多目标机器学习钢的四种力学性能

魏清华¹, 熊杰², 孙升^{1*}, 张统一^{1*}

1. 上海大学材料基因组工程研究院, 上海 200444;

2. 香港理工大学机械工程系, 香港 999077

*E-mail: mgissh@t.shu.edu.cn; zhangty@shu.edu.cn

收稿日期: 2020-11-29; 接受日期: 2021-01-08; 网络版发表日期: 2021-03-05

国家重点研发计划(编号: 2018YFB0704400)、国家自然科学基金(批准号: 12072179)、高等学校学科创新引智计划(编号: D16002)和香港理工大学(编号: 1-ZE8R和G-YBDH)资助项目

摘要 预报钢材的力学性能对其研发和应用都具有重要作用. 本文对360条源自日本国立材料研究所(NIMS)数据库的钢材数据, 使用正则化线性回归、随机森林和人工神经网络三种机器学习方法对钢材的疲劳强度、拉伸强度、断裂强度和硬度四种力学性能进行多目标机器学习, 即使用一组特征变量来描述这四种目标函数, 并使用MultiTaskLasso、特征重要性和复相关系数三种方法对特征变量的重要程度进行了排序, 也用“面积法”来平衡模型复杂度(特征个数)与预测精度, 且通过线性回归给出了基于化学成分、制备工艺参数、夹杂物参数的解析表达式, 对四种力学性能具有较好的预测能力. 作为对比, 本文还使用原子半径、价电子数和鲍林电负性等9种原子尺度特征替代9种元素特征后进行多目标机器学习建模和预测, 计算结果显示原子尺度特征同样具备较好的预测能力.

关键词 机器学习, 多目标预测, 特征选择, 模型复杂度, 钢材力学性能预测

1 引言

寻找材料的化学成分、制备工艺、微观结构与宏观性能之间的关系, 是材料研发的科学途径. 由于材料的多样性和复杂性, 完全理解、掌握和调控化学成分、制备工艺、微观结构与宏观性能之间的关系一直是材料科学和工程界的追求和巨大挑战^[1]. 近年来, 随着人工智能、机器学习的迅猛发展, 数据挖掘和机器学习正在作为一种有效的方法快速地进入材料研发和生产的各个环节^[1~8], 并由此诞生了材料信息学. 材料

信息学作为材料基因组工程的核心, 旨在将材料研发的成本减少一半、时间缩短一半^[9], 将材料研发和生产更加信息化、智能化. 材料信息学由材料数据驱动, 探索材料的化学成分、制备工艺、微观结构与宏观性能之间的关系, 指导材料研发和生产的创新并发展材料科学的新模型和新理论. 材料信息学和机器学习已经在材料研发的诸多领域都取得了长足的进展^[10~18]. 例如, Wei等人^[10]运用机器学习揭示了纳米多孔石墨烯的新导热机制, 使用遗传算法发现随机无序结构有可能提高热导率, 找到了孔随机排布结构热导率高于

引用格式: 魏清华, 熊杰, 孙升, 等. 多目标机器学习钢的四种力学性能. 中国科学: 技术科学, 2021, 51: 722–736

Wei Q H, Xiong J, Sun S, et al. Multi-objective machine learning of four mechanical properties of steels (in Chinese). Sci Sin Tech, 2021, 51: 722–736, doi: 10.1360/SST-2020-0475

周期排布的例子. Liu等人^[11]通过分而治之的自适应机器学习方法对镍基单晶高温合金的蠕变断裂寿命进行了分析, 研究了影响蠕变寿命的主要因素, 为镍基高温合金的反向设计提供了有效指导. Tong等人^[5]通过五种机器学习方法建立了镁合金孪晶形核预测模型. 在机器学习对钢材力学性能预测的应用中, Xiong等人^[1]通过机器学习方法对钢材的多种力学性能进行单目标预测, 并设计了一种新的抗疲劳钢; Agrawal和Choudhary^[2]使用多种机器学习算法对钢材疲劳强度进行预测, 并开发了一个钢材疲劳强度预测的网站; Wang等人^[3]使用套索回归(least absolute shrinkage and selection operator, Lasso)对工模具钢硬度进行了分析和预测. 但是, 这些模型^[1~3]都是单目标机器学习, 即一个目标需要一组对应的特征量来描述, 多个目标需要多组不同的特征量来描述. 而多目标机器学习仅用一组特征变量来描述多个目标, 可以在不同目标间具有相互冲突的特征变量上, 更好地优化多个材料力学性能, 因而多目标机器学习更有利于材料研发和生产的创新和发展.

本文使用正则化线性回归、随机森林和正则化人工神经网络三种算法对钢材的疲劳强度、拉伸强度、断裂强度和硬度四种力学性能进行多目标预测, 从而能用一组特征量同时预报钢材的这四种力学性能. 还通过线性回归给出这四种力学性能与特征之间的具有较强预测能力的解析式. 在考虑和不考虑模型复杂度(特征个数)两种条件下的特征筛选中, 筛选出两组最佳特征子集. 本文还使用原子半径和、价电子数和鲍林电负性等9种原子尺度特征替代9种元素特征后进行多目标机器学习和建模, 结果表明, 这些原子尺度特征对这四种力学性能同样具有较好的预测能力.

