工程科学学报
Chinese Journal of Engineering

基于机器学习的材料弹性性能预测及可视化分析

林轩杰 江汉同 李倩 周玉玲 臧怀娟 任永生 詹曙 马文会

Prediction of the elastic properties of materials based on machine learning and visualization analysis

LIN Xuanjie, JIANG Hantong, LI Qian, ZHOU Yuling, ZANG Huaijuan, REN Yongsheng, ZHAN Shu, MA Wenhui

引用本文:
林轩杰, 江汉同, 李倩, 周玉玲, 臧怀娟, 任永生, 詹曙, 马文会. 基于机器学习的材料弹性性能预测及可视化分析[J]. 工程科学学报, 2024, 46(6): 1120–1129. doi: 10.13374/j.issn2095–9389.2023.08.10.003
LIN Xuanjie, JIANG Hantong, LI Qian, ZHOU Yuling, ZANG Huaijuan, REN Yongsheng, ZHAN Shu, MA Wenhui. Prediction of the elastic properties of materials based on machine learning and visualization analysis[J]. *Chinese Journal of Engineering*, 2024, 46(6): 1120–1129. doi: 10.13374/j.issn2095–9389.2023.08.10.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.08.10.003

工程科学学报, 第 46 卷, 第 6 期: 1120−1129, 2024 年 6 月
Chinese Journal of Engineering, Vol. 46, No. 6: 1120−1129, June 2024
https://doi.org/10.13374/j.issn2095-9389.2023.08.10.003; http://cje.ustb.edu.cn

基于机器学习的材料弹性性能预测及可视化分析

林轩杰[1,2)], 江汉同[1,2)], 李　倩[1,2)], 周玉玲[1,2)], 臧怀娟[1,2)], 任永生[3,4)]✉, 詹　曙[1,2)], 马文会[3,4)]

1) 合肥综合性国家科学中心人工智能研究院, 合肥 230601　2) 合肥工业大学计算机与信息学院, 合肥 230601　3) 真空冶金国家工程实验室, 昆明 650093　4) 昆明理工大学冶金与能源工程学院, 昆明 650093

✉通信作者, E-mail: ryssdy@126.com

摘　要　在工程材料的应用中, 弹性模量是重要的性能参数, 找到特定弹性性能的材料是新材料合成领域的热点问题, 如何快速且准确的预测弹性在工程上具有重要意义. 通过实际实验测量大量材料的弹性性能并不现实. 因此, 通过计算机模拟筛选材料数据, 选出候选材料, 再通过实际实验进行验证, 是一种理想的新材料发现方法. 目前材料性能预测的主要计算方法是基于第一性原理的高通量计算, 这类方法效率低下, 难以高效地完成大批量的材料筛选任务. 而基于材料统计学的机器学习预测方法, 可通过大数据挖掘, 快速预测材料性能, 成为一种有可能替代高通量计算的方案. 本文将特征选择方法和机器学习模型进行组合, 从中选择最有效的弹性模量预测组合方案, 并设计交互界面对输入特征和材料弹性性能的关系进行可视化分析. 实验表明 Pearson/RFE 和 GBDT 的组合模型性能最好, 同时通过可视化分析发现每原子能量、熔点、密度等特征对于预测结果的影响较大. 这些重要的特征可以从特征–目标关系中初步预测弹性模量的范围, 目标属性值也可反过来估计材料的重要特征. 这些研究成果可应用于探究弹性的影响因素、预测大批量材料性能和可视化分析指导材料合成.

关键词　特征选择; 机器学习; 弹性性能预测; 可视化; 数据挖掘

分类号　TP181;TQ127.2

Prediction of the elastic properties of materials based on machine learning and visualization analysis

LIN Xuanjie[1,2)], *JIANG Hantong*[1,2)], *LI Qian*[1,2)], *ZHOU Yuling*[1,2)], *ZANG Huaijuan*[1,2)], *REN Yongsheng*[3,4)]✉, *ZHAN Shu*[1,2)], *MA Wenhui*[3,4)]

1) Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230601, China

2) School of Computer and Information, Hefei University of Technology, Hefei 230601, China

3) National Engineering Laboratory of Vacuum Metallurgy, Kunming 650093, China

4) Faculty of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming 650093, China

✉Corresponding author, E-mail: ryssdy@126.com

ABSTRACT　The elastic modulus is an important performance parameter that measures the ability of materials to resist deformation and is critical for assessing their reliability and stability. Thus, the elastic modulus serves as a guide in engineering design and material selection, and finding materials with specific elastic properties is a hot issue in the field of novel materials synthesis. Predicting elasticity quickly and accurately is of great significance in engineering. It is not practical to measure the elastic properties of many materials using practical experiments because this requires a significant amount of time and cost. For many material samples, each sample needs to be tested and analyzed, which is a time-consuming and expensive task. Thus, screening material data through computer simulation,

收稿日期: 2023−08−10

基金项目: 国家自然科学基金青年科学基金资助项目(52104303); 安徽省高校协同创新项目(GXXT-2022-041); 安徽省科技重大专项(2023z020006)

choosing candidate materials, and then confirming them through actual experiments is an ideal method for new material discovery. Currently, the main calculation methods for material performance prediction are first-principles high-throughput calculation, which is inefficient and difficult to efficiently complete the high-volume material screening. Machine learning prediction methods based on material statistics can rapidly predict material properties through big data mining, which has become a possible alternative to high-throughput computing. In this work, the feature selection method and machine learning model are integrated to choose the most effective combination scheme for elastic modulus prediction, and an interactive interface is developed to perform a visual analysis of the relationship between input features and elastic properties of materials. For the analysis of the prediction results, the root mean square error (R_{mse}) and *R*-Square (R^2) are employed as evaluation indicators for the performance of the prediction model. The experiment shows that the Pearson/RFE/LASSO-GBDT combination model possesses the best performance. On the other hand, by visualization analysis, it is revealed that the energy of each atom, melting point, density, and other characteristics have a great effect on the prediction results. These important characteristics can be utilized to preliminarily predict the range of elastic moduli from the feature–target relationship, and the value of target attributes can be used for the estimation of important characteristics of materials. These findings can be applied to investigate the influencing factors of elasticity, predict the properties of large quantities of materials, and guide the synthesis of materials by visualization analysis. This work has certain significance for guiding the discovery of novel materials and exploring the influencing factors of material properties.

