基于机器学习的工模具钢硬度预测

王家豪¹, 孙升¹, 何燕霖², 张统一^{1*}

1. 上海大学材料基因组工程研究院, 上海 200444;

2. 上海大学材料科学与工程学院, 上海 200444

* E-mail: zhangty@shu.edu.cn

收稿日期: 2019-02-16; 接受日期: 2019-04-25; 网络版发表日期: 2019-10-08

国家重点研发计划(编号: 2017YFB0701604, 2018YFB0704400)资助项目

摘要 硬度是工模具钢性能的一个重要指标. 本文通过机器学习中的层次聚类和LASSO回归方法, 给出了工模具钢硬度-成分关系的解析表达式. 层次聚类将79种牌号工模具钢的合金元素成分-硬度数据集按欧式距离分为了4簇, 其中高铬钢和低铬钢两簇包含了几乎所有的数据. 然后我们对高铬钢和低铬钢两簇数据分别进行LASSO回归, 画出LASSO路径, 并利用留一法交叉验证得到具有最佳泛化预测能力的成分-硬度的解析公式. 之后引入电负性、原子半径变化率、价电子数、电子亲和能和第一电离能等原子尺度的特征, 重新进行LASSO回归, 得到了新的硬度公式. 结果显示机器学习方法可以成功预报工模具钢的硬度. 同时, 使用原子尺度的特征为钢铁材料的性能研究提供了新的思路.

关键词 机器学习, 工模具钢, 硬度预测

1 引言

随着计算机计算能力的提升和算法的改进, 机器学习作为一种有效的方法已经进入材料的各个领域. 如Xiong等人^[1]利用机器学习的方法预测了大块非晶的机械性能和玻璃形成能力; Agrawal和Choudhary^[2]使用400组NIMS的数据, 利用数据挖掘的方法, 研究了钢铁的疲劳与成分和工艺之间的关系; Aourag^[3]利用机器学习的方法对铝和稀土元素组成的化合物的机械性能进行了预测; Dey等人^[4,5]利用机器学习方法对时效硬化铝合金的机械性能进行了研究. 本文利用机器学习的方法对工模具钢硬度进行了预测.

模具工业水平的高低是衡量国家制造业水平的重要标志^[6], 而工模具钢是模具的重要组成部分^[7]. 目前, 我国工模具钢的发展非常不均衡, 中低端产品供过于求, 而高端的模具钢接近50%依赖进口^[8]. 因此提升工模具钢性能对于在“十三五”期间, 加快我国制造业从“大国”向“强国”发展具有重要的推动作用. 制约我国工模具钢发展的关键因素包括: 生产工艺水平较低、缺乏有关技术标准的规范与引导等^[8].

众所周知, 硬度是工模具钢的一个重要性能指标, 而硬度基本上由合金钢热处理组织结构决定, 特别与工模具钢中马氏体的含量密切相关. 决定马氏体含量的重要因素分别是热处理时的冷却速率、奥氏体开始

引用格式: 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

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, doi: 10.1360/SST-2019-0060

向马氏体转变的温度(马氏体点或马氏体转变温度 M_s)和奥氏体全部转变为马氏体的温度(马氏体终了温度 M_f)。冷却速率是由热处理工艺所决定。对于一个工模具钢, 热处理时的冷却速率必须大于一个临界值, 才能抑制奥氏体向其他热力学稳态相的转变, 实现奥氏体向马氏体的转变。但马氏体转变温度 M_s 和马氏体终了温度 M_f 与冷却速率无关, 完全取决于工模具钢的化学成分。马氏体转变温度越高, 在淬火过程中越容易形成马氏体, 从而工模具钢硬度越高。另一方面, 如果工模具钢是完全由马氏体相组成, 工模具钢将太脆、韧性不够好。由于缺乏包含冷却速率的热处理工艺数据、每种牌号工模具钢的连续冷却C曲线数据(continuous cooling transformation, CCT)和工模具钢其他性能的数据, 本文只探究化学成分对于硬度的影响, 而没有研究热处理工艺和材料微结构对于硬度的作用。

前人有过一些关于使用成分预测硬度的报道。其中早期研究多采用数理统计的方法来建立硬度与成分之间的函数关系, 从而完成性能预报。Just^[9]在1969年提出了直接计算端淬曲线的公式, 其中描述碳的质量分数在0.1%~0.6%之间的钢的淬火端的硬度公式如下:

$$J_0 = 60\sqrt{C} + 20 \text{ (HRC)},$$

式中, J_0 和 C 分别代表淬火端的硬度和碳元素质量分数, HRC是洛氏硬度C标尺测出的硬度。Just^[9]认为在距离淬火端的4个单位长度内(每个单位长度为1.5875 mm), 钢铁的硬度值近似相等, 都为 J_0 。余柏海^[10]基于测量数据和分析, 得到适用于碳质量分数在0.1%~1.5%之间的钢种的淬火端硬度公式:

$$J_0 = \frac{30}{1+C}(C(4.5-C)+1) \text{ (HRC)}.$$

金满等人^[11,12]根据端淬曲线导数的变化规律, 用线性试探法, 建立了关于端淬曲线的微分方程, 之后解出关于硬度的分布函数。然后基于碳的质量分数在0.05%~1.3%之间钢种的硬度数据, 得到的结构钢淬火端硬度公式为

$$J_0 = 66.5 - 47.8\exp(-4C) \text{ (HRC)}.$$

以上公式给出了钢中含碳量与其硬度值的定量关系, 对碳钢的设计研发起到了促进作用, 但对高合金含量的工模具钢而言尚有局限性。当前快速发展的人工智能机器学习提供了一些强大的技术手段, 用以探究特征量之间复杂的非线性关系。近年来已有大量利用

人工神经网络的方式预测成分-硬度关系的研究。例如Vermeulen等人^[13]利用前馈神经网络研究了Nedstaal BV提供在表1成分范围内的4000组数据, 通过训练得到预测模型, 该模型在整个样本中预测值与实际值的平均误差为2 HRC。Filetin等人^[14]使用人工神经网络研究了同来源的40种不同成分钢的数据, 其钢的成分范围如表1所示, 得到的模型预测值与实际值的平均误差为0.0045 HRC。由伟等人^[15]收集了文献中共111组钢的数据, 其钢的成分范围如表1所示, 用神经网络的方法进行处理, 得到的预测模型的均方根误差为0.0868 HRC。

人工神经网络方法能够得到预测精度很高的模型, 但是该方法无法得到显式的解析公式。基于此, 本文拟采用两种机器学习的方法——层次聚类和LASO回归, 尝试得到影响钢铁硬度的主要特征量。除了钢铁中合金元素的质量分数之外, 还在待选特征中引入元素的电负性、原子半径、价电子数、电子亲合能和第一电离能。通过LASSO回归, 选出最佳的特征量。然后利用最小二乘法回归, 研究这些特征量与硬度之间的显式函数关系, 从而建立了可较为准确预测高品质工模具钢硬度值的解析公式, 为实现其理性设计提供参考, 对加快此类钢种的研发具有重要的现实意义。

表1 Vermeulen等人^[13]、Filetin等人^[14]和由伟等人^[15]使用数据集的成分范围(单位为质量分数wt%)

元素	含量		
	Vermeulen等人 ^[13]	Filetin等人 ^[14]	由伟等人 ^[15]
碳	0.09~1.04	0.12~0.7	0.09~0.15
锰	0.35~1.54	0.12~1.2	0.91~1.61
磷	0~0.24	—	—
硫	0~0.23	—	—
硅	0.01~0.34	0.12~1.24	0.23~0.43
铜	0.01~1.26	0.001~0.34	—
铬	0~1.67	0.14~1.96	0~0.62
镍	0.01~1.7	0.14~2.76	0~0.69
钼	0~0.45	0.02~0.35	0~0.54
铝	0~0.066	0.002~0.34	—
钒	—	0.08~0.13	—
硼	—	—	0~0.002
钛+铌+钽	—	—	0~0.16
铁	其余	其余	其余

2 数据来源与方法

本文所使用的数据整理自国家标准工模具钢^[16]. 其数据集共包含了79种不同牌号的工模具钢的成分范围和硬度范围等. 本文取各元素的含量中值和硬度最小值来表示合金成分和硬度值. 处理后的数据见表S1.

本文采用均方根误差(RMSE)作为衡量结果优劣的标准, 其定义如下:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}},$$

其中, $\hat{y}_i$ 和 y_i 分别代表回归公式的预测值与数据真实值, n 为数据数. 本文所使用的机器学习方法为层次聚类和LASSO回归. 层次聚类方法分为自上而下和自下而上两种方式: 自上而下是指, 起初将所有个体分为一簇, 然后按照某种距离的远近, 将这个大簇不断地细分下去, 直至达到某个终止条件; 自下而上的方式是指, 起初将每一个个体看作一簇, 然后将距离比较近的两个簇合并, 不断重复, 直至所有个体都被包含在最大的簇内. 为了减少计算量, 本文所使用的层次聚类方法是采用自下而上的方式进行的. 两个簇之间距离的计算公式如下:

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)},$$

其中, u, v 代表任意两个簇, $u[i], v[j]$ 分别代表 u 和 v 中的元素, $\sum_{ij} d(u[i], v[j])$ 代表 u 中的每个元素 i 与 v 中的每个元素 j 之间的距离之和, $|u|$ 和 $|v|$ 分别代表 u 和 v 中包含的元素数目. 计算元素与元素之间距离的计算公式为欧式距离, 公式如下:

$$d(u[i], v[j]) = \sqrt{\sum_{m=1}^n (u[i]_m - v[j]_m)^2},$$

其中, $u[i]_m$ 和 $v[j]_m$ 分别代表 $u[i]$ 和 $v[j]$ 的第 m 个维度的值, 即元素的质量百分比.