2 数据来源

本文使用的360个样本数据和Xiong等人^[1]所用的数据相同, 源于日本国立材料研究所(NIMS)数据库^[19]的钢材数据集. 这些数据包含钢材的力学性能、化学元素成分、制备工艺和夹杂物参数, 钢材力学性能为在室温测量的旋转弯曲疲劳强度(以下简称疲劳强度)、拉伸强度、断裂强度和硬度. 疲劳强度为钢材在经受 10^7 次交变载荷作用时不产生断裂的最大加载应力区间. 其化学成分包含9种化学元素(C、Si、Mn、

P、S、Ni、Cr、Cu、Mo), 工艺参数为从钢锭到棒材的压缩比、热处理温度. 夹杂物参数指非金属夹杂物的面积占比, 包括dA(塑性加工导致的夹杂物)、dB(不连续排列的夹杂物)和dC(孤立的夹杂物). 借鉴Just模型^[20], 碳的平方根 $\sqrt{C}$ 也作为变量. 这样, 这360个样本数据具有17个特征变量, 每个样本数据中包括10个化学元素变量、1个压缩比变量、3个热处理温度变量、3个夹杂物参数变量和3个力学性能目标. 表1中给出了每个特征变量的最小值、最大值和均值.

3 机器学习算法及评价方法

3.1 算法介绍及算法参数设置

本文使用的MultiTaskLasso、Lasso、线性回归、随机森林和人工神经网络(ANN)算法均来自Scikit-learn中的开源算法包^[21], 算法原理介绍详见附录第1节, 算法参数除特别说明外均用设置的默认值. 在数据预处理过程中, 除随机森林外, 其他算法的建模及特征筛选前均将特征进行了归一化处理. 计算结果表明, 是否对特征进行归一化处理对MutitaskLasso和Lasso算法的预测精度基本无影响, 但归一化后可以使得特征系数在数量级上差别不大, 从而使得MutitaskLasso回归路径图更加清晰; 人工神经网络算法若不对特征进行归一化处理, 模型预测效果很差; 随机森林是基于决策树的算法, 特征是否进行归一化对决策树算法的预测精度基本无影响. 在随机森林中, Bootstrap方法用有放回随机抽样, 在节点分裂时从全部 P 个特征中随机抽取 $p(p < P)$ 个特征; 在人工神经网络中, 随机值用作系数 β 的初始值. 随机森林中的随机数种子random_state设置为1, 人工神经网络中该值设置为100.

在随机森林中, 当决策树个数达到一定数量时, 模型的预测精度将收敛. 本文在建模及特征筛选过程中, 模型预测精度均在决策树个数达到1000之前已收敛, 故决策树个数n_estimators设置为1000. 在对决策树每个节点进行二分时, 从全部 P 个特征中随机抽取一个特征个数为 $p(p < P)$ 的特征子集, 再从这个子集中选择一个最优特征来进行分裂, p 值过大或者过小, 都会降低模型的预测能力. 本文在基于不同特征集建模时, 通过暴力搜索来确定最佳 p 值, 即确定参数max_features的最佳取值, 特征集RF_Best_Subset, RF_C_Subset,

表 1 360个样本数据的17个特征变量^{a)}

Table 1 17 Features of the 360 NIMS fatigue data

特征名称	特征描述	最小值(min)	最大值(max)	均值(Mean)	标准差(StdDev)
NT	正火温度(°C)	825	900	865.6	17.37
QT	淬火温度(°C)	825	865	848.2	9.86
TT	回火温度(°C)	550	680	605	42.4
C (x_1)	碳元素重量百分比(wt.%)	0.28	0.57	0.407	0.061
$\sqrt{C}(\sqrt{x_1})$	碳元素重量百分比的开根号(wt%)	0.53	0.75	0.64	0.05
Si (x_2)	硅元素重量百分比(wt%)	0.16	0.35	0.258	0.034
Mn (x_3)	锰元素重量百分比(wt%)	0.37	1.3	0.849	0.294
P (x_4)	磷元素重量百分比(wt%)	0.007	0.031	0.016	0.005
S (x_5)	硫元素重量百分比(wt%)	0.003	0.03	0.014	0.006
Ni (x_6)	镍元素重量百分比(wt%)	0.01	2.78	0.548	0.899
Cr (x_7)	铬元素重量百分比(wt%)	0.01	1.12	0.556	0.419
Cu (x_8)	铜元素重量百分比(wt%)	0.01	0.22	0.064	0.045
Mo (x_9)	钼元素重量百分比(wt%)	0	0.24	0.066	0.089
RR	压缩比(%)	420	5530	971.2	601.4
dA	塑性加工导致的夹杂物面积占比(%)	0	0.13	0.047	0.032
dB	不连续排列的夹杂物面积占比(%)	0	0.05	0.003	0.009
dC	孤立的夹杂物的面积占比(%)	0	0.04	0.008	0.01

a) 铁元素重量百分比为 $x_{10} = 100 - \sum_{i=1}^9 x_i$.

Processing_only, Composition_only和AF_RF_Best_Subset的最佳取值分别为3, 2, 3, 3和7. 在进行特征筛选时, max_features的取值为变量 $\text{int}(\log_2 p)+1$, 其中 $\text{int}()$ 表示舍去小数点的整数, p 为特征个数.

