状的分析,得到每个表型的预测准确性,并找出与基因型数据关联性最高的表型,为油菜的育种与改良提供科学依据。

1 材料与方法

1.1 实验材料

本文实验使用的基因型数据来源于中国农业科学院油料研究所提供的甘蓝型油菜60k Brassica Infinium® SNP芯片,由472份甘蓝型油菜种质的基因数据组成,每份油菜样本包含19 924个与表型相关的SNP位点。数据集信息如表1所示。

表型数据为6个产量性状,包括SN、SP、TKW、PH、SOC和YP。

Table 1 Dataset information

表1 数据集信息

ID	Bn-A01-p467816	Bn-A01-p622550	...	Bn-scaff_23186_1-p340682
1182	AC	TT	...	TT
1183	AA	TT	...	TT
1188	CC	CC	...	GG
...	...	...	...	...
7426	CC	CC	...	GG

1.2 研究方法

本文旨在利用优化SVR和MLP堆叠模型对甘蓝型油菜基因型数据及TKW、PH、SOC等表型数据进行分析。通过基因数据预测成熟期油菜籽粒的表型数据,探索基因与表型间的复杂关系,从而提高油菜育种效率和产量。主要步骤包括数据预处理、特征选择、模型构建、模型堆叠、绩效评估,每个步骤均独立应用于各表型。研究框架如图1所示。

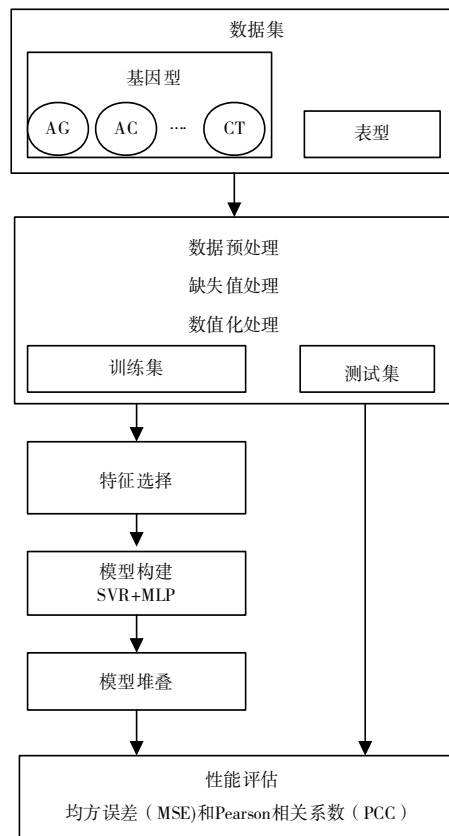


Fig. 1 Research framework

图1 研究框架

2 实验方案设计

2.1 数据预处理

(1)缺失值处理。一般样本数据都会存在缺失值,影响模型的训练和预测精度。很多机器学习算法无法处理缺失值,因此需要在预处理阶段处理掉缺失值。常用方法

包括使用均值、中位数、众数等填补缺失值,或直接删除包含缺失值的样本。本文研究中基因数据的缺失值比例较低,因此通过删除包含缺失值的样本来简化数据预处理流程,确保算法顺利执行并提高计算效率。

(2)数值化处理。基因数据由碱基对构成,而许多机器学习算法需要数值输入数据。因此,本文将每种基因型编码为1~13的整数值,即'AA': 1、'GG': 2、'CC': 3、'TT': 4、'NN': 5、'AC': 6、'AG': 7、'AT': 8、'TA': 9、'CG': 10、'GC': 11、'TC': 12、'TG': 13^[14]。数值编码可降低数据维度、减少数据稀疏性、提高计算效率,并增强数据的可视化 and 解读能力。