LASSO回归是一种带惩罚的线性回归方法, 常规的线性公式为 $\mathbf{X}(\theta) = \sum \theta_i \cdot x_i + C$, 这里 $\mathbf{X}(\theta)$ 为目标变量, θ_i 为待回归的系数, x 为特征量, C 为常数. LASSO回归的损失函数为

$$L = (\mathbf{X}(\theta) - \mathbf{Y})^T (\mathbf{X}(\theta) - \mathbf{Y}) + \alpha \|\theta\|_1,$$

$\mathbf{Y}$ 代表真实值, 矢量 $\alpha \|\theta\|_1$ 称为惩罚项, α 是可以人为调

节的正参数, 称为缩减系数. LASSO回归得到的结果是让该损失函数达到最小值的 θ . 该损失函数要求在系数的绝对值之和小于一个给定参数的条件下, 数据的预测值与实际值之差的平方和最小. LASSO回归可以将不为零的系数 θ_i 的个数缩减, 如果 α 足够大, 甚至可以使所有 $\theta_i=0$, 因而LASSO回归可以进行特征量选择, 达到在众多特征量中找出主要特征量, 简化公式的目的. 参数 α 是LASSO回归中的重要参数, 其决定了惩罚项在损失函数中的权重, 使用不同的 α 会得到不同的模型. 把 α 作为横坐标, 各个特征量的系数作为纵坐标作图, 就得到了LASSO的路径图. 随着 α 的增大, 所有系数都会变小, 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的那个特征量, 是LASSO认为最重要的特征量, 依次类推. LASSO路径图一般将 $-\log_{10}(\alpha)$ 作为横坐标, 这时, 最先出现系数的特征量为最重要的特征量, 依次类推.

本文所使用的留一法交叉验证求测试集均方根误差的方法是指: 将包含 n 个样本的数据集分为两个部分, 其中含有 $n-1$ 个样本的部分为训练集, 用来训练模型. 并使用训练后的模型对剩下的1个样本进行预测. 如此循环 n 次, 共得到 n 个预测值, 将这 n 个预测值和对应的真实值按照均方根误差的公式求得均方根误差, 这个均方根误差称为留一法交叉验证的测试集均方根误差.

3 结果与分析

首先使用层次聚类的方法对数据成分含量进行聚类分析. 图1表示层次聚类结果. 其中横坐标的数字是对数据的编号, 纵坐标表示聚类簇之间的距离. 例如, 左边的2组数据被分成的簇和右边77组数据被分成的簇之间的距离大约为16.

当使用合金成分作为特征量时, 每一个特征量相当于一个维度, 这79种工模具钢共包含15种元素, 形成了一个15维空间. 由特征量所形成的空间称为特征空间. 数据分析可以得到该元素特征空间中各元素的取值范围如表2所示.

本数据集中一个具体钢种中含有的合金元素数量是不同的, 可从3种变化到11种, 并且每个元素变化的步长也不同. 因此该特征空间内存在超过 2^{64} 种可能的钢种. 特征空间太大和数据太稀疏会导致公式的泛化

图1 (网络版彩图)使用层次聚类的方法处理79组元素质量分数数据得到的结果. 横坐标为钢的编号, 纵坐标为聚类簇之间的距离

表2 各成分质量分数的取值范围、步长及其出现的次数

元素 (特征量)	最小值 (wt%)	最大值 (wt%)	步长 (wt%)	含该元素的数据数
C (x_1)	0.17	2.15	0.005	79
Mn (x_2)	0.2	15.5	0.05	79
P (x_3)	0	0.015	0.005	30
S (x_4)	0	0.015	0.005	30
Si (x_5)	0.15	2	0.005	79
Ni (x_6)	0	2	0.005	29
Cr (x_7)	0	18	0.005	70
Mo (x_8)	0	5	0.005	40
Al (x_9)	0	2.8	0.1	4
Co (x_{10})	0	0.5	0.5	14
V (x_{11})	0	2	0.005	39
Cu (x_{12})	0	0.15	0.15	17
W (x_{13})	0	6.5	0.05	23
Ca (x_{14})	0	0.05	0.05	1
Nb (x_{15})	0	0.275	0.275	6

$$\text{Fe} = 1 - \sum_{i=1}^{15} x_i$$

能力较弱. 这是材料信息学所面临的重大挑战. 缩小特征空间取值范围可以使得到的公式预测性能更好. 缩小特征空间取值范围有两种方式, 一种方式是减小特征的变化范围, 即在不改变特征空间维度的情况下, 减小每一个维度的大小; 另一种方法为减少特征数, 即降低特征空间的维度.