在人工神经网络中, max_iter为最大迭代次数, alpha为 L_2 范式惩罚项系数, max_iter和alpha分别设置为20000和 1×10^{-5} . 在激活函数选取过程中, 当使用原始特征时, 激活函数使用sigmoid函数时预测性能最佳, 因此在建模时激活函数 f_1 和 f_2 均使用sigmoid函数. 但由于使用sigmoid函数后, 计算耗时过长, 因此对原始特征进行筛选时, 激活函数 f_1 和 f_2 均使用耗时较短的ReLU函数; 当使用原子尺度特征时, 激活函数使用ReLU函数时预测性能最佳, 因此在建模及特征筛选过程中, 激活函数 f_1 和 f_2 均使用ReLU函数. 在建模及特征筛选过程中, 激活函数的使用见表2. 在隐藏层参数优化过程中, 通过网格搜索确定最佳隐藏层参数, 设置最大层数为2, 每层神经元最大个数设置为200, 搜索步长为5, 隐藏层大小hidden_layer_sizes的取值见表2.

3.2 评价指标

K 折交叉验证将 N 个样本的数据集随机均分为 K 份, 将其中的 $K-1$ 份作为训练集, 剩余的1份作为验证集, 并使用训练后的模型对剩余的1份样本集进行预测, 得出预测值. 如此循环 K 次, 得到 K 份预测值. 这 K 份预测值给出了全部数据的验证集预测误差. 当 $K=10$ 时, 称为10折交叉验证法(10 Fold); 当 $K=N$ 时, 每个样本都被单独留出一次作为验证集, 这种交叉验证方法称为留一法(leave one out). 本文在特征筛选过程中, 使用10折交叉验证法和留一法分别进行计算, 发现这三种算法留一法的预测精度均高于10折交叉验证法, 故本文最终选取的交叉验证方法为留一法. 完成交叉验证后, 有三种方法选取模型用于预测新样本. 第一种为使用交叉验证的 K 个模型分别预测新样本, 取 K 个模型预测值均值作为输出值; 第二种为取交叉验证 K 个模型中验证集预测误差最小的模型用于预测新样本; 第三种为交叉验证模型的算法超参数确定后, 使用全部 N 个样本数据根据确定的超参数再训练出一个模

表 2 不同特征集上神经网络参数hidden_layer_sizes的最佳值^{a)}

Table 2 Optimum values of hidden_layer_sizes of ANN on different feature sets

特征集	ANN_Best_Subset	ANN_C_Subset	AF_ANN_Best_Subset	特征筛选过程
f_1, f_2	Sigmoid	Sigmoid	ReLU	ReLU
hidden_layer_sizes	(25, 25)	(25, 25)	(100)	(100)

a) hidden_layer_sizes ($a_1, a_2, \dots, a_n$)表示共有 n 层隐藏层, 其中 $a_i(i=1, 2, \dots, n)$ 表示第 i 层神经元的个数.

型, 用此模型预测新样本.

本文使用皮尔逊相关系数 R 和相对均方根误差 $RRMSE$ 作为预测性能评价指标, 其定义如下:

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^N (y_i - \bar{y}_i)^2 \cdot \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}}_i)^2}},$$

$$RRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}}{\bar{y}_i},$$

式中, $\hat{y}_i$, y_i 和 $\bar{y}_i$ 分别代表预测值、测量值和测量值的均值, N 为样本数量.

4 结果与分析

4.1 特征重要性排序

4.1.1 基于MultiTaskLasso算法的特征重要性排序

MultiTaskLasso多目标回归和Lasso单目标回归均可通过调节参数 α 来筛选特征. 随着 α 的增大, 各特征系数都会变小, 当 α 增大到某一值的时候, 所有系数都会变为0, 系数最后变为0的那个特征, 就是最重要的特征^[3], 依次类推. Lasso回归路径图一般使用 $-\log_{10}(\alpha)$ 为横坐标, 这时, 最先出现非零系数的特征为最重要的特征, 依次类推. MultiTaskLasso多目标回归路径如图1所示, 图1(a)~(d)分别表示疲劳强度、拉伸强度、断裂强度和硬度对应的特征系数随 α 的变化, 随着 α 值从 1×10^2 减小至 1×10^{-4} (横坐标从左至右), 系数不为零的特征个数逐渐增加. 图2为MultiTaskLasso多目标回归和Lasso单目标回归在 $-\log_{10}(\alpha)$ 分别为-1, 0, 0.5, 1时特征系数非零项位置比较图. 在多目标回归时, 4个目标用同一组特征量描述, 因而对应的特征非零系数项位置完全一致; 而在单目标回归时, 每个目标性能用一组特征量描述, 所以4个目标性能对应的特征非零系数项位置不一致. 基于MultiTaskLasso方法给出的

相关特征排序见表3.

4.1.2 基于特征重要性和复相关系数的特征排序

基于随机森林算法给出的特征重要性排序如图3(a)所示. 复目标 y_c 为每个单目标归一化后等权重相加, 即 $y_c = y_1 + y_2 + \dots + y_T$. 复目标 y_c 和特征之间的皮尔逊相关系数 R 叫作多个目标与特征之间的复相关系数, 基于复相关系数绝对值大小的特征排序如图3(b)所示. 特征基于MultiTaskLasso、特征重要性和复相关关系的三种排序中, Cr元素均排在第一位, 即为最重要的特征, 三种方法排序结果中未出现特征重要程度完全冲突的特征.

4.2 特征筛选及预测

4.2.1 特征筛选方法介绍

缩小特征空间有两种基本方式, 一是减少特征的变化范围, 二是减少特征数, 即降低特征空间的维度^[3], 也可以两者共用. 本文通过减少特征数量来缩小特征空间. 本文对17个原始特征计算了特征之间的相关性, 如附录中图A3所示. 结果表明这17个原始特征间的相关系数绝对值都小于0.95, 因而特征之间的相关性没有从这17个原始特征中排除掉任何一个特征.