KEY WORDS feature selection; machine learning; elastic property prediction; visualization; data mining

随着科技的快速发展, 航空航天、人工智能、人造太阳等这些曾经不为人知的新兴词汇如今已从概念逐渐变为现实, 影响着人类的生活与生产方式. 新兴领域取得的突破伴随着新产物的出现, 都和新材料的研究与发展有着密不可分的联系[1]. 新材料是新兴产业发展的基础, 几乎所有的重大科技革新都与新材料的出现有关[2]. 2011 年美国提出“材料基因组计划 (MGI)[3], 材料基因组融合了材料的高通量计算、高通量制备、高通量检测及数据库系统[4], 形成新材料创新发展的基础条件和能力. 材料基因工程是材料领域的颠覆性前沿技术, 将对材料研发模式产生革命性的变革, 全面加速材料从设计到工程化应用的进程, 大幅度提升新材料的研发效率[5]. 材料基因工程是材料领域的颠覆性前沿技术, 将对材料研发模式产生革命性的变革, 全面加速材料从设计到工程化应用的进程, 大幅度提升新材料的研发效率, 缩短研发周期, 降低研发成本, 促进工程化应用. 高通量计算是以量子力学中的薛定谔方程作为理论基础, 以材料内部粒子空间分布 (晶体学信息文件 CIF) 作为输入, 设置所需参数, 预测材料的相关属性 (包括动力学、热力学和结构性质等)并导入材料数据库. 然而高通量计算需要消耗大量的计算资源与时间[6], 不同元素构成的材料, 如果元素种类、配比、空间结构等存在差异, 生成材料的性能也大不相同, 仅仅采用高通量计算进行盲目的检索, 并不是一种明智的举措. 近些年来, 随着基于第一性原理 (DFT)[7−8] 模拟仿真方法[9−10] 的成熟运用, 产生了大量的基于高通量计算的材料属性数据, 材料数据库也不断扩展. 同时, 机器学习、大数据挖掘的快速兴起, 使得基于大数据和机器学习的材料性能预测方法逐渐成为了可能[11−12]. 从已有的材料数据库中进行数据挖掘, 提取有用的特征, 将这些信息应用于未知材料领域的性能预测, 与 DFT 相比可以使计算量下降多个数量级.

目前, 将机器学习方法应用于材料性能预测领域的研究取得了一定的进展. Ramprasad 等[13] 概括了一些数据驱动的材料信息学的成功应用, 特别指出特征描述符的重要性, 认为选择合适的特征描述符在确定控制复杂现象的关键物理因素方面非常有效, Fujimura 等[14] 基于第一性原理 (DFT), 将理论和实验数据集相结合, 利用机器学习方法以预测材料在 373 K 的导电性, 可以加速锂超离子导体材料设计. Meredig 等[15] 利用 DFT 理论计算得到数据库并构建机器学习模型, 由此产生的模型可以预测材料的热力学稳定性, 与 DFT 计算结果进行对比, 发现拟合精度较好的同时, 计算量降低 6 个数量级, 计算速度得到了较大的提升, 说明机器学习方法适用于高通量计算材料筛选. Xu 等[16] 利用组分替换策略, 在已有的高通量计算材料数据的基础上产生大批量相似新型材料体系, 通过皮尔逊线性相关性系数图谱进行特征选择, 使用基本的机器学习回归算法, 有效且准确地预测了类金刚石组分替换产物的带隙值. Wang 等[11] 总结了基于机器学习的热点材料的研究进展, 指出了有限的数据量是机器学习方法实现应用的主要障

碍. 材料研究进展到目前为止, 通过结合机器学习技术与高通量计算手段来研究材料体系大都是进行性能预测, 对于特征描述符对材料性能影响的可视化分析的研究较少. 在材料性能预测中, 特征描述符是预测材料性能的决定性因素[17], 找到特征和目标属性之间隐藏的映射关系是机器学习的主要目标. 通过可视化可以更加直观了解和分析特征描述符在材料属性中的影响程度和相关性, 对于选择合适的特征描述符提升材料属性预测的准确性和理解材料内在机理具有重要意义.

本文通过实验, 将不同的特征选择方法和机器学习模型进行组合和比较, 寻找最优的组合方案来预测材料的弹性模量[18], 并设计交互式可视分析系统, 将各个特征描述符对材料弹性预测的性能影响进行直观表示, 探究单个特征描述符对弹性模量进行范围预测和弹性模量反推特征可行域的可能性, 为加快材料粗筛速度和深入了解弹性性能的影响因素提供参考.

1　研究方法

本文通过开源的无机化合物材料数据库 Materials Project[19] 获取材料数据集. 数据库内含基于第一性原理高通量计算得到的材料属性, 选择弹性模量中剪切模量 Voigt 及 Reuss 的平均值 G_v 和 G_r,体积模量 Voigt 及 Reuss 的平均值 K_v 和 K_r 四个参数作为目标属性, 选择数据库中含有的材料信息作为输入, 首先进行不同方式的特征选择操作, 构建对于弹性模量具有有效预测能力的特征子空间, 然后在训练集上学习获得不同类型的机器学习模型来预测目标属性, 计算并比较预测精度, 得到最优的性能预测组合. 然后通过可视化的手段分析各特征描述符对弹性模量预测的影响程度, 找出对于弹性模量预测最重要的特征描述符, 探究重要的特征描述符独立预测弹性模量范围和弹性模量反推特征可行域的可能. 实验流程如图 1 所示, 特征选择方法、机器学习模型和实验的具体流程将在下文进行介绍.