减小特征的变化范围可以采取将元素质量分数差别较大的钢种分开进行回归. 根据图1, 将距离在7以内的数据认为是1簇, 这样整个数据集分为4簇, 即图1中4

种不同颜色的分支. 其中第一簇和第三簇分别包含2组和3组数据点, 数据量太小无法进行回归. 对剩余两簇数据分别进行处理. 本文借鉴Just模型, 引入 $\sqrt{C}$ 特征量. 对于第二簇, 共18组数据, 即编号为3~20钢. 首先将LASSO回归的路径画出, 结果如图2所示, 从图中可以看出各个元素在LASSO回归过程中的重要性排序. 注意图2表示各个元素的系数随着 α 取值变化而变化.

既然元素的系数随 α 的值而变化, 每个特征量出现, 我们在图2中用平行于纵坐标的虚线来表示. 图3所示为留一法交叉验证测试集上的均方根误差和整体均方根误差随着 α 变化的折线图. 图3中每条虚线都与图2中的虚线对应的 α 取值相等. 将第 k 根虚线所对应的 α 作为有 k (有可能大于 k , 因为在有些虚线处出现了多个特征量, 如图2中的第4根虚线就出现了两个特征)个特征量的公式所使用的 α , 如第1根虚线对应的 α 为1个特征量的公式所使用的 α . 从图3可以看出, 虽然随着 α 的减小, 整体均方根误差在不断减小, 在右侧箭头所指向的点之后基本保持不变. 而留一法均方根误差在几处出现极小值, 选取全局最小值对应的 α 值(左侧箭头所指向的点, 即第6根虚线所对应的 α)作为合理的 α 取值, 这时特征量的系数是留一法交叉验证均方根误差全局最小值对应的系数. 第二簇数据中成分的取值范围如表3所示, 且需要满足所有元素质量分数之和为1.

图4是第二簇数据用图3左侧箭头对应的 α 时留一法交叉验证预测值与实际值的散点图. 从图4可以看到有2组数据预测值和测量值偏差较大. 它们分别对应编号为5和6的钢. 可能造成的原因是铬元素的含量较高. 这两组数据铬含量分别为16.5%和18%, 其他大部分数据点的铬含量在7%~16%之间.

第二簇18组数据使用图3中左侧箭头对应的 α 进行LASSO回归, 得到的硬度表达式为

$$H(\text{HRC}) = -0.53\omega_{\text{Cr}} + 1.53\omega_{\text{Mo}} + 2.45\omega_{\text{Si}} + 13.53\omega_{\sqrt{C}} + 1.56\omega_{\text{Ni}} - 0.52\omega_{\text{Mn}} + 46.87,$$

其中, ω 为质量分数, 下标为元素符号, 如 ω_{C} 代表C元素的质量分数. 此公式的均方根误差为1.6615 HRC. LASSO回归确定了特征数目, 然后再使用最小二乘法对LASSO回归得到的特征再次进行回归, 对系数进行优化^[17]. 使用以上6个特征量, 进行最小二乘法回归

图2 (网络版彩图)使用元素质量分数和碳的质量分数的平方根作为特征量,第二簇18组数据的LASSO回归路径。从左到右依次出现的特征量反映了在LASSO回归过程中的重要性排序。由于系数的分散性太大,为了使每一条线都清晰可见,对个别特征的系数进行了一定比例的放大或缩小。主图中铬的系数为实际的4倍,钼的系数为实际的2倍,硅的系数为实际的3倍,根号碳的系数为实际的1/5,镍的系数为实际的2倍,钨的系数为实际的4倍,钴的系数为实际的4倍,铜的系数为实际的1/3,磷的系数为实际的1/13,硫的系数为实际的1/22。左下角小图为实际系数的路径图

图3 (网络版彩图)第二簇18组数据,留一法交叉验证均方根误差和整体均方根误差随 α 的变化图。虚线对应的 α 为新特征出现时的 α ,左侧箭头所指向的点对应的 α 为留一法均方根误差全局最小值对应的 α ,右侧箭头所指向的点对应的 α 为整体均方根误差不再变化时对应的 α

后得到的公式如下:

$$H(\text{HRC}) = -0.40\omega_{\text{Cr}} + 1.36\omega_{\text{Mo}} + 4.96\omega_{\text{Si}} + 14.69\omega_{\sqrt{\text{C}}} + 2.88\omega_{\text{Ni}} - 5.09\omega_{\text{Mn}} + 44.60.$$

此公式的均方根误差为1.5466 HRC.