MultiTaskLasso多目标回归方法自身便具备特征筛选功能, 图4(a)所示为MultiTaskLasso在10折交叉验证、留一法交叉验证、无交叉验证情况下 $1-R$ 值随着 α 变化的折线图. 基于图3(a)所示的随机森林特征重要性排序, 可以得到如图4(b)所示的随机森林 $1-R$ 值对应特征数量变化的折线图. 类似的, 基于图3(b)中的复相关系数排序, 也能得到人工神经网络 $1-R$ 值对应特征数量变化的折线图, 如图4(c)所示.

从图4(a)~(c)中可以看出, 随着特征数量在一定范围内的增加, 模型的预测精度会不断提高, 超过某一值后, 模型的预测精度反而会有所下降, 泛化能力降低.

图1 (网络版彩图)等权重考虑这4个目标性能,进行MultiTaskLasso多目标回归时,疲劳强度(a)、拉伸强度(b)、断裂强度(c)和硬度(d)对应特征的系数值随 α 的变化

Figure 1 (Color online) Four objectives are considered with equal weight in MultitaskLasso, and the feature coefficient varies with the value of α , corresponding to (a) fatigue strength, (b) tensile strength, (c) fracture strength and (d) hardness.

一般选取预测精度变化折线中 $1-R$ 最小值,也就是模型预测精度最高时,所对应的特征集作为最佳特征集。如果要进一步降低模型的复杂度,可以牺牲一些预测精度,这样需要在预测精度和模型复杂度(特征个数)之间找到一个平衡点。本文通过计算预测精度变化折线上的点与坐标轴所围成矩形面积(在本文中记为 $Area$)来平衡预测精度和特征个数,选取 $Area$ 最小时所对应的特征集作为最佳特征集,这种方法称为“面积法”。

4.2.2 基于MultiTaskLasso的特征筛选与建模

如图4(a)所示,在利用MultiTaskLasso多目标回

归进行特征筛选时,留一法时交叉验证的 $1-R$ 值在 $-\log_{10}(\alpha)=0.9$ 时取得最小值,此时对应的特征集记为MTL_Best_Subset,此特征集包含16个特征(Cr, TT, Mo, Ni, QT, $\sqrt{C}$, dB, Mn, Cu, dC, RR, P, Si, C, NT, S);留一法时交叉验证的 $Area$ 值在 $-\log_{10}(\alpha)=0.6$ 时取得最小值,对应的特征子集记为MTL_C_Subset,此特征集包含8个特征(Cr, TT, Mo, Ni, QT, $\sqrt{C}$, dB, Mn)。

使用特征子集MTL_C_Subset(8个特征),线性回归全部360个样本数据,拟合出疲劳强度(Fas, MPa)、拉伸强度(TS, MPa)、断裂强度(FrS, MPa)和硬度(H , HV)4个目标性能的线性解析式:

图 2 MultiTaskLasso多目标回归和Lasso单目标回归在 $-\log_{10}(\alpha)$ 分别为-1, 0, 0.5, 1时17个特征的特征系数非零项位置图Figure 2 When $-\log_{10}(\alpha)$ is -1, 0, 0.5, and 1, the location of non-zero coefficients in MultitaskLasso and Lasso.

表 3 基于MultiTaskLasso方法的特征重要性排序

Table 3 Features ranking based on MultiTaskLasso

特征排序	特征名称	特征排序	特征名称	特征排序	特征名称
1	Cr	7	dB	13	Si
2	TT	8	Mn	14	C
3	Mo	9	Cu	15	NT
4	Ni	10	dC	16	S
5	QT	11	RR	17	dA
6	$\sqrt{C}$	12	P		

$$\begin{aligned} \text{Fas(MPa)} = & 105.82\text{Cr} - 0.87\text{TT} + 179.89\text{Mo} + 26.12\text{Ni} \\ & + 0.02\text{QT} + 492.16\sqrt{C} - 520.17\text{dB} \\ & + 25.30\text{Mn} + 584.38, \end{aligned} \quad (1)$$

$$\begin{aligned} \text{TS(MPa)} = & 211.77\text{Cr} - 1.83\text{TT} + 402.45\text{Mo} + 47.82\text{Ni} \\ & + 0.99\text{QT} + 1211.36\sqrt{C} - 843.85\text{dB} \\ & + 80.95\text{Mn} + 166.05, \end{aligned} \quad (2)$$

图3 (网络版彩图)等权重考虑多个目标时, 特征重要性排序(a)和复相关系数排序(b).

Figure 3 (Color online) Features ranking based on feature importance (a) and multiple (b) correlation, when considering four objectives with equal weight.

图4 (网络版彩图)在留一法交叉验证、10折交叉验证和无交叉验证时, MultiTaskLasso的(1-R)均值随着 α 变化的折线图(a1), 以及随机森林(b1)和人工神经网络(c1)的(1-R)均值随着特征数量变化的折线图. MultiTaskLasso (a2)、随机森林(b2)和人工神经网络(c2)多目标回归折线上的点、坐标轴和坐标轴原点所构成的矩形面积Area随横坐标取值变化的折线图

Figure 4 (Color online) When using leave-one-out cross validation, 10-fold cross validation and no cross validation, the (1-R) value of MultiTaskLasso varies with the value of α (a1), and the (1-R) values of RandomForest (b1) and ANN (c1) vary with different numbers of features used. The Area values of MultiTaskLasso (a2), RandomForest (b2) and ANN (c2) vary with different numbers of features used, and the Area is the rectangular area consisting of the points on the polyline, the axes, and the origin of the axes.