1.1　数据集

文中采用的硅材料数据集来源于 MP(Materials Project)数据库. MP 数据库是由加州伯克利大学劳伦斯实验室(Lawrence Berkeley National Laboratory)及麻省理工学院(MIT)在 2011 年发起的项目, 旨在通过计算所有已知材料的属性, 挖掘材料特性, 加速材料研究的创新. 本实验中选取的是 V2021.03.22 版本, 数据库中共收录无机化合物 131613 种, 可为深度学习提供大量可靠的材料数据. MP 数据库还可通过材料应用编程接口(MAPI)和开源 Python 材料基因组学(Pymatgen)[20] 材料分析包提供材料的相关数据和分析. 因此, 数据获取步骤是利用 MAPI 从 MP 数据库中检索硅材料的相关信息作为机器学习的目标输出. 本文选择了 20 种常见的金属非金属元素构成的材料. 然后, 对候选材料进行筛选, 因为 MP 数据库中并不是所有材料都计算了弹性模量, 因此弹性模量属性为空的材料需要去除. 除此之外, 因为数据库中材料的弹性模量是通过高通量计算得到的理论值, 并没有考虑晶体系统的 Born–Huang 弹性稳定性标准[21], 如果高通量计算的弹性张量特征值是负的, 说明违反了稳定性标准, 表明该化合物在零温度下机械不稳定或者计算错误, 考虑到机械不稳定或者弹性张量计算错误的材料会对预测结果造成干扰, 本文还将弹性模量有警告提示的材料去除. 最后, 再筛选去除重复检索的化合物 (例如 $CuFeS_2$ 可由 Cu 元素检索得到, 但会被 Fe 元素重复检索) 得到 1143 种无机化合物作为数据集.

1.2　特征描述符

利用 MP 数据库和 Pymatgen 获取数据集材料的 40 个特征描述符, 这些特征描述符可分为 3 个方面: 14 个化合物基本属性 (能量, 每原子能量, 原子形成能, 体积, 原子数, Ehull, 空间群编号, 密度, 总磁化强度, 最大电负性, 元素数, 最低未占分子轨道, 金属特性, 平均电负性), 20 个元素属性 (原

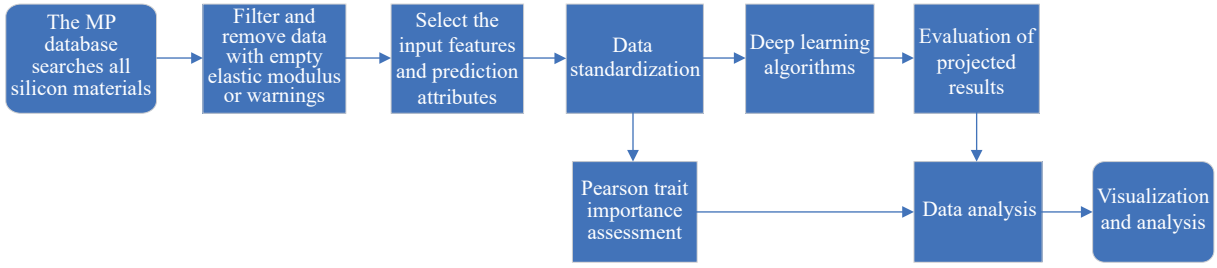

图 1　实验流程图

Fig.1　Experimental flowchart

子序数, 熔点; 分别填充到 10 个维度), 6 个晶体空间结构属性 (键长, 角度). 其中, 因为构成化合物的元素种类并不都是相等的 (例如 SiO_2 由两种元素构成, 而 $CuSO_4$ 则有三种), 为满足数据等长, 设置构成材料的元素数量为 10 个单位长度, 若构成化合物的元素种类少于 10 种, 用“0”对“原子序数”和“熔点”进行填充;“键长”和“角度”分别有 3 个子属性: a, b, c 和 α, β, γ.这 40 个特征描述符作为实验的输入特征集, 然后从 MP 数据库中获取化合物的 G_r, G_v, K_r, K_v 数据, 分别选取其中的一种作为预测目标, 构成实验的输出目标集.

1.3 特征选择

特征输入是机器学习的关键环节, 不同的特征描述符对于预测目标的影响是不同的. 例如在预测材料稳定性时, 分子内部相邻两个原子 (或离子) 间的相互作用力强弱起到了决定性作用. 特征描述符之间也可能会有一定的相关性, 例如粒子平均能量由分子总能量和粒子数量决定. 特征描述符的细粒度也有差异性, 这里的细粒度可以理解为影响因素的复杂程度, 例如带隙相较于体积, 是更加精细的特征描述符. 如果需要提高预测目标的准确性, 特征描述符必须足够精细且与预测目标强相关, 以便模型能够对细节进行学习. 一般来说, 特征描述符越精细. 独立特征数量越多, 预测准确度也越高, 但拟合难度也更高, 更加费力. 相反, 较为粗糙的特征集合, 可以对预测目标进行范围估计, 用于材料的快速初始筛选. 因此, 材料性能预测任务的实质就是预测模型学习特征描述符与目标性能之间的映射关系, 其中每个特征都作为一个因素影响到预测结果. 特征选择的目标就是找到一个特征子空间, 使特征之间尽量相互独立且能较好的表征预测结果, 从中除去无关变量和冗余变量, 改善预测性能的同时减小计算量. 特征选择的方式主要有三种: 过滤式选择、包裹式选择和嵌入式选择.