2.2 融合特征选择

RFE是一种基于包装策略的特征选择方法,通过反复训练和消除特征来选择最具预测能力的特征。其核心思想是利用一个学习模型来评估特征的重要性,并在每轮迭

代中移除影响最小的特征,直至最终剩下最重要的特征。RFE本身并不依赖于具体的学习模型,其可以与任何回归模型、分类模型或回归分析工具一起使用。

由于基因型数据具有高维性,建模前需要在训练数据集上进行降维,以提高预测准确性并减少过拟合。本文结合RFE与随机森林特征重要性评估进行油菜基因型数据的特征选择,具体流程如图2所示。首先,使用RFE对所有特征进行递归筛选,逐步消除对表型预测影响较小的特征,保留最具预测能力的基因数据;其次,在RFE筛选后的特征集上训练随机森林模型,计算每个特征的重要性分数,并对其进行排序;最后,结合RFE随机森林的评估结果,选择出最终的最重要特征集,以提高后续模型的预测精度和稳定性。以上过程确保了所选特征集具有较强的预测能力,优化了模型的训练效果^[15]。

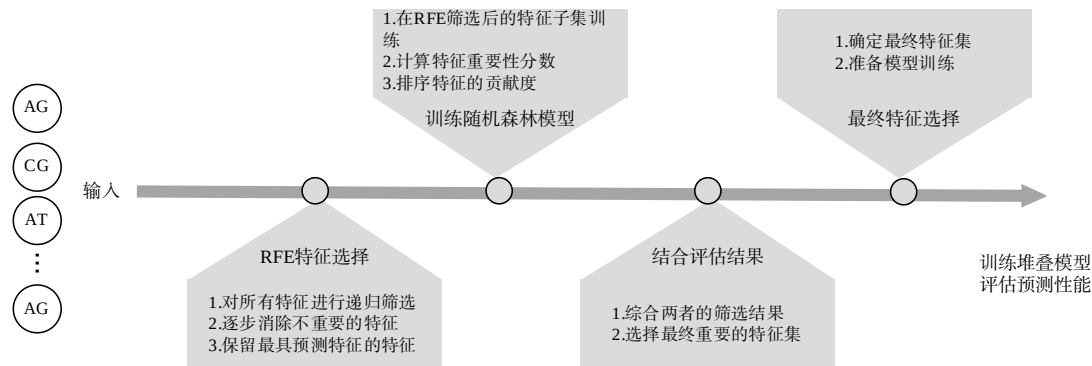


Fig. 2 Fusion feature selection flow
图2 融合特征选择流程

2.3 预测模型

2.3.1 SVR

SVR是一种建立在SVM基础上的函数回归算法,采用非线性单元隐藏层的前馈网络结构^[16]。在实验过程中可能会由于参数选择不当而导致原始SVR模型在处理高维数据时性能不佳。为此,本文使用网格搜索方法优化SVR模型的超参数,通过系统搜索正则化参数 C 、 ϵ 与核函数参数 γ 的最佳组合来提升模型的预测能力;同时采用径向基函数(Radial Basis Function, RBF)核作为SVR模型的核函数,用于捕捉数据之间复杂的非线性关系。最后,采用交叉验证方法评估SVR模型的性能,以确保其在不同数据划分下的稳定性和预测性能。优化SVR流程见图3。

2.3.2 MLP

MLP是一种前馈神经网络,包含输入层、隐藏层和输出层^[17]。然而,MLP模型在处理高维数据时容易出现过拟合问题,并且在优化过程中容易陷入局部最优解。为了克服这些挑战,本文提出一种优化的MLP网络结构,对传统MLP模型进行以下改进:①在每个隐藏层的输出后引入Dropout层,通过随机丢弃部分神经元输出的方式防止过拟合,提升模型的泛化能力;②引入自注意力机制(Self-Attention Mechanism)来提升模型对关键特征的关注意度;

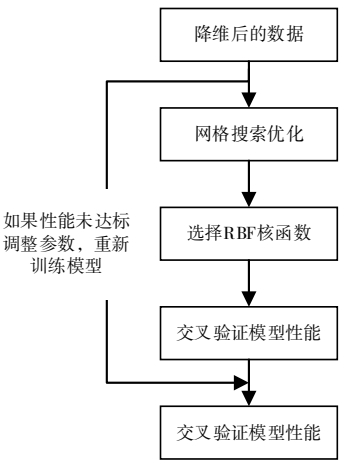


Fig. 3 Optimized SVR flow
图3 优化SVR流程

③优化训练过程,采用不同的优化算法加速模型训练并提高收敛速度。优化的MLP结构见图4。

2.3.3 模型堆叠

本文提出一种堆叠模型(Stacking)架构,通过结合多个学习模型来增强对油菜表型的预测能力。具体结构见图5。堆叠模型的核心思想为通过一级模型(SVR)与二级模型(MLP)之间的协作,利用不同模型的优势来弥补单一