$$\begin{aligned} \text{FrS(MPa)} = & 226.93\text{Cr} - 1.17\text{TT} + 376.99\text{Mo} + 49.32\text{Ni} \\ & + 0.73\text{QT} + 592.27\sqrt{C} - 984.38\text{dB} \\ & + 90.32\text{Mn} + 1106.92, \end{aligned} \quad (3)$$

$$\begin{aligned} H(\text{HV}) = & 64.23\text{Cr} - 0.58\text{TT} + 130.91\text{Mo} + 12.00\text{Ni} + 0.22\text{QT} \\ & + 355.34\sqrt{C} - 320.04\text{dB} + 22.68\text{Mn} + 163.35, \end{aligned} \quad (4)$$

式中, 元素成分的单位为质量百分比(wt%), 特征TT和QT的单位为摄氏度($^{\circ}\text{C}$), 特征dB的单位为面积百分比. 解析式(1)~(4)的预测精度 R 分别为0.9535, 0.9685, 0.9561, 0.9696, 相对均方根误差 $RRMSE$ 分别为4.36%, 3.85%, 2.38%, 3.61%. 从评价指标 R 和 $RRMSE$ 可以看出, 这4种力学性能与特征之间的解析式均具有较高的预测能力. 在这4个解析式中, 元素Cr, Mo, Ni和Mn的系数均为正值, 说明这4种元素为强化元素, 即增加这4种元素的含量有利于提高钢材的疲劳强度、拉伸强度、断裂强度和硬度(规律1); 特征TT和dB的系数均为负值, 特征QT的系数均为正值, 说明降低回火温度(TT)、减少不连续排列的夹杂物(dB)、提高淬火温度(QT)有利于提高钢材的这4种力学性能(规律2). 当特征取值在表1中给定的范围时, 规律1和规律2均成立, 能给实验提供指导, 具有一定的物理冶金意义. 图5(a)

为使用线性回归基于MTL_Best_Subset和MTL_C_Subset两种特征集时做线性回归时的留一法交叉验证 R 和 $RRMSE$ 值. 基于MTL_Best_Subset特征集时, 这四个目标性能的预测值与实验测量值对比的散点图如图5(b)~(e)所示. 从图3(a)中可以看出, 使用包含16个特征的MTL_Best_Subset特征子集时, 多目标回归的 R 均值为0.9627, 而使用包含8个特征的MTL_C_Subset特征子集时, 多目标回归的 R 均值仍高达0.9599, 所用特征减少了8个, 但模型预测能力未出现显著下降.

4.2.3 基于随机森林的特征筛选与建模

如图4(b)所示, 在利用随机森林多目标回归进行特征筛选时, 留一法交叉验证的 $1-R$ 值在特征数为8时取得最小值, 此时对应的特征集记为RF_Best_Subset (Cr, TT, NT, Mo, QT, Mn, Ni, C); 留一法交叉验证的

图5 (网络版彩图) (a) 线性回归基于MTL_Best_Subset和MTL_C_Subset两种特征子集时的留一法交叉验证 R 和 $RRMSE$ 值; 基于MTL_Best_Subset特征集时, 线性回归疲劳强度(b)、拉伸强度(c)、断裂强度(d)和硬度(e)的预测值与实验测量值的对比图

Figure 5 (Color online) The R and $RRMSE$ values of the (a) LinearRegression models using feature set MTL_Best_Subset and MTL_C_Subset with leave-one-out cross validation. The comparison between the predicted and measured values of the LinearRegression models based on feature set MTL_Best_Subset, corresponding to fatigue strength (b), tensile strength (c), fracture strength (d) and hardness (e).

Area值取得最小值时, 对应的特征子集RF_C_Subset仅包含3个特征(Cr, TT, NT).

图6(a)~(e)为随机森林基于RF_Best_Subset和RF_C_Subset两个特征子集时, 留一法交叉验证 R 和RRMSE, 以及基于RF_Best_Subset特征子集时的四个目标性能预测值与实验测量值的对比图. 从图6(a)可以看出, 在RF_C_Subset特征集下, 随机森林算法仅使用3个特征时仍具有较强的预测能力, 此时多目标回归的 R 均值可达到0.9431.

如图7所示, 对于疲劳强度, 在使用特征子集RF_Best_Subset(8个特征)时, 随机森林模型的 R 值为0.9639; 在仅考虑工艺参数和夹杂物参数(Processing_Only, 7个特征)时, 模型的 R 值为0.9407, 预测性能仅出现轻微下降; 当仅考虑化学成分(Composition_Only, 10个特征)时, 随机森林模型的 R 值为0.6559, 散点图中预测值与测量值偏离较大, 模型预测能力偏弱, 这说明热

处理工艺和夹杂物对疲劳强度有较大影响.

4.2.4 基于人工神经网络的特征筛选与建模

如图4(c)所示, 在利用人工神经网络多目标回归进行特征筛选时, 留一法交叉验证的 $1-R$ 的最小值在包含14个特征(Cr, Mo, TT, dB, S, Cu, NT, Ni, dA, Mn, Si, P, RR, C)的特征子集ANN_Best_Subset上取得最小值; 留一法交叉验证的Area值取得最小值时, 对应的特征子集ANN_C_Subset仅包含3个特征(Cr, Mo, TT). 图8(a)~(e)为人工神经网络基于ANN_Best_Subset和ANN_C_Subset两个特征子集, 留一法交叉验证 R 和RRMSE值, 以及基于ANN_Best_Subset特征子集时的四个目标性能预测值与实验测量值的对比图. 从图8(a)中可以看出, 人工神经网络在仅使用3个特征时仍具有较强的预测能力, 此时多目标回归的 R 均值可达到0.9187.