图1中的第四簇共含56组数据,即编号为19~74的钢。第四簇数据中成分的取值范围如表4所示,且需要满足所有元素质量分数之和为1.此特征空间取值范围与第二簇数据特征空间取值范围的主要差别在于铬元素的质量分数。此组公式适用于预测铬元素质量分数

表3 第二簇数据成分范围

元素	最小值 (wt%)	最大值 (wt%)	步长 (wt%)
C	0.17	2.15	0.005
Mn	0.2	0.75	0.05
P	0	0.015	0.005
S	0	0.015	0.005
Si	0.2	1.05	0.005
Ni	0	2	0.005
Cr	7	18	0.005
Mo	0	2.4	0.005
Co	0	0.5	0.5
V	0	2	0.005
Cu	0	0.15	0.15
W	0	0.7	0.05
Fe		其余	

在0~5.18 wt%的钢种,定义为低铬钢,而第二簇数据的公式适用于预测铬质量分数在7~18 wt%的钢种,定义为高铬钢。

图5表示这56组数据LASSO回归的路径。图5显示了各个元素在LASSO回归过程中的重要性排序。再次强调,每个元素的系数并不是一个常数值,它的值随着 α 的值在变化。图6所示为留一法交叉验证测试集上的均方根误差和整体均方根误差随着 α 变化的折线图。

图 4 (网络版彩图)第二簇18组数据使用图3左侧箭头对应的 α 时,得到的留一法交叉验证的预测值与实际值的散点图.圆点代表预测值和实际值差的绝对值在5 HRC以内,星号代表预测值和实际值差的绝对值在5 HRC以外

表 4 第四簇数据成分范围

元素	最小值 (wt%)	最大值 (wt%)	步长 (wt%)
C	0.27	1.35	0.005
Mn	0.2	2.15	0.05
P	0	0.015	0.005
S	0	0.015	0.005
Si	0.15	2	0.005
Ni	0	1.6	0.005
Cr	0	5.125	0.005
Mo	0	3.1	0.005
Al	0	0.5	0.1
Co	0	0.5	0.5
V	0	1	0.005
Cu	0	0.15	0.15
W	0	3	0.05
Nb	0	0.275	0.275
Fe		其余	

图6中每条虚线都与图5中的虚线对应的 α 取值相等.从图6可以看出,随着 α 的减小,整体均方根误差在不断减小,在右侧的箭头所指向的点之后,整体均方根误差基本保持不变.而在图中左侧箭头所指向的点所对应的 α 时,留一法均方根误差最小.因此,取此时的 α 对应的模型比较合理.

图7表示56组以质量分数作为特征的数据,使用图

6中左侧的箭头所指向的点对应的 α 得到的模型的留一法交叉验证的预测值与实际值散点图.从图7中可以看出,只有一组数据偏差较大.这组数据对应编号为35的钢.此组数据没有非常明显的特征,可能是由于数据本身引起的误差.使用56组数据,左侧箭头所指向的点对应的 α 得到的模型为

图 5 (网络版彩图)使用元素质量分数和碳质量分数的平方根作为特征量,第四簇56组数据的LASSO回归路径.从左到右依次出现的特征量反映了在LASSO回归过程中的重要性排序.由于系数的分散性太大,为了使每一条线都清晰可见,对个别特征的系数进行了一定比例的放大或缩小.主图中铬的系数为实际的25倍,根号碳的系数为实际的1/5,锰的系数为实际的15倍,钨的系数为实际的10倍,镍的系数为实际的15倍,碳的系数为实际的1/3,钼的系数为实际的5倍,钴的系数为实际的10倍,铌的系数为实际的4倍,磷的系数为实际的1/5,钙的系数为实际的1/40.左下角小图为实际系数的路径图

图6 (网络版彩图)第四簇56组数据, 留一法交叉验证均方根误差和整体均方根误差随 α 的变化图。虚线对应的 α 为新特征出现时的 α , 左侧箭头所指向的点对应的 α 为留一法均方根误差全局最小值对应的 α , 右侧箭头所指向的点对应的 α 为整体均方根误差不再变化时对应的 α

图7 (网络版彩图)第四簇56组数据, 图6中左侧箭头所对应的 α 对应的模型的留一法交叉验证的预测值与实际值的散点图。圆点代表预测值和实际值差的绝对值在5 HRC以内, 星号代表预测值和实际值差的绝对值在5 HRC以外

$$H(\text{HRC}) = -0.45\omega_{\text{W}} - 58.73\omega_{\text{C}} + 1.12\omega_{\text{Mo}} - 0.99\omega_{\text{Mn}} + 0.88\omega_{\text{Ni}} - 0.32\omega_{\text{Cr}} + 121.59\omega_{\sqrt{\text{C}}} + 0.24\omega_{\text{Co}} - 0.43\omega_{\text{Si}} - 1.04\omega_{\text{V}} + 16.22\omega_{\text{S}} + 1.74\omega_{\text{Nb}} - 5.05\omega_{\text{Cu}} - 3.69\omega_{\text{Al}} + 0.75,$$

此公式的均方根误差为1.2425 HRC. 使用以上14个特征进行最小二乘法回归, 得到的硬度公式为

$$H(\text{HRC}) = -0.43\omega_{\text{W}} - 61.25\omega_{\text{C}} + 1.08\omega_{\text{Mo}} - 1.01\omega_{\text{Mn}} + 1.00\omega_{\text{Ni}} - 0.33\omega_{\text{Cr}} + 126.09\omega_{\sqrt{\text{C}}} + 0.59\omega_{\text{Co}} - 0.47\omega_{\text{Si}} - 1.11\omega_{\text{V}} + 17.52\omega_{\text{S}} + 2.34\omega_{\text{Nb}} - 6.39\omega_{\text{Cu}} - 4.30\omega_{\text{Al}} + 1.18.$$

此公式的均方根误差为1.2396 HRC.

