Advances in Applied Mathematics 应用数学进展, 2024, 13(4), 1643-1647
Published Online April 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.134156

Hans汉斯

基于机器学习试析孟德尔随机化研究中R^2值预测模型

伊力米努尔·艾克拜尔

新疆大学外国语学院，新疆 乌鲁木齐

收稿日期：2024年3月25日；录用日期：2024年4月22日；发布日期：2024年4月29日

摘　要

孟德尔随机化研究在现代遗传学领域具有重要意义。它利用自然发生的基因突变作为工具，探究基因变异与生物特性之间的因果关系，从而克服了传统观察性研究中可能存在的混杂因素干扰，为生物特性机制的揭示提供了有力支持。然而与表型相关的研究数据中R^2值很难获取，国内外公共数据库中也常缺失。因此本文以我国生物信息中心(CNCB)数据库中甘蓝型油菜(oilseed rape)开花时间相关的基因数据为学习素材，通过采取多种机器学习算法，试对比不同模型预测R^2值的适用性。

关键词

遗传学，机器学习，孟德尔随机化

Analysis of the Prediction Model of R^2 Value in Mendelian Randomization Study Based on Machine Learning

Yiliminuer Aikebaier

School of Foreign Languages, Xinjiang University, Urumqi Xinjiang

Received: Mar. 25th, 2024; accepted: Apr. 22nd, 2024; published: Apr. 29th, 2024

Abstract

The study of Mendelian randomization is of great significance in modern genetics. It uses naturally occurring gene mutation as a tool to explore the causal relationship between gene variation and traits, thus overcoming the possible confounding factors in traditional observational studies and

文章引用: 伊力米努尔·艾克拜尔. 基于机器学习试析孟德尔随机化研究中 R^2 值预测模型[J]. 应用数学进展, 2024, 13(4): 1643-1647. DOI: 10.12677/aam.2024.134156

providing strong support for the revelation of disease pathogenesis. However, R^2 value is difficult to obtain in research data related to phenotype, and is often missing in public databases at home and abroad. In this paper, the genetic data related to the flowering time of oilseed rape in the CNCB database was used as learning materials, and various machine learning algorithms were adopted to compare the applicability of different models to predict R^2 values.

Keywords

Genetics, Machine Learning, Mendelian Randomization

1. 引言

孟德尔随机化(Mendelian Randomization, MR)是一种基于遗传变异的方法，用于推断因果关系。其基于孟德尔遗传定律，利用自然发生的基因变异来探究基因与复杂表型之间的因果关系。核心思想在于，个体的基因变异是随机且自然发生的，这些变异可以视为自然实验，从而避免了传统观察性研究中可能存在的混杂因素和反向因果关系问题[1]。

该方法利用与暴露密切相关的遗传变异(通常以单核苷酸多态性(Single Nucleotide Polymorphism, SNP)的形式)作为工具变量(Instrumental Variable, IV)，通过遗传变异的特性来评估暴露因素与结局之间的因果关系。使用“暴露”一词来指代假定的因果风险因素，有时也称为中间表型，它可以是生物标志物(Biomarker)、人体测量指标(Physical Measurement)或任何其他可能影响结果的风险因素(Risk Factor)。通常情况下，结局是疾病，但并不局限于疾病[2]。

在生物信息学研究中，表型相关的研究数据往往扮演着至关重要的角色，其中 R^2 表示工具变量解释暴露因素的程度。然而，一个普遍存在的问题是，与表型相关的 R^2 值并不容易获取，无论是在国内还是国外的公共数据库中，R^2 值常常缺失。因此本文以我国生物信息中心(CNCB)数据库中甘蓝型油菜(oilseed rape)开花时间相关的基因数据为学习素材，利用多种经典的机器学习模型，包括线性回归、决策树、随机森林和梯度提升回归树来预测 R^2 值，以此来评估各种算法在预测 R^2 值方面的适用性。试图为生物信息学领域的研究者提供一些有益的参考和借鉴。