图6 (网络版彩图) (a) 随机森林基于RF_Best_Subset和RF_C_Subset两种特征子集时的留一法交叉验证 R 和RRMSE值. 基于RF_Best_Subset特征集时, 随机森林的预测值与实验测量值的对比图: (b) 疲劳强度; (c) 拉伸强度; (d) 断裂强度; (e) 硬度

Figure 6 (Color online) The R and RRMSE values of the (a) RandomForest models using feature set RF_Best_Subset and RF_C_Subset with leave-one-out cross validation. The comparison between the predicted and measured values of the RandomForest models based on feature set RF_Best_Subset, corresponding to fatigue strength (b), tensile strength (c), fracture strength (d) and hardness (e).

图 7 (网络版彩图)随机森林在RF_Best_Subset(a)和RF_C_Subset(b)两种特征子集上, 以及仅考虑工艺参数和夹杂物参数(c)、仅考虑化学成分(d)时, 多目标预测中疲劳强度预测结果与测量值的对比图

Figure 7 (Color online) The comparison between the predicted and measured values of fatigue strength by RandomForest models based on feature set. (a) RF_Best_Subset; (b) RF_C_Subset; (c) Processing_Only; (d) Compositions_Only.

4.2.5 特征筛选结果及讨论

交叉验证的最高预测精度时的特征集作为最佳特征集, 由此MultiTaskLasso筛选出16个特征, 随机森林筛选出8个特征, 人工神经网络筛选出14个特征. 由于冗余特征的存在, 三种算法在最高预测精度时对应的特征集均不是由全部17个特征所组成的特征集. 基于筛选出的特征, 三种算法预测的4个目标性能 R 的均值分别为0.9627, 0.9731, 0.9777. 牺牲一些预测精度, 可大幅度减少模型复杂度. 使用“面积法”取预测精度变化折线上的点与坐标轴所围成的矩形面积最小时所对应的特征集作为最佳特征集. MultiTaskLasso筛选出8个特征, 随机森林筛选出3个特征, 人工神经网络筛选出3个特征. 这时, 三种算法预测的这4个目标性能 R 的均值分别为0.9577, 0.9431, 0.9187. 这些结果表明, 通过“面积法”能大幅度减少特征个数, 同时保持较高的预测精度.

4.3 基于原子尺度特征的建模

为了进一步提高模型的泛化能力, 本文使用了如表4所示的基于化学成分计算的原子半径和、价电子数和鲍林电负性等9个原子尺度特征. 这9个原子尺度特征加上7个工艺参数和夹杂物参数特征, 组成一个共有16个特征的新特征集. 表4中, r_i , r_{Fe} 分别为元素 i 和铁元素的原子半径; VEC_i , VEC_{Fe} , VEC_C 分别为元素 i , 铁及碳元素的价电子; χ_i , χ_{Fe} , χ_C 分别为元素 i , 铁及碳元素的鲍林电负性; a_i 是元素 i 的原子百分比,

$$a_i = \frac{x_i / M_i}{\sum_{j=1}^{10} x_j / M_j}. \text{ 选择这些原子尺度特征是因为铁和}$$

碳作为钢铁中最重要的两种元素^[3], 其他元素与它们电负性的差异可以表示与他们结合能力的强弱; 而相对铁元素的原子半径差反映了原子尺度上的应变, 即合金元素的引进造成的晶格畸变; 价电子数可以反映一种元素可以用来成键的电子数. 这些原子尺度特征

图8 (网络版彩图) (a) 神经网络基于ANN_Best_Subset和ANN_C_Subset两种特征子集时的留一法交叉验证 R 和 $RRMSE$ 值。基于ANN_Best_Subset特征集时, 人工神经网络的预测值与实验测量值的对比图: (b) 疲劳强度; (c) 拉伸强度; (d) 断裂强度; (e) 硬度

Figure 8 (Color online) The R and $RRMSE$ values of the (a) ANN models using feature set ANN_Best_Subset and ANN_C_Subset with leave-one-out cross validation. The comparison between the predicted and measured values of the ANN models based on feature set ANN_Best_Subset, corresponding to fatigue strength (b), tensile strength (c), fracture strength (d) and hardness (e).

都可能和钢材的疲劳强度、拉伸强度、断裂强度和硬度相关。

用相同的方法, 对上述原子尺度特征进行重要性排序。基于MultiTaskLasso方法给出的相关特征排序如表5所示, 基于特征重要性和复相关系数的特征排序如图9所示。同时, 在这些特征排序的基础上, 结合线性回归、随机森林和人工神经网络, 进行基于原子尺度特征的特征筛选与建模。

图10(a)所示为MultiTaskLasso的 R 值随着 α 变化的折线图, MultiTaskLasso的留一法交叉验证 R 均值在 $-\log_{10}(\alpha)=0.9$ 时取得最大值0.9614, 所筛选出的最佳特征子集AF_MTL_Best_Subset(不包含 $d_{\chi-Fe}$)共包含15个原子尺度特征。