从图2和5可以看出, 高铬钢中铬是首要作用元素, 而低铬钢中钨是首要作用元素. 在这两类钢中都起到重要作用的元素是碳. 分析其原因在于, 马氏体是钢铁材料中的主要强化相, 而其硬度取决于其含碳量, 因此, 一般而言, 碳含量是影响硬度的主要元素. 不过, 工模具钢基体碳含量的增加会导致马氏体开始转变温度降低, 减弱其最终获得马氏体的能力, 即在同样的工艺条件下, 碳含量的增加有可能因减少组织中的马氏体含量而降低其硬度. 而某些可以显著提高淬透性的元素, 如较高含量铬的加入, 对硬度的影响超过了碳, 造成在高铬钢数据集中铬的影响占了首要地位; 此外, 在低铬钢数据集中, 含钨的数据较多, 钨和碳容易形成碳化钨硬质点从而显著影响硬度, 因此, 钨的影响要高于其他元素.

为了使公式更具有泛化性. 除了元素含量特征外, 还选取关于原子半径和电负性等原子尺度的特征. 使用这些特征量还可以避免由于某些元素出现次数太少, 无法得到其对硬度的影响. 本文共探究17个特征量与硬度之间的关系. 这些特征量为碳的原子百分比(a_{C})、碳的原子百分比的平方根($\sqrt{a_{\text{C}}}$)、相对铁(EN_{Fe})、碳(EN_{C})的电负性, 相对铁(α_{Fe})、碳(α_{C})原子的半径变化率, 相对铁(IP_{Fe})、碳(IP_{C})第一电离能、相对铁(EA_{Fe})、碳(EA_{C})电子亲和能以及元素价电子数(ven_{all}), 定义如下:

$$\text{EN}_{\text{Fe}} = \sum (\text{MEN}_i - \text{MEN}_{\text{Fe}}) \times a_i,$$

$$\text{EN}_{\text{C}} = \sum (\text{MEN}_i - \text{MEN}_{\text{C}}) \times a_i,$$

$$\alpha_{\text{Fe}} = \sum ((R_i - R_{\text{Fe}}) / R_{\text{Fe}}) \times a_i,$$

$$\alpha_{\text{C}} = \sum ((R_i - R_{\text{C}}) / R_{\text{C}}) \times a_i,$$

$$\text{IP}_{\text{Fe}} = \sum (\text{MIP}_i - \text{MIP}_{\text{Fe}}) \times a_i,$$

$$\text{IP}_{\text{C}} = \sum (\text{MIP}_i - \text{MIP}_{\text{C}}) \times a_i,$$

$$\text{EA}_{\text{Fe}} = \sum (\text{MEA}_i - \text{MEA}_{\text{Fe}}) \times a_i,$$

$$\text{EA}_{\text{C}} = \sum (\text{MEA}_i - \text{MEA}_{\text{C}}) \times a_i,$$

$$\text{ven}_{\text{all}} = \sum \text{ven}_i \times a_i,$$

其中, MEN_i 代表第*i*个元素的电负性, MEN_{C} , MEN_{Fe} 分别代表碳元素和铁元素的电负性, R_i 代表第*i*个元素的半径, R_{C} 和 R_{Fe} 分别代表碳元素和铁元素的半径, MIP_i ,

MIP_{Fe} , MIP_C 分别代表第*i*个元素、铁元素和碳元素的第一电离能, MEA_i , MEA_{Fe} , MEA_C 分别代表第*i*个元素、铁元素和碳元素的电子亲和能, ven_i 代表第*i*个元素的价电子数, a_i 代表第*i*个元素的原子百分比. 定义式如下:

$$a_i = \frac{\omega_i / M_i}{\sum_i \left(\frac{\omega_i}{M_i} \right)},$$

其中, ω_i 和 M_i 分别为第*i*种元素的质量分数和摩尔质量. 选择这些特征量是因为铁和碳作为钢铁中最重要的两种元素, 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、碳半径变化率的定义式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; 第一电离能和电子亲和能可以反映某种元素的失电子能力和得电子能力. 这些特征量都可能与硬度有关. 由于电负性有鲍林标度和艾伦标度等, 原子半径有金属半径、共价半径、离子半径等. 因此, 上面的一些特征量又可细分为: 相对铁鲍林电负性(PEN_{Fe})、相对碳鲍林电负性(PEN_C)、相对铁艾伦电负性(AEN_{Fe})、相对碳艾伦电负性(AEN_C)、相对铁共价半径变化率($CR\alpha_{Fe}$)、相对碳共价半径变化率($CR\alpha_C$)、相对铁金属半径变化率($MR\alpha_{Fe}$)、相对碳金属半径变化率($MR\alpha_C$)、相对铁离子半径变化率($IR\alpha_{Fe}$)、相对碳离子半径变化率($IR\alpha_C$). 这些特征量都

是由上面的定义式求出的, 只是选择的标度不同.