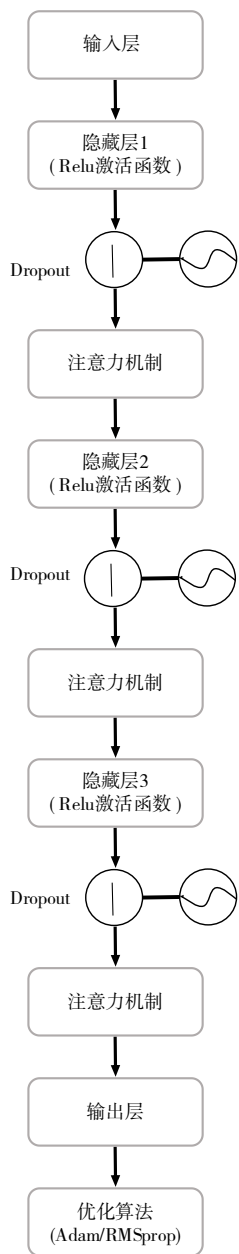


Fig. 4 Optimized MLP structure

图4 优化的MLP结构

模型的缺点。构建步骤为：

(1)构建一级模型。使用优化后的SVR模型作为一级模型,通过交叉验证评估SVR的预测效果,并保存其在训练集上的预测结果作为新的特征,为二级模型提供由一级模型产生的预测信息,而不仅仅是原始的输入数据特征。

(2)构建二级模型。将SVR模型的预测结果作为特征输入到优化后的MLP模型中,训练MLP模型,使其学习如何校正SVR模型的预测误差,进一步提升预测精度。

(3)训练堆叠模型。在堆叠过程中,MLP模型基于SVR模型的预测结果进行二次训练,校正单一模型的误差,优化预测准确性。采用交叉验证等方法评估堆叠模型的性能,并与单独的SVR和MLP模型进行比较。

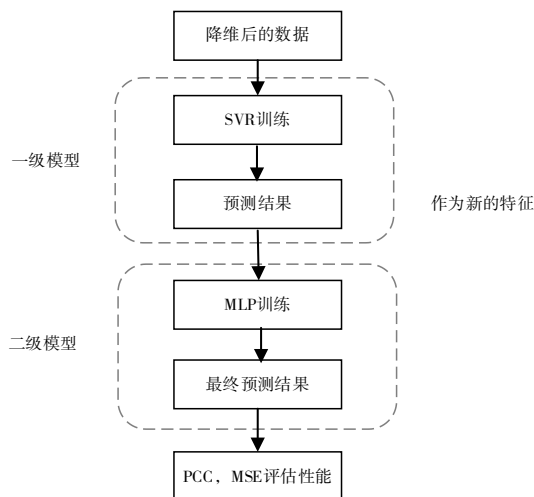


Fig. 5 Stacking model structure

图5 堆叠模型结构

2.3.4 模型性能评价

使用实测表型与预测表型之间的均方误差(Mean Squared Error, MSE)来评估模型性能。计算公式为：

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

式中： y_i 和 $\hat{y}_i$ 分别表示检验中的预测值和实测值。MSE值越小,表明模型在预测表型时性能和泛化能力越好。因此,通过最小化MSE可以优化模型,使其更好地拟合数据,提高预测的可靠性和准确性。

使用皮尔逊相关系数(Pearson Correlation Coefficient, PCC)衡量性状数值之间的相关性,以反映和比较不同降维与预测模型的性能。PCC是一个标准化统计量,是GP问题最常用的指标^[18]。计算公式为：

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

式中： x_i, y_i 为第*i*个样本两个变量的取值； $\bar{x}$ 和 $\bar{y}$ 表示两个变量的平均值。 R 的取值范围为-1~1,越接近1表示两个变量的相关程度越高,0则表示没有相关性。