图10(b)和(c)所示为在利用随机森林多目标回归进行特征筛选时, 留一法交叉验证 R 均值在使用15个

特征时取得最大值0.9701, 此时对应的特征集记为AF_RF_Best_Subset(不包含 d_C); 在利用神经网络多目标回归进行特征筛选时, 留一法交叉验证 R 均值在使用15个特征时取得最大值0.9614, 此时对应的特征集记为AF_ANN_Best_Subset(不包含 $d_{\chi-Fe}$)。图11(a)为线性回归基于特征子集AF_MTL_Best_Subset、随机森林基于特征子集AF_MTL_Best_Subset、人工神经网络基于特征子集AF_ANN_Best_Subset, 进行多目标预测时的留一法交叉验证 R 值; 在图9(b)~(e)中, 将随机森林模型对于四个目标性能的预测值与实验测量值进行了对比。

用线性回归、随机森林、人工神经网络和交叉验证法筛选出的特征集进行多目标预测, 当使用元素作为特征时, 预测的这4个目标性能 R 的均值分别为0.9627, 0.9731, 0.9630; 而当使用泛化性能更好的原子

表 4 原子尺度特征

Table 4 Atomic features used in this work

特征	特征描述	公式
a_{Fe}	铁元素的原子个数百分比	a_{10}
t_r	总原子半径	$\sum_{i=1}^{10} a_i r_i$
$d_{r-\text{Fe}}$	原子半径差(铁基)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - \frac{r_i}{r_{\text{Fe}}}\right)^2}$
t_{VEC}	总价电子数	$\sum_{i=1}^{10} a_i \text{VEC}_i$
$d_{\text{VEC}-\text{Fe}}$	价电子数差(铁基)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - \text{VEC}_i / \text{VEC}_{\text{Fe}}\right)^2}$
$d_{\text{VEC}-\text{C}}$	价电子差(碳基)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - \text{VEC}_i / \text{VEC}_{\text{C}}\right)^2}$
t_{χ}	总鲍林电负性	$\sum_{i=1}^{10} a_i \chi_i$
$d_{\chi-\text{Fe}}$	电负性差(铁基)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - \chi_i / \chi_{\text{Fe}}\right)^2}$
$d_{\chi-\text{C}}$	电负性差(碳基)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - \chi_i / \chi_{\text{C}}\right)^2}$

图 9 (网络版彩图)基于随机森林特征重要性(a)及复相关系数(b)的原子尺度特征排序

Figure 9 (Color online) Atomic features ranking based on feature importance (a) and multiple correlation (b), when considering four objectives with equal weight.

表 5 原子尺度特征基于MultiTaskLasso方法的特征重要性排序

Table 5 Atomic features ranking based on MultitaskLasso

特征排序	特征名称	特征排序	特征名称	特征排序	特征名称	特征排序	特征名称
1	$d_{\chi-\text{C}}$	5	$d_{\text{VEC}-\text{Fe}}$	9	TT	13	RR
2	tr	6	NT	10	tVEC	14	dC
3	$d_{\text{VEC}-\text{C}}$	7	dr-Fe	11	aFe	15	dB
4	$d_{\chi-\text{Fe}}$	8	t_{χ}	12	QT	16	dA

图 10 (网络版彩图)当加入原子尺度特征后,在留一法交叉验证、10折交叉验证和无交叉验证时,MultiTaskLasso的 R 均值随着 α 变化的折线图(a);随机森林(b)和人工神经网络(c)的 R 均值随着特征数量变化的折线图

Figure 10 (Color online) When using Leave-One-Out cross validation, 10-fold cross validation and no cross validation with atomic features, the $1-R$ value of MultiTaskLasso varies with the value of α (a); the $1-R$ values of RandomForest (b) and ANN (c) vary with different number of features used.

图 11 (网络版彩图) (a) 基于原子尺度特征,线性回归、随机森林、人工神经网络三种算法,在不考虑模型复杂度的情况下,最佳特征子集的多目标预测精度.基于AF_RF_Best_Subset特征子集时,随机森林的预测值与实验测量值的对比图: (b) 疲劳强度; (c) 拉伸强度; (d) 断裂强度; (e) 硬度

Figure 11 (Color online) The (a) multi-objective prediction accuracy of LinearRegression, RandomForest and ANN with the best feature subset which is selected with considering the model complexity, based on atomic features. The comparison between the predicted and measured values of the RandomForest models based on the best feature set AF_RF_Best_Subset, corresponding to fatigue strength (b), tensile strength (c), fracture strength (d) and hardness (e).

尺度特征时,这4个目标性能 R 的均值分别为0.9614, 0.9701, 0.9627. 这表明使用原子尺度特征,这三种算

法预测的 R 下降幅度均小于1%,模型预测性能基本保持不变.

5 结论

本文使用正则化线性回归、随机森林和人工神经网络等三种多目标机器学习方法, 对钢材的疲劳强度、拉伸强度、断裂强度和硬度进行了预测, 并通过留一交叉验证法对模型性能进行评估. 本文研究的这四个目标性能是强正相关, 因而多目标机器学习模型的预测表现较好, 其主要结论如下.

(1) 使用MultiTaskLasso回归路径中特征系数不为零次序、随机森林特征重要性和复相关系数三种方法, 在等权重考虑四个目标时, 对特征重要程度进行排序. 取模型预测精度最高时所对应的最佳特征集, 线性回归在包含16个特征的特征子集上交叉验证 R 值可达到最大值0.9627, 随机森林在包含8个特征的特征子集上交叉验证 R 值可达到最大值0.9731, 人工神经网络在包含14个特征的特征子集上交叉验证 R 值可达到最大值0.9777.