1.3.1 过滤式特征选择

过滤式特征选择按照发散性或者相关性对各个特征进行评分, 设定阈值或者待选择阈值的个数来选择特征. 本文中的特征属性既有离散型也有连续型, 而且除了补零部分外, 特征取值基本不同, 所以采用单变量特征选择, 主要原理是分别单独计算每个变量的某个统计指标, 根据该指标来判断哪些指标重要, 剔除那些不重要的指标. 本文采用了皮尔森相关系数来分析并识别出具有高度相关的特征, 这种方法较简单且易于运行, 通常对于理解数据有较好的效果. 皮尔森相关系数的计算公式如下:

$$\rho_{x,y}=\frac{(X-\mathrm{mean}(X))(Y-\mathrm{mean}(Y))}{\mathrm{std}(X)\mathrm{std}(Y)} \tag{1}$$

式中, 假设有两个变量 X 和 Y, mean() 代表变量的均值, std() 代表变量的标准差. 这里计算出的相关系数严格来说是样本相关系数, 还需要从样本推论到总体, 做假设检验. 假设输入 (X) 和输出 (Y) 都满足正态分布, 利用样本提供的信息对提出的假设进行检验. 下面令:

$$f=\frac{\rho^2}{1-\rho^2}\cdot(n-2) \tag{2}$$

式中, n 表示每个组别的样本大小, ρ 代表皮尔森相关系数. 如果 X 和 Y 都服从正态分布, 则 f 服从 $F(1,n-2)$, 这个统计量用来检验正态假定下总体中两个变量之间的相关性. f 值越大, 说明特征和目标参数之间的相关性就越大. 因此, 根据 f 值的大小进行特征选择.

1.3.2 包裹式特征选择

本文采用了递归特征消除 (Recursive feature elimination, RFE), 使用一个基模型来进行多轮训练, 每轮训练后移除若干权值系数的特征, 再基于新的特征集进行下一轮训练. 主要操作为对特征含有权重的预测模型, RFE 通过递归减少考察的特征集规模来选择特征. 首先, 预测模型在原始特征上训练, 每个特征指定一个权重. 然后, 那些具有最小绝对值权重的特征从特征集中去除. 如此往复递归, 直至剩余的特征数量达到所需的特征数量.

1.3.3 嵌入式选择

基于机器学习模型的特征选择方法也是一种主流的方法. 有些机器学习方法本身就有对特征进行打分的机制, 很容易将其运用到特征选择任务中. 通过学习模型在训练过程中自动进行了特征选择, 当维数较大样本较少时, 容易陷入过拟合. 故加入正则项, 使用 L1 范数作为惩罚项使得大部分特征对应的系数为 0, 更容易得到稀疏的解, 从而减少特征的维度以简化任务. 常用的稀疏预测模型有 LASSO 和 LinearSVC. 其中 LASSO 运用于回归任务, LinearSVC 应用于分类任务, 本文采用 LASSO 进行特征选择.

1.4 可视化分析

对特征描述符的重要性进行评价, 有助于分析影响材料弹性模量预测的关键特征属性, 还能

够了解材料空间结构、电子结构、能量等属性与弹性性能之间的内在规律. 本文设计了材料弹性预测可视化系统, 通过图表直观显示特征描述符对弹性模量的影响程度, 进而判断特征对目标变量的重要性. 除此之外还对训练集和测试集的数据可视化, 表征弹性模量的变化范围, 显示机器学习方法的拟合性能. 还通过可视化分析, 探究重要的特征描述符独立应用于弹性模量范围预测的可能, 为指定性能要求的新材料初步快速筛选提供了新的解决思路.

1.5　回归模型

根据数据库提供的丰富材料特征数据, 通过机器学习的模型学习关于特征描述符–目标属性的映射关系, 预测未知目标属性的值. 由于材料的弹性模量均为连续值, 因此对弹性模量进行预测属于机器学习中典型的回归问题. 文中使用回归任务中较为常用的 4 种机器学习回归预测模型: LASSO[22−23]、SVR[24]、GBDT[25]、MLP[26]. 这四种算法具有不同的模型特性和优势, LASSO 回归的特点是在拟合广义线性模型的同时进行变量筛选和复杂度调整, 具有良好的线性表达能力;SVR 可以用来处理非线性数据, 通过选择不同核函数的非线性映射将数据投影至特征空间, 然后在特征空间使用线性回归, 具有低维的计算成本而实际的回归效果表现在高维上;GBDT 在传统机器学习算法里是对真实分布拟合的最好的几种算法之一, 预测性能较好; MLP 是当前机器学习领域普遍应用的算法, 具有很强的自适应学习能力, 能处理复杂的非线性系统. 探索这些算法在弹性预测任务中的表现, 与特征提取方法组合, 获得最优的预测模型.

1.5.1　LASSO 回归

LASSO 是以缩小特征集为思想的压缩估计方法. 它在损失函数中引入了正则化 L1 范数惩罚项, 减少输入变量数量进而控制模型的复杂度, 可以解决线性回归出现的过拟合问题. LASSO 回归的损失函数为:

$$\mathrm{Loss}_{\mathrm{LASSO}}=\frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{k}\left|w_j\right|\right] \tag{3}$$

式中, m 表示样本大小, $h_\theta()$ 是回归模型, w 是回归系数. λ 是正则化参数, 作用是控制平衡拟合训练的目标和保持参数值较小.

1.5.2　SVR

SVR 是支持向量机(Support vector machine, SVM)对回归问题的一种运用. SVR 回归是要找到一个回归平面, 让一个集合的所有数据到该平面的距离最近; SVR 认为只要 $f(x)$ 与 y(预测值和真实值)的偏离程度不要太大, 既可以认为预测正确, 不用计算损失. 具体的说就是设置阈值 τ, 只计算 $|f(x)-y|>\tau$ 的数据点的 loss. SVR 的目标函数为:

$$T_{\mathrm{SVR}}=\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}l_\tau\left(f(x_i)-y_i\right) \tag{4}$$

式中, C 为正则化常数, ω 为划分超平面的法向量, l_τ 为损失函数:

$$l_\tau=\begin{cases}|f(x)-y|-\tau, & |f(x)-y|>\tau\\ 0, & |f(x)-y|\leqslant\tau\end{cases} \tag{5}$$

由于特征空间维数可能很高, 高维计算通常是困难的, 所以需要设计核函数, 使得非线性回归问题在经过核函数的转换后可以变成一个近似线性回归的问题. SVR 引入核函数之后, 可重写为:

$$f(x)=\sum_{i=1}^{m}a_iy_ik(x,x_i)+b \tag{6}$$

式中, $k(x,x_i)$ 为核函数, a_i 为拉格朗日系数, $a_i\geqslant0$, b 为偏置系数. 常用的核函数有线性核、多项式核、高斯核, 根据具体问题选择性能最优的核函数.