3 实验结果与分析

对油菜基因数据进行缺失值和数值化处理后,分别采用单独的SVR、MLP和堆叠模型对6个表型进行预测实验。各表型的PCC和MSE值分别见图6、图7。可以看出,堆叠模型对6个表型的PCC和MSE值均优于单一模型(SVR、MLP),具有更强的泛化能力和稳定性。其中,SOC基因表型与基因型数据的关联性最高,PCC值达到0.83,MSE值为0.07,可能与SOC的遗传基础和基因型特征的稳定性相关。总体而言,堆叠模型在6个油菜表型预测中均表现良好。通过引入融合特征选择方法和多种优化策略,模型能够更好地处理高维基因型数据,捕捉基因型数据与表型之间的复杂关系,显著提升预测性能和泛化能力。

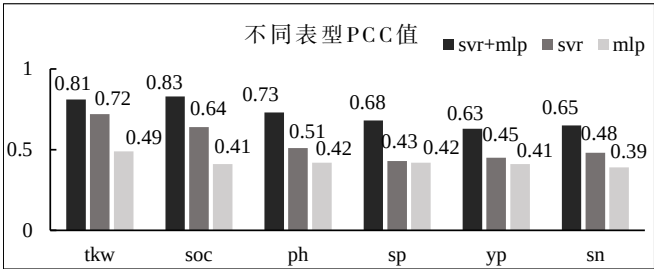


Fig. 6 PCC values for different phenotype predictions

图6 各表型预测PCC值

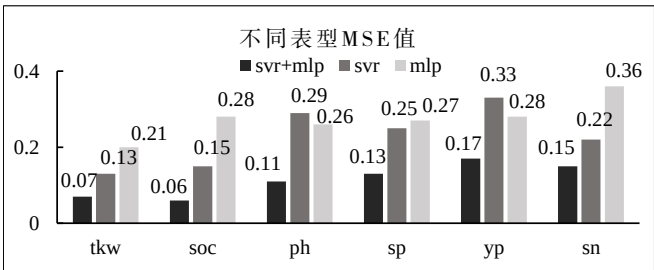


Fig. 7 MSE values for different phenotype predictions

图7 各表型预测MSE值

以SOC表型为例,堆叠模型的MSE和PCC值随epoch的变化趋势如图8所示。可以看出,MSE在初期阶段迅速下降,表明模型通过特征融合策略快速捕捉到关键基因型特征,显著降低了预测误差,并进入了一个显著优化的阶段。随后,MSE的下降逐渐趋缓,在一定的训练轮次后达到一个平稳状态,说明模型已经达到了收敛,进一步的训练不再显著改善误差水平。PCC则随着迭代次数的增加而快速上升,从初期的较低水平迅速攀升至约0.83,随后趋于稳定。这种表现表明堆叠模型在训练过程中逐步提高了预测值与真实值之间的线性相关性,在捕捉关键特征与表型之间关系方面表现较好。