2. 材料与方法

2.1. 数据来源

数据选取自我国生物信息中心(CNCB)数据库中甘蓝型油菜(oilseed rape)开花时间相关的基因数据1000 条，其中包含突变基因 ID，*P*-values 和 R^2 值。

首先，检查数据集中的重复项、无效值或异常值。对于异常值，先使用 matplotlib 库中散点图对数据进行可视化，大致了解数据中异常值的情况。再使用 Python 中 Scipy 库的 Z-score 方法(通常，Z-score 绝对值大于 3 的值被认为是异常值)，进行筛选和清理。

其次，缺失值的存在可能影响数据分析的准确性。使用 Python 中 Pandas 库函数 dropna()，直接删除含有缺失值的行或列。

最后，检查数据的分布特性。绘制直方图，观察数据的分布情况，数据总量达到千条，涵盖多个维

度。在数值型数据中，分布呈现偏态特征。

2.2. 方法

通过上文原始数据分布可视化，发现原始数据呈现出稍带曲线的线性关系，因此选取线性回归模型，同时选取决策树回归来探索其可能有的非线性关系。

因此，本文选取的机器学习模型分别是线性回归、决策树回归、梯度提升回归树和随机森林回归。线性回归通过拟合最佳直线来预测目标变量，简单直观，适用于线性关系的数据；决策树回归基于树形结构进行回归预测，易于理解，能处理非线性关系；梯度提升回归树通过迭代添加弱学习器(决策树)来优化预测，提升性能；随机森林回归构建多棵决策树并集成其预测结果，提高预测精度和稳定性[3]。以上四种模型的优缺点见表 1。

Table 1. Advantages and disadvantages of four machine learning models
表 1. 四种机器学习模型优缺点

学习模型	优点	缺点
线性回归 (Linear Regression)	□易于理解和实现 □计算简单且速度快 □能够捕捉变量之间的线性关系	□对非线性关系建模能力较差。 □对异常值敏感 □假设输入特征之间不相关(即不存在多重共线性)
决策树回归 (Decision Tree Regression)	□能够捕捉非线性关系 □对特征的选择和转换不敏感 □容易理解和可视化	□可能会过拟合 □不稳定，不同的样本集可能导致不同的树结构 □对连续特征的处理可能不够精细。
梯度提升回归树 (Gradient Boosting Regression Trees)	□能够处理复杂的非线性关系 □通常比单个决策树具有更好的性能 □对异常值和噪声的鲁棒性较好	□计算成本较高 □可能对超参数敏感 □难以解释单个预测的来源
随机森林回归 (Random Forest Regression)	□具有较高的预测性能 能够处理高维数据 □对特征选择和转换不敏感 □能够评估特征的重要性	□计算成本较高 □对于某些数据集可能会过拟合 □模型较难解释

生成四种模型后，计算模型的平均绝对误差(Mean Absolute Error)、均方误差(Mean Squared Error)和均方根误差(Root Mean Squared Error)，用来评估模型的优良性。再用 GridSearch 进行模型优化，通过穷举所有可能的参数组合来寻找最佳超参数组合，并使用交叉验证方法评估每个组合的性能，最终选择性能最优的超参数组合作为模型的最终超参数，以实现更好的模型性能。

3. 结果

在评估模型的过程中，有重要的三个参数分别是平均绝对误差(MAE)衡量预测值与真实值之间的平均绝对差异，直观反映预测误差大小；均方误差(MSE)计算预测误差的平方均值，对较大误差更敏感和均方根误差(RMSE)是 MSE 的平方根，与数据规模相关，常用于比较不同数据集上的模型性能。通过训练数据集生成的四个模型参数罗列如表 2。

通过表格，可明显看到线性回归模型的平均绝对误差是最小的，其相应的均方根误差也是最小；随机森林回归模型的平均绝对误差是最大的，其相应的均方根误差也是最大。可见在四种模型中，线性回归模型对该类数据集的拟合度最高，随机森林回归模型对该类数据集的拟合度最低。

Table 2. Four model evaluation parameters
表 2. 四种模型评估参数

模型评估参数	线性回归	决策树回归	梯度提升回归树	随机森林回归
Mean Absolute Error	3.848	4.958	5.051	5.058
Mean Squared Error	17.524	133.303	144.932	148.843
Root Mean Squared Error	4.186	11.546	12.039	12.201