1.5.3　GBDT

GBDT 全称梯度下降树, 是通过采用加法模型以及不断减小训练过程产生的残差来达到将数据分类或者回归的算法. 经过多轮迭代, 每轮迭代产生一个弱分类器, 每个分类器在上一轮分类器的残差基础上进行训练. 训练的实质是通过降低偏差来不断提高最终分类器的精度. GBDT 算法可以看成是 M 棵决策树组成的加法模型, 其计算公式如下:

$$f(x,p)=\sum_{m=0}^{M}K_m\mathrm{Tree}_m(x,p_m) \tag{7}$$

式中, x 为输入样本, p 为模型参数, Tree 为分类回归树, K 为每棵树的权重.

1.5.4　MLP

MLP 是一种前馈人工神经网络模型, 其将输入的多个数据集映射到单一的输出的数据集上. MLP 的一个重要特点就是多层, 第一层称之为输入层, 最后一层称之有输出层, 中间的层称之为隐藏层. MLP 并没有规定隐藏层的数量, 因此可以根据各自的需求选择合适的隐藏层层数. 因为在上述预测弹性模量的任务中, 数据集并不大且维度较低, 因此为了避免过于复杂的神经网络造成过拟合现象, 本文只涉及了一个隐藏层, 选择 MSE 损失作为回归损失函数.

1.6 评价指标

文中主要采用均方根误差 R_{mse} 和拟合优度 R^2 作为预测模型性能评价指标. R_{mse} 衡量的是真实值和预测值之间的误差, R^2 的最大取值为 1, 取值越接近 1 表明拟合程度越好, 其计算方法如下:

$$R_{\mathrm{mse}}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \tag{8}$$

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2} \tag{9}$$

式中, y 代表真实值, $\hat{y}$代表模型预测值, $\bar{y}$代表样本均值, m 代表样本个数.

2 实验结果与分析

采用上文提到的 3 种特征选择方法和 4 种机器学习模型进行组合, 总共得到 9 种组合方式来对材料的弹性模量进行预测. 选用 MP 数据集中常用元素构成的 1143 种无机化合物作为数据集, 通过十折交叉验证将数据集随机分成 10 组, 以 R_{mse} 和 R^2 作为评价指标, 显示模型的预测性能.

2.1 预测结果与分析

基于材料训练集, 对 9 种不同的回归组合算法进行网格搜索调参, 产生各自的最优回归模型, 将其在测试集上的弹性模量预测结果与 MP 数据库中直接高通量计算的弹性模量进行比较, R_{mse} 和 R^2 指标如表 1、表 2 所示. 图 2 为材料的 4 种弹性模量在 Pearson-GBDT 模型下的真实值−预测值图, 较为直观地展示了预测结果.

结合表 1、表 2 中可以看出, Pearson、RFE 和 LASSO 这三种特征提取方法的效果基本相同, Pearson 和 RFE 稍好于 LASSO. 在回归模型选择上, GBDT 对于弹性模量的预测效果最好, SVR 与

表 1 组合模型对 4 种弹性模量的预测结果 (R_{mse} 值)

Table 1 Prediction results of the four elastic moduli of the combined model (R_{mse})

Feature extraction	Regression model	Model parameters	R_{mse}			
			G_{r}	G_{v}	K_{r}	K_{v}
Pearson	LASSO	λ=0.01	25.27	24.48	29.26	28.12
Pearson	SVR	kernel='rbf', C=1000, gamma=0.01, epsilon=0.01	19.66	18.11	19.15	17.90
Pearson	GBDT	n_estimators=1500, max_features=None, subsample=0.6, learning_rate=0.03	**4.31**	**3.87**	4.62	4.45
RFE	LASSO	λ=0.01	25.27	24.48	29.26	28.12
RFE	SVR	kernel='rbf', C=1000, gamma=0.01, epsilon=0.01	19.66	18.11	19.15	17.90
RFE	GBDT	n_estimators=1500, max_features=None, subsample=0.6, learning_rate=0.03	4.37	4.03	**4.54**	**4.38**
LASSO	SVR	kernel='rbf', C=1000, gamma=0.01, epsilo n= 0.01	19.66	18.11	19.15	17.90
LASSO	GBDT	n_estimators=1500, max_features=None, subsample= 0.6, learning_rate=0.03	4.325	3.93	4.56	4.35
—	MLP	hidden_layer_sizes=(30), activation='relu', solver='adam', max_iter=10000	18.12	16.09	21.14	19.17

表 2 组合模型对 4 种弹性模量的预测结果 (R^2值)

Table 2 Prediction results of four elastic moduli of the combined model (R^2 value)

Feature extraction	Regression model	Model parameters	R^2			
			G_{r}	G_{v}	K_{r}	K_{v}
Pearson	LASSO	λ=0.01	0.57	0.62	0.83	0.83
Pearson	SVR	kernel='rbf', C=1000, gamma=0.01, epsilon=0.01	0.72	0.77	0.9	0.91
Pearson	GBDT	n_estimators=1500, max_features=None, subsample=0.6, learning_rate=0.03	**0.77**	**0.8**	**0.9**	**0.91**
RFE	LASSO	λ=0.01	0.57	0.62	0.83	0.83
RFE	SVR	kernel='rbf', C=1000, gamma=0.01, epsilon=0.01	0.72	0.77	0.9	0.91
RFE	GBDT	n_estimators=1500, max_features=None, subsample=0.6, learning_rate=0.03	**0.77**	**0.8**	**0.9**	**0.91**
LASSO	SVR	kernel='rbf', C=1000, gamma=0.01, epsilon=0.01	0.72	0.76	0.9	0.91
LASSO	GBDT	n_estimators=1500, max_features=None, subsample=0.6, learning_rate=0.03	0.76	0.8	0.9	0.91
—	MLP	hidden_layer_sizes=(30), activation='relu', solver='adam', max_iter=10000	0.71	0.75	0.9	0.9