首先将这17个特征量的LASSO回归路径画出, 结果如图8所示, 从图中可以看出各个特征在LASSO回归过程中的重要性排序. 图9表示均方根误差随 α 的变化, 随着 α 的减小, 整体均方根误差在不断减小, 在右侧箭头所指向的点之后基本保持不变, 而留一法交叉验证均方根误差在图中左侧箭头所指向的点取得全局最小值, 取此时的 α 作为最终模型的 α 比较合理.

图10是对于全部79种工模具钢, 使用图9中左侧箭头所指向的点对应的 α 进行LASSO回归的留一法交叉验证的预测值与实际值的散点图. 从图10中可以看出有5个点的预测值与实际值差值的绝对值大于5 HRC. 它们分别对应编号为1, 2, 5, 35和72的钢. 其中编号为1, 2的钢为高锰钢, 锰的质量分数在15%以上, 其他编号钢的锰的质量分数均在2.15%以下. 另外3个编号的钢从质量分数和原子尺度角度的分析, 它们与其他数据点之间并无明显差别. 预测误差较大的原因仍在探索中. 使用图9中左侧箭头所指向的点对应的 α , 对79组数据进行LASSO回归得到的硬度公式如下:

$$\begin{aligned} H(\text{HRC}) = & -22.66a_C + 53.59\sqrt{a_C} \\ & + 23.85PEN_{Fe} + 1.34ven_{all} \\ & - 0.02IP_{Fe} + 0.11EA_{Fe} \\ & - 0.02IP_C + 6.82, \end{aligned}$$

其中, a_C 代表碳的原子百分比, $\sqrt{a_C}$ 代表碳原子百分

图8 (网络版彩图)使用原子尺度作为特征量, 全部79组数据的LASSO回归路径. 从左到右依次出现的特征量反映了在LASSO回归过程中的重要性排序. 由于系数的分散性太大, 为了使每一条线都清晰可见, 对个别特征的系数进行了一定比例的放大或缩小. 主图中价电子数(ven_{all})的系数为实际值的6倍, 相对铁第一电离能(MIP_{Fe})的系数为实际值的200倍, 相对铁电子亲和能(MEA_{Fe})的系数为实际的100倍, 相对碳第一电离能(MIP_C)的系数为实际值的400倍. 左下角小图为实际系数的路径图

图9 (网络版彩图)全部79组数据, 留一法交叉验证均方根误差和整体均方根误差随 α 的变化图。虚线对应的 α 为新特征出现时的 α , 左侧箭头所指向的点为留一法均方根误差全局最小值对应的 α , 右侧箭头所指向的点为整体均方根误差不再变化时对应的 α

图10 (网络版彩图)全部79组数据, 图9中左侧箭头所对应的 α 对应的模型的留一法交叉验证的预测值与实际值的散点图。圆点代表预测值和实际值差的绝对值在5 HRC以内, 星号代表预测值和实际值差的绝对值在5 HRC以外

比的平方根、 PEN_{Fe} 代表相对铁鲍林电负性、 ven_{all} 代表价电子数、 IP_{Fe} 代表相对铁第一电离能、 EA_{Fe} 代表相对铁电子亲合能、 IP_C 代表相对碳第一电离能。此公式的均方根误差为2.4846 HRC。

使用以上7个特征量进行最小二乘法回归, 得到的硬度公式为

$$H(\text{HRC}) = -28.24a_C + 64.55\sqrt{a_C} + 26.76PEN_{Fe} + 1.88ven_{all} - 0.03IP_{Fe} + 0.12EA_{Fe} - 0.02IP_C - 1.85.$$

此公式的均方根误差为2.4593HRC。

为了与以元素质量分数作为特征对比, 使用以上17个特征量对第二簇和第四簇数据分别进行LASSO回归。图11为第二簇数据的原子尺度特征, 进行LASSO回归, 均方根误差随 α 的变化图, 箭头所指向的点为留一法均方根误差取全局最小值(2.2382 HRC)所对应的 α (此时对应的整体均方根误差为1.6147 HRC)。图12为第四簇数据的原子尺度特征, 进行LASSO回归, 均方根特征随 α 的变化图。从图中可以看出, 留一法均方根误差并没有出现最小值, 而是稳定在了1.7588 HRC附近, 此时对应的整体均方根误差为1.3923 HRC。