通过 GridSearch 进行模型优化，全面搜索参数空间，选定了最佳参数组合以构建新模型。四种模型的性能评估参数现已整理至表 3，这些参数直观地反映了模型在测试集上的预测效果。由表 3 可知，经过优化后的四个模型在评估参数上均呈现出显著的下降趋势，这充分说明了通过 GridSearch 进行模型优化和参数选择的有效性。

Table 3. Four model evaluation parameters after model optimization
表 3. 模型优化后四种模型评估参数

模型评估参数	线性回归	决策树回归	梯度提升回归树	随机森林回归
Mean Absolute Error	2.984	4.537	4.961	4.973
Mean Squared Error	10.537	111.636	139.813	141.826
Root Mean Squared Error	3.246	10.566	11.824	11.909

线性回归模型经过优化后，其平均绝对误差降低至 2.984，这意味着模型的预测值与真实值之间的平均绝对差异显著减少。同时，其均方根误差也降低至 3.246，进一步反映了模型在预测上的准确性得到了提升。这样的优化结果对于线性回归模型来说是非常显著的，显示出模型在数据拟合上的能力得到了有效增强。随机森林回归模型经过参数优化后，该模型的平均绝对误差降至 4.973，虽然相较于线性回归模型仍然较高，但相较于优化前的结果已经有了明显的改进。同时，其均方根误差也回落到 11.909，虽然数值上仍然较大，但考虑到随机森林模型本身的复杂性和对数据的适应能力，这样的降幅也是值得肯定的。

在对比四个模型的评估参数降幅时，可以发现线性回归模型的评估参数降幅最大，这可能是由于线性回归模型对于参数的敏感性较高，通过优化参数可以显著提升其性能。而随机森林回归模型的评估参数降幅相对较小，这可能与随机森林模型本身的稳定性和对参数的鲁棒性有关。

综上所述，通过 GridSearch 进行模型优化和参数选择，成功降低了四个模型的评估参数，提升了模型的预测性能。这一结果不仅验证了优化方法的有效性，也为后续模型的应用和推广提供了坚实的基础。

4. 总结

本研究旨在利用机器学习技术，深入探索孟德尔随机化研究中的 R^2值预测模型。通过构建和训练一系列机器学习模型，试图揭示遗传变异与表型变异之间复杂关系的潜在规律，为精准医疗和遗传学研究提供有力工具。

在研究过程中，发现机器学习模型在预测 R^2值方面展现出一定的优势，能够较准确地捕捉遗传变异对表型变异的贡献度。然而，研究也存在一定的局限性。首先，数据集的规模和质量对模型性能具有显著影响。目前，可用的孟德尔随机化研究数据相对有限，且可能存在噪声和偏差，这在一定程度上限制了模型的泛化能力。其次，机器学习模型的复杂性和可解释性之间存在权衡。为了获得更高的预测精度，

可能需要采用更为复杂的模型结构，但这往往牺牲了模型的可解释性。

日后在模型的优化方面，首先，应积极收集更多的孟德尔随机化研究数据，并对数据进行预处理和质量控制，以提高数据的质量和可靠性。其次，探索更先进的机器学习算法和模型结构，以平衡模型的复杂性和可解释性，同时提升预测精度。

参考文献

[1] Burgess, S., Daniel, R.M., Butterworth, A.S., *et al*. (2014) Network Mendelian Randomization: Using Genetic Variants as Instrumental Variables to Investigate Mediation in Causal Pathways. *International Journal of Epidemiology*, **44**, 484-495. https://doi.org/10.1093/ije/dyu176

[2] Stephen, B. (2021) Mendelian Randomization: Methods for Causal Inference Using Genetic Variants. Taylor & Francis Group, Oxford.

[3] Lipton, Z.C. (2018) The Mythos of Model Interpretability. *ACM Queue*: *Architecting Tomorrows Computing*, **16**, 31-57. https://doi.org/10.1145/3236386.3241340