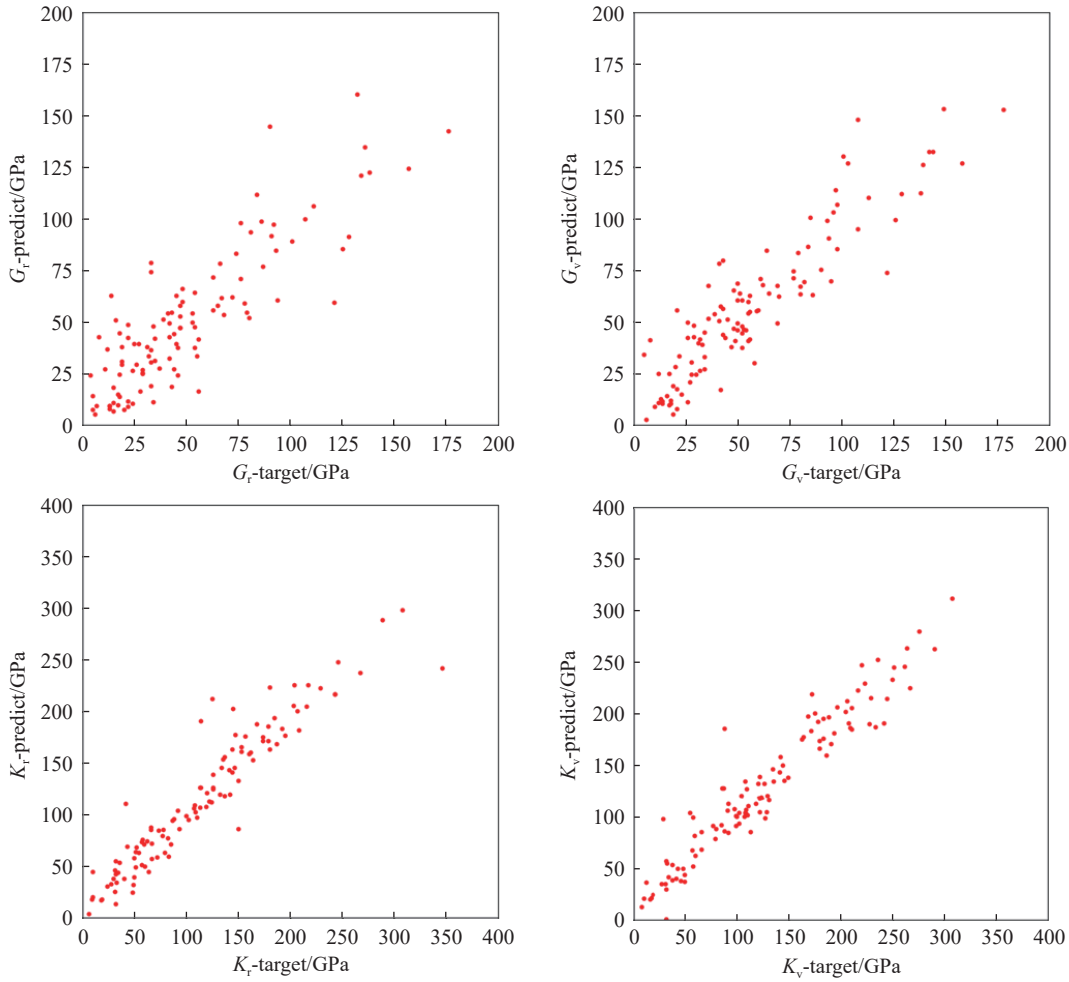

图 2　4 种弹性模量的最佳预测模型真实值–预测值图

Fig.2　Plot of true and predicted values of the best prediction model for four elastic moduli

MLP 的效果较好, 而 LASSO 的预测效果最差. 最佳的模型组合是 Pearson 或 RFE 与 GBDT 的预测模型, 对 K_r 和 K_v 的预测拟合优度分别达到了 0.90 和 0.91, G_r 和 G_v 的预测性能相比于体积模量 K 稍差, 但也达到了 0.77 和 0.80. GBDT 通过采用加法模型以及不断减小训练过程产生的残差来达到将数据回归, 其优点是可以灵活处理各种类型的数据, 包括连续值和离散值, 在相对较少的调参时间情况下, 预测的准确率也比较高, 在使用一些健壮的损失函数, 对异常值得鲁棒性非常强. 而在实验中, 数据集材料包含 40 个特征, 这些特征描述符可分为 3 个方面, GBDT 能灵活处理这些特征描述符, 从而得到最优的预测. 总的来说, 组合模型对于弹性模量的预测表现较好, 其中特征提取方法对于性能的预测影响不大, 预测性能的主要影响在于回归模型的选择上, 体积模量 K 的预测性能好于剪切模量 G, 我们认为这是因为材料的体积模量与单位体积化学键的键能和密度关系很大, 而输入的特征描述符包括了体积、键能和密度的相关信息, 因此模型较容易从中学习特征–性能的映射关系, 而影响剪切模量的因素相对比较复杂, 输入端缺失部分重要的特征描述符, 是导致剪切模量预测性能差于体积模量的主要原因.

2.2　可视化分析

下面通过可视化的方式, 分析影响无机化合物弹性模量预测的关键材料属性, 通过皮尔森特征提取获得各特征属性的 f 值, f 值大小与特征重要性成正比, 可作为特征得分, 从而绘制对应的特征重要性评价图. 如图 3 所示, 直观显示了影响弹性性能预测的重要材料属性.