使用RFE进行特征选择,利用随机森林回归评估所选特征的重要性,得到对预测SOC表型前10个特征的重要

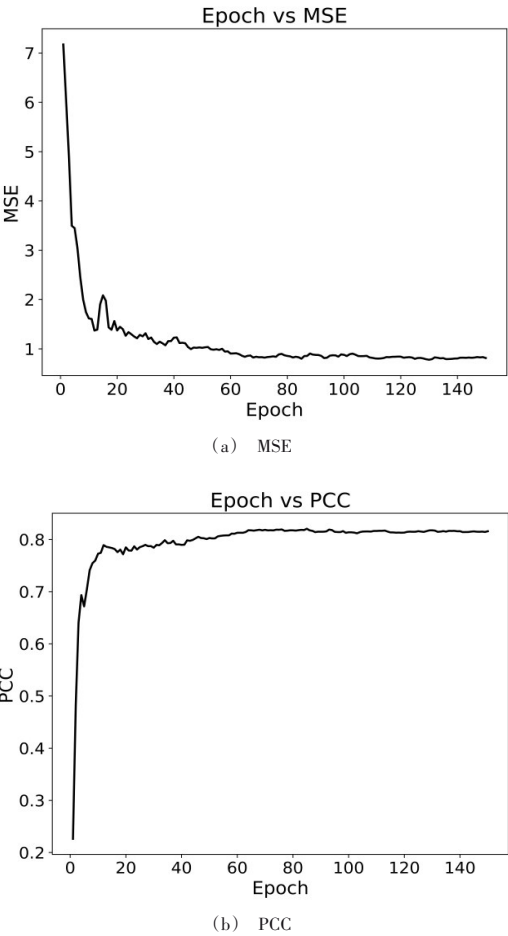


Fig. 8 Plot of the model's MSE and PCC values changing with epoch

图8 模型MSE和PCC值随epoch的变化趋势

性结果,以更全面地理解哪些基因对于SOC表型预测具有重要影响。由图9可以看出,Bn-A03-p248970、Bn-scaff_15576_1-p140859等SNP在模型中对SOC的影响最为显著。以上结果为理解油含量的遗传基础提供了新视角,有

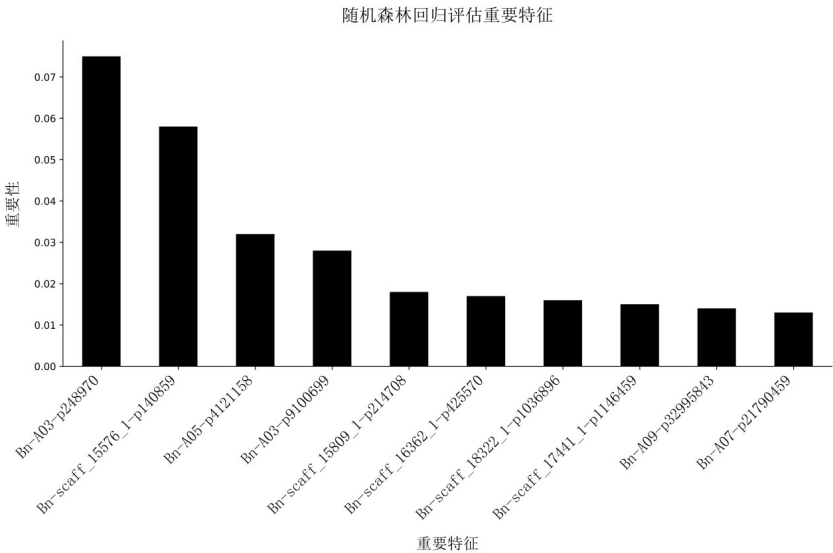


Fig. 9 Random forest regression evaluation of important features

图9 随机森林回归评估重要特征

助于揭示油含量表型的分子机制,为分子育种提供指导。

4 结语

SVR擅长处理高维数据并捕捉其中的非线性关系,而MLP可以通过深度网络结构进一步增强对复杂非线性关系的建模能力。本文充分利用SVR和MLP各自的优势,使其相互补充,形成一个新的预测堆叠框架,两者结合不仅可以克服单一模型在处理高维基因型数据和复杂表型关系时的局限性,而且能够更全面地捕捉基因型与表型之间的复杂关系,提高预测准确性。在油菜基因预测表型的研究中,堆叠模型的表现优于SVR和MLP这两个单一模型,展现出较好的预测精度和泛化能力。研究结果为GP在油菜育种中的应用提供了参考,展现出堆叠模型在复杂性状预测中的潜在应用价值。未来研究可引入更多基因、表型数据及环境变量,进一步优化模型结构,提高预测精度。

参考文献:

- [1] HE Z H, LI B, LIU Y. Current status and future prospects of China's rapeseed industry[J]. *Agricultural Economics*, 2021, 47(1): 23-30.
何中华, 李波, 刘洋. 中国油菜产业发展现状及前景[J]. *农业经济*, 2021, 47(1): 23-30.
- [2] National Bureau of Statistics. China Agricultural Yearbook—statistical data on rapeseed production in China[M]. Beijing: China Agriculture Press, 2020: 125-130.
国家统计局. 中国农业年鉴——中国油菜生产统计数据[M]. 北京: 中国农业出版社, 2020: 125-130.
- [3] XIE K, CHENG X, LIU N. Current status and outlook of rapeseed production in China[J]. *Journal of Oil Crops Science*, 2020, 42(5): 673-680.
谢凯, 程旭, 刘娜. 中国油菜生产现状及展望[J]. *油料作物学报*, 2020, 42(5): 673-680.
- [4] LIU H, WANG Y, ZHANG J. Analysis and outlook of China's edible oil market[J]. *Agricultural Economic Research*, 2021, 49(2): 14-22.
刘辉, 王燕, 张杰. 中国食用油市场分析与展望[J]. *农业经济研究*, 2021, 49(2): 14-22.
- [5] ZHANG C P, GONG R L, ZHONG H, et al. Integrated multi-locus genome-wide association studies and transcriptome analysis for seed yield and yield-related traits in *Brassica napus*[J]. *Frontiers in Plant Science*, 2023, 14: 1153000.
- [6] MEUWISSEN T H E, HAYES B J, GODDARD M E. Prediction of total genetic value using genome-wide dense marker maps[J]. *Genetics*, 2001, 157(4): 1819-1829.
- [7] YANG J, BENYAMIN B, MCEVOY B P, et al. Common SNPs explain a large proportion of the heritability for human height[J]. *Nature Genetics*, 2010, 42(7): 565-569.
- [8] LIU J, ZHANG D, SUN R, et al. Identification of quantitative trait loci for thousand seed weight in *Brassica napus*[J]. *Theoretical and Applied Genetics*, 2016, 129(4): 657-669.
- [9] LIU Y, XU H, ZHANG X. Genotype-by-environment interactions for thousand seed weight in *Brassica napus*[J]. *BMC Genomics*, 2019, 20(1): 1-10.
- [10] ZHAO J, WU X, WANG Y, et al. Prediction of thousand seed weight in rapeseed using machine learning models[J]. *Frontiers in Plant Science*, 2021, 12: 654347.
- [11] XIE F, YANG L, LI H, et al. Molecular marker-assisted selection for improvement of thousand seed weight in rapeseed[J]. *Euphytica*, 2020, 216(8): 136.
- [12] YANG L, LIU Z, CHEN Y, et al. Feature selection in wheat genotype data using mutual information method[J]. *Journal of Genomics and Crop Science*, 2019, 25(3): 145-154.
- [13] XIE W, ZHANG Q, CHEN T, et al. Recursive feature elimination for yield-related trait prediction in rapeseed genotypes[J]. *BMC Plant Biology*, 2022, 22(1): 78-86.
- [14] HE X Y, LIU C H, GUAN W J. Prediction of rapeseed thousand-seed weight based on feature selection[J]. *Journal of Wuhan Polytechnic University*, 2022, 41(5): 34-39.
何旭燕, 刘昌华, 管文杰. 基于特征选择实现油菜基因预测千粒重值[J]. *武汉轻工大学学报*, 2022, 41(5): 34-39.
- [15] ZHAO Q Y, YE C M, LU Y B. Importance evaluation of micro-motion features based on random forest[J]. *Modern Defense Technology*, 2022, 50(4): 124-131.
赵庆媛, 叶春茂, 鲁耀兵. 基于随机森林的微动特征重要性评估研究[J]. *现代防御技术*, 2022, 50(4): 124-131.
- [16] GAN L Q, LIU Y H. Air quality index prediction based on BP-SVR combination model[J]. *Software Guide*, 2020, 19(10): 80-83.
甘露情, 刘媛华. 基于BP-SVR组合模型的空气质量指数预测[J]. *软件导刊*, 2020, 19(10): 80-83.
- [17] LI Q C, CHEN J, ZHAO Q L, et al. Prediction of wheat grain zinc content and safety utilization zoning based on deep learning[J/OL]. *Journal of Agro-Environment Science*, 2024-06-25. <http://kns.cnki.net/kcms/detail/12.1347.s.20240415.1643.007.html>.
李清彩, 陈娟, 赵庆令, 等. 基于深度学习的小麦籽粒锌含量预测及安全利用分区[J/OL]. *农业环境科学学报*, 2024-06-25. <http://kns.cnki.net/kcms/detail/12.1347.s.20240415.1643.007.html>.
- [18] ZHANG X Z, WANG Y F, SHI W S. pCAMP: performance comparison of machine learning packages on the edges[DB/OL]. <https://arxiv.org/abs/1906.01878>.

(责任编辑:尹晨茹)