(2) 通过“面积法”来平衡预测精度与模型复杂度(特征个数), 使得在大幅度减少特征个数时, 预测精度不会出现显著下降. 此时, 线性回归在包含8个特征的特征子集上交叉验证 R 值为0.9577, 随机森林在仅包含3个特征的特征子集上交叉验证 R 值为0.9431, 人工神经网络在仅包含3个特征的特征子集上交叉验证 R 值也可达到0.9187.

(3) 钢材疲劳强度、拉伸强度、断裂强度和硬度的解析式(1)~(4)表明, 在成分配比上, 增加Cr, Mo, Ni和Mn这4种元素含量, 有利于提升这4种力学性能; 在制备工艺中, 降低回火温度、减少不连续排列的夹杂物、提高淬火温度同样有利于提升这4种力学性能.

(4) 使用原子半径之和、价电子数和鲍林电负性等9种原子尺度特征替代元素特征后, 三种算法同样具有较好的预测能力, 所得模型交叉验证的 R 值均大于0.96.

补充材料

本文的补充材料见网络版techcn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

- Xiong J, Zhang T Y, Shi S Q. Machine learning of mechanical properties of steels. *Sci China Tech Sci*, 2020, 63: 1247–1255
- Agrawal A, Choudhary A. An Online tool for predicting fatigue strength of steel alloys based on ensemble data mining. *Int J Fatigue*, 2018, 113: 389–400
- Wang J H, Sun S, He Y L, et al. Machine learning prediction of the hardness of tool and mold steels (in Chinese). *Sci Sin Tech*, 2019, 49: 1148–1158 [王家豪, 孙升, 何燕霖, 等. 基于机器学习的工模具钢硬度预测. 中国科学: 技术科学, 2019, 49: 1148–1158]
- Ramprasad R, Batra R, Pailan G, et al. Machine learning in materials informatics: Recent applications and prospects. *Npj Comput Mater*, 2017, 3: 54
- Tong Z, Wang L, Zhu G, et al. Predicting twin nucleation in a polycrystalline Mg alloy using machine learning methods. *Metall Mater Trans A*, 2019, 50: 5543–5560
- Chen Y, Tian Y, Zhou Y, et al. Machine learning assisted multi-objective optimization for materials processing parameters: A case study in Mg alloy. *J Alloys Compd*, 2020, 844: 156159
- Agrawal A, Choudhary A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater*, 2016, 4: 053208
- Xiong J, Shi S Q, Zhang T Y. A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys. *Mater Des*, 2020, 187: 108378
- Green M L, Choi C L, Hattrick-Simpers J R, et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl Phys Rev*, 2017, 4: 011105
- Wei H, Bao H, Ruan X. Genetic algorithm-driven discovery of unexpected thermal conductivity enhancement by disorder. *Nano Energy*, 2020, 71: 104619
- Liu Y, Wu J, Wang Z, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine

- learning. *Acta Mater*, 2020, 195: 454–467
- 12 Wang C, Fu H, Jiang L, et al. A property-oriented design strategy for high performance copper alloys via machine learning. *Npj Comput Mater*, 2019, 5: 87
- 13 Yi S, Jones R. Machine learning framework for predicting reliability of solder joints. *Solder Surf Mt Tech*, 2019, 32: 82–92
- 14 Chowdhury A, Kautz E, Yener B, et al. Image driven machine learning methods for microstructure recognition. *Comput Mater Sci*, 2016, 123: 176–187
- 15 Wei H, Zhao S, Rong Q, et al. Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods. *Int J Heat Mass Transfer*, 2018, 127: 908–916
- 16 Hansen K, Montavon G, Biegler F, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J Chem Theor Comput*, 2013, 9: 3404–3419
- 17 Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 2016, 533: 73–76
- 18 Xiong J, Zhang T Y, Shi S Q. Machine learning prediction of elastic properties and glass-forming ability of bulk metallic glasses. *MRS Commun*, 2019, 9: 576–585
- 19 Yamazaki M, Xu Y. NIMS materials database and its search engine (Factual databases). *J Inf Sci Technol Assoc*, 2009, 59: 177–182
- 20 Just E. New formulas for calculating hardenability curves. *Metal Prog*, 1969, 95: 87–88
- 21 Swami A, Jain R. Scikit-learn: Machine learning in python. *J Mach Learn Res*, 2013, 12: 2825–2830

Multi-objective machine learning of four mechanical properties of steels

WEI QingHua¹, XIONG Jie², SUN Sheng¹ & ZHANG TongYi¹

¹ *Materials Genome Institute, Shanghai University, Shanghai 200444, China;*

² *Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China*

The prediction of mechanical properties of steels is highly significant for the development and applications of steels. In the present work, we used three multi-objective machine learning algorithms, Regularized linear regression, Random forest and Artificial neural network, to predict four mechanical properties of steels, including fatigue strength, tensile strength, fracture strength and hardness, based on 360 data samples from the Japan National Institute of Material Science (NIMS) database. Three methods, MultiTaskLasso, feature importance and Multiple correlation, were applied to rank features. The “Area method” was proposed to balance the model complexity (number of features) and prediction accuracy. The linear regression, with selected features, of the entire data leads to four analytic formulas, which predict very accurately the four mechanical properties of steels. Furthermore, nine atomic features, such as atomic radius, valence electron number, and Pauling electronegativity, of the involved elements in the steels were replaced the elements in the multi-objective machine learning and the results show that the atomic features are able to accurately predict the four mechanical properties of steels as well.

machine learning, multi-objective prediction, feature selection, model complexity, steels

doi: [10.1360/SST-2020-0475](https://doi.org/10.1360/SST-2020-0475)