其中最重要的 4 个特征描述符的相对影响率在表 3 中进行了总结. 为探究单个特征描述符与弹性性能预测的关系, 本文还绘制单个特征值–弹性模量预测值散点图. 如图 3 所示, 我们发现一个重要的特征描述符可以从特征–目标的映射关系中推导出目标属性的范围. 反过来, 一个大致的属

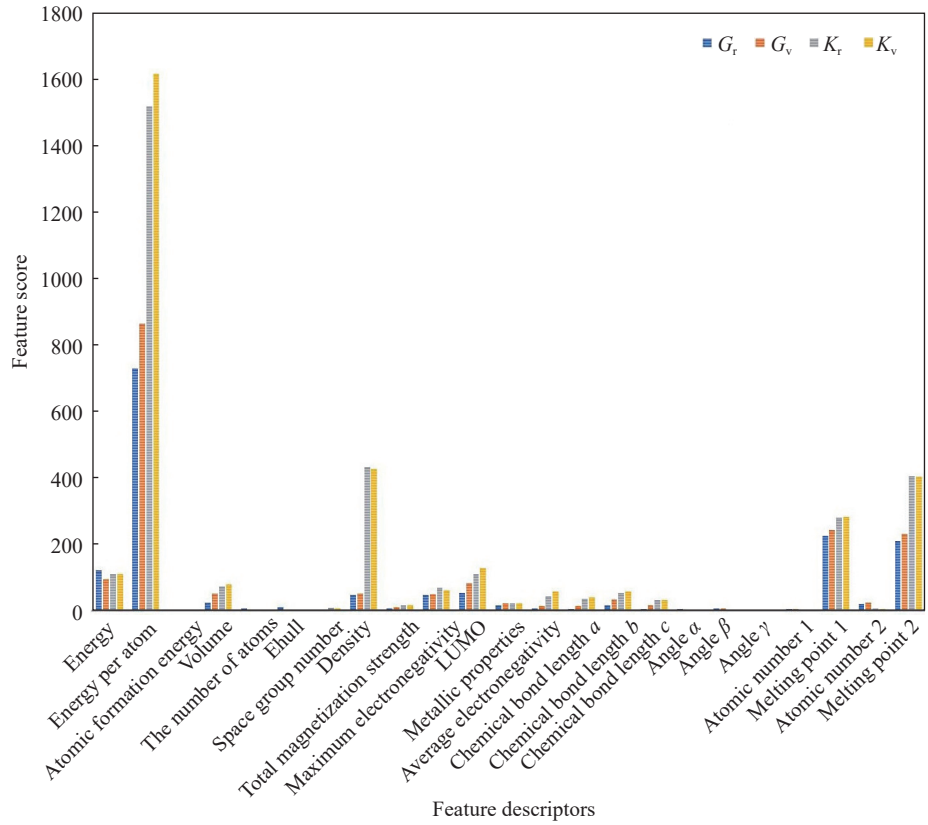

图 3 特征重要性评价图

Fig.3 Evaluation of feature importance

表 3 最重要的 4 个特征描述符的相对影响率和排序

Table 3 Relative impact rates and rankings of the four most important feature descriptors

Elastic modulus	Serial number	Feature descriptors	Relative impact rate/%
G_r	1	Energy per atom	46.1
	2	Melting point 1	14.3
	3	Melting point 2	13.4
	4	Energy	7.8
G_v	1	Energy per atom	47.1
	2	Melting point 1	13.3
	3	Melting point 2	12.6
	4	Energy	5.2
K_r	1	Energy per atom	47.0
	2	Density	13.4
	3	Melting point 2	8.7
	4	Melting point 1	8.6
K_v	1	Energy per atom	48.2
	2	Density	12.7
	3	Melting point 2	12.0
	4	Melting point 1	8.4

性范围也可以用来粗略估计关键特征的相应数值. 例如每原子能量–K_v 的预测关系图 4(c) 中, 选择每原子能量 2 eV, 则 K_v 的范围大致在 0 ~ 100 GPa; 同样若已知 K_v 为 300 GPa, 则材料的每原子能量大概在 8 ~ 10 eV 之间. 这个规律可以用于材料弹性预测的粗筛 (不需要较为精确的目标值), 降低对输入特征数的要求. 还可以根据所需的弹性目标值, 反推材料的关键特征, 从而应用于新材料的发现与合成.

2.3 交互界面

为方便用户对材料性能进行分析研究, 本文提供了一个用于材料弹性预测和分析的可视分析界面, 帮助用户从中获取材料的相关信息与预测分析. 系统交互界面如图 5 所示, 其分为 4 个区域. 区域 A 是数据导入和参数设置的主要区域, 该模块内含常用材料数据集的下载地址, 用户可以导入相关数据集, 也可以根据自己的需求自定义数据集. 用户可根据自身需求选择对应的特征和弹性性能并输入相关数值, 作为单特征–弹性模量范围预测的输入参数. 区域 B 是预测模块, 会根据区域 A 的参数设置预测材料弹性性能, 还可以计算特征值和弹性预测范围, 应用于材料的筛选任务.

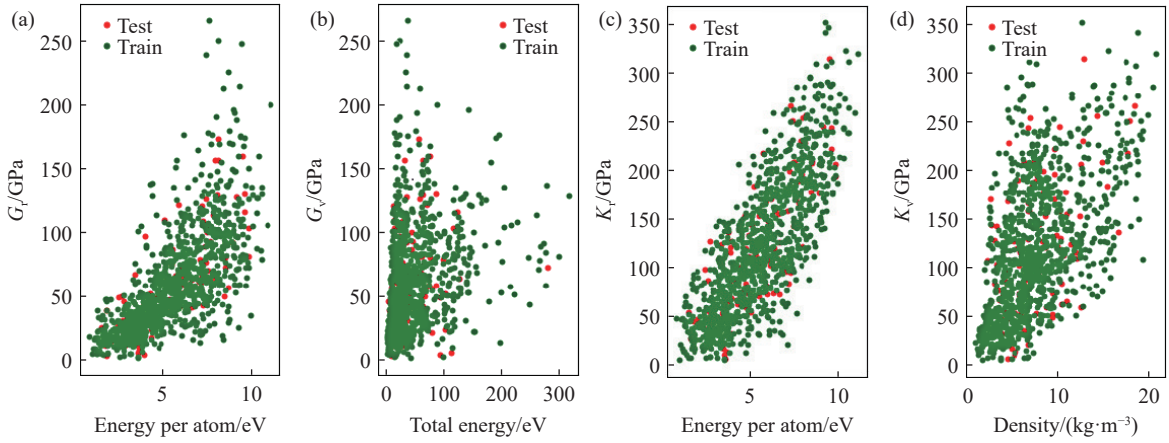

图 4　单特征–弹性模量预测图. (a) 每原子能量–G_v 预测关系; (b) 总能量–G_v 预测关系; (c) 每原子能量–K_v 预测关系; (d) 密度–K_v 预测关系

Fig.4　Single feature–modulus of the elasticity prediction plot: (a) prediction of the energy–G_v relationship per atom; (b) prediction of the total energy–G_v relationship; (c) prediction of the energy–K_v relationship per atom; (d) prediction of the density–K_v relationship

图 5　弹性预测可视化系统交互界面

Fig.5　Elastic prediction visualization system interface

区域 C 是图片选择模块, 显示了数据分析的不同类别, 用户在此区域进行选择, 结果将展示在可视化分析模块中. 区域 D 是可视化分析模块, 用户可获取弹性模量的真实值—预测值图、特征重要性评价图、单特征描述符–弹性模量预测图和特征相对影响图, 这些统计图是模型预测性能和特征–弹性模量相关性的直观显示, 方便用户了解弹性预测的深层次.

3　结语

文中使用了 3 种特征选择方法和 4 种机器学习模型组成 9 种组合模型, 对 MP 数据库中常用元素构成的 1143 种无机化合物的弹性模量进行了预测, 并进行模型性能比较.

(1) 实验发现 3 种特征选择方法对于预测的性能影响差异很小, 而不同的机器学习模型存在较

大差异, 其中 GBDT 的预测性能在 4 种方法中是最优的, Pearson/RFE-GBDT 组合模型的预测结果最好, 对 K_r 和 K_v 预测的拟合优度分别达到了 0.90 和 0.91, GBDT 能灵活处理数据集中不同类型的特征描述符, 从而得到最优的预测.

(2) 对特征描述符的预测重要性进行量化分析, 发现每原子能量、元素熔点、总能量、密度等特征描述符, 在弹性性能预测的过程中具有较大影响, 重要的特征符可以单独作用于弹性预测, 得到材料的弹性预测范围, 弹性模量也可反推材料关键特征属性的可行区间, 此发现有助于加快材料的粗筛和新材料的发现.

(3) 设计了针对材料弹性预测的可视分析系统, 该系统集数据收集与可视分析于一体, 将实验用到的材料数据库、机器学习模型、数据分析算法嵌入其中, 方便用户根据自身需求获取相关数据, 增强材料弹性性能研究的分析能力, 对新型功能材料的研发具有重要意义.

参 考 文 献

[1] Dobrzański L A. Significance of materials science for the future development of societies. *J Mater Process Technol*, 2006, 175(1-3): 133

[2] Lin S. New materials promote high-tech revolution. *Scien Tech Waves*, 1994(6): 36
(林森. 新材料推动高科技革命. 科技潮, 1994(6): 36)

[3] Patel P. Materials Genome Initiative and energy. *MRS Bull*, 2011, 36(12): 964

[4] de Pablo J J, Jones B, Kovacs C L, et al. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr Opin Solid State Mater Sci*, 2014, 18(2): 99

[5] Chen W. High-throughput computing for accelerated materials discovery. *Computational Mater Syst Design*, 2018: 169

[6] Greeley J, Jaramillo T F, Bonde J, et al. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat Mater*, 2006, 5(11): 909

[7] Nityananda R, Hohenberg P, Kohn W. Inhomogeneous electron gas. *Resonance*, 2017, 22(8): 809

[8] Jain A, Hautier G, Moore C J, et al. A high-throughput infrastructure for density functional theory calculations. *Comput Mater Sci*, 2011, 50(8): 2295

[9] Surhone L M, Tennoe M T, Henssonow S F. Vienna ab-initio simulation package (VASP): The guide. Betascript Publishing, 2003

[10] Wang T, Hu S L. High performance computing in materials science. *J Front Comput Sci Technol*, 2017, 11(2): 185
(王涛, 胡双林. 材料科学中的高性能计算. 计算机科学与探索, 2017, 11(2): 185)

[11] Wang X D, Sheng Y, Ning J Y, et al. A critical review of machine learning techniques on thermoelectric materials. *J Phys Chem Lett*, 2023, 14(7): 1808

[12] Liu Y, Zhao T L, Ju W W, et al. Materials discovery and design using machine learning. *J Materiomics*, 2017, 3(3): 159

[13] Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: Recent applications and prospects. *NPJ Comput Mater*, 2017, 3: 54

[14] Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. *Adv Energy Mater*, 2013, 3(8): 980

[15] Meredig B, Agrawal A, Kirklin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys Rev B*, 2014, 89(9): 094104

[16] Xu Y L, Wang X M, Li X, et al. New materials band gap prediction based on the high-throughput calculation and the machine learning. *Sci Sin (Technol)*, 2019, 49(1): 44
(徐永林, 王香蒙, 李鑫, 等. 基于高通量计算及机器学习的新材料带隙预测. 中国科学:技术科学, 2019, 49(1): 44)

[17] Ghiringhelli L M, Vybiral J, Levchenko S V, et al. Big data of materials science: Critical role of the descriptor. *Phys Rev Lett*, 2015, 114(10): 105503

[18] de Jong M, Chen W, Angsten T, et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci Data*, 2015, 2: 150009

[19] Jain A, Ong S P, Hautier G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater*, 2013, 1(1): 011002

[20] Ong S P, Richards W D, Jain A, et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput Mater Sci*, 2013, 68: 314

[21] Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat Commun*, 2017, 8: 15679

[22] Ranstam J, Cook J A. LASSO regression. *Br J Surg*, 2018, 105(10): 1348

[23] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodol)*, 1996, 58(1): 267

[24] Smola A J, Schölkopf B. A tutorial on support vector regression. *Stat Comput*, 2004, 14(3): 199

[25] Friedman J H. Greedy function approximation: A gradient boosting machine. *Ann Statist*, 2001, 29(5): 1189

[26] Ruck D W, Rogers S K, Kabrisky M, et al. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans Neural Netw*, 1990, 1(4): 296