使用原子尺度特征可以预测整个数据集中的数据, 并且可以推广到新元素加入的情况。

在本文所使用的数据集上, Just模型、余柏海模型和金满模型的均方根误差分别为18.2886 HRC, 6.8069 HRC和5.4330 HRC。本文得到的适用于全体数据集的

图11 (网络版彩图)第二簇18组原子尺度数据, 留一法交叉验证均方根误差和整体均方根误差随着 α 的变化图。箭头所指向的点为留一法均方根误差全局最小值对应的 α

图12 (网络版彩图)第四簇56组原子尺度数据, 留一法交叉验证均方根误差和整体均方根误差随着 α 的变化图。箭头所指向的点为留一法均方根误差不再变化时对应的 α

公式的均方根误差为2.4593 HRC, 适用于18组数据的公式的均方根误差为1.5466 HRC, 适用于56组数据的均方根误差为1.2396 HRC, 这些结果均优于前人得到的公式.

4 结论

本文通过层次聚类的方法将整个数据集分为了四簇, 将其中数据量较多的两簇数据使用LASSO回归, 并利用留一法交叉验证确定了最佳的特征量数目. 得

到了两个分别适用于铬质量分数在0~5.125 wt%和铬质量分数在7~18 wt%的工模具钢硬度-合金成分的解析公式.

本文还引入原子尺度的特征, 进行LASSO回归, 并画出17个特征在整个数据集上的LASSO路径, 利用留一法交叉验证确定了最佳的特征量数目. 得到一个适用于本文所有数据的硬度公式. 本文得到的公式在本文所使用的数据集上的预测效果优于前人的模型. 引入原子尺度的这些特征量, 给研究影响钢铁硬度的因素提供了新的思路.

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

- 1 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, 8: 1–10
- 2 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
- 3 Aourag H. Prediction of superlattices ultra hard aluminum rare-earth intermetallic compounds. *Jnl Comp Theo Nano*, 2014, 11: 589–595
- 4 Dey S, Dey P, Datta S. Design of novel age-hardenable aluminium alloy using evolutionary computation. *J Alloys Compd*, 2017, 704: 373–381
- 5 Dey S, Sultana N, Dey P, et al. Intelligent design optimization of age-hardenable Al alloys. *Comput Mater Sci*, 2018, 153: 315–325
- 6 陈再枝, 蓝德年. 模具钢手册. 北京: 冶金工业出版社, 2002
- 7 刘波. GB/T1299-2014《工模具钢》标准浅析. *工具技术*, 2015, 49: 82–86
- 8 戴强. 我国高档模具钢标准研制情况综述. *世界金属导报*, 2017, B11: 1–3
- 9 Just E. New formulas for calculating hardenability curves. *Metal Progress*, 1969, 95: 87–88
- 10 余柏海. 计算淬透性及机械性能的非线性方程. *钢铁*, 1985, 3: 40–49
- 11 金满, 连建设, 江中浩. 描述钢淬透性的一个新数学模型. *金属学报*, 2006, 3: 265–272
- 12 金满, 连建设, 江中浩. 结构钢端淬曲线预测新方法. *金属学报*, 2006, 4: 405–410
- 13 Vermeulen W G, van der Wolk P J, de Weijer A P, et al. Prediction of Jominy hardness profiles of steels using artificial neural networks. *J Mater Eng Perform*, 1996, 5: 57–63
- 14 Filetin T, Majetić D, Žmak I. Prediction the Jominy curves by means of neural network. In: *Proceedings of the 4th ASM Heat Treatment and Surface Engineering Conference in Europe*. Florence, 1998. 353–361
- 15 由伟, 赵玮玮, 赖惠先. 用人工神经网络研究钢的硬度的影响因素. *钢铁研究学报*, 2013, 1: 34–38
- 16 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 工模具钢(GB/T1299-2014)
- 17 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer, 2009

Machine learning prediction of the hardness of tool and mold steels

WANG JiaHao¹, SUN Sheng¹, HE YanLin² & ZHANG TongYi¹

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

² *School of Materials Science & Engineering, Shanghai University, Shanghai 200444, China*

Hardness is a major indicator of the quality of tool and mold steels (TMS). Analytic formulas of hardness versus compositions were proposed here by using hierarchical clustering (HC) and LASSO regressions, based on data of 79 brands of TMS. HC presents two large groups of TMS, which contain high and low concentrations of chromium. Then, LASSO regressions were applied on each group and the regression result with the lowest root mean square error in leave-one-out cross-validation was taken out as the theoretical prediction formula of hardness versus chemical composition feature. Furthermore, atomic features were selected from electronegativity, atomic radius, valence electron number, electron affinity, first ionization energy, etc. LASSO regressions of the data with the atomic features give another prediction formula. These results demonstrate the powerful ability of machine learning in the design of TMS.

machine learning, tool and mold steels, hardness prediction

doi: [10.1360/SST-2019-0060](https://doi.org/10.1360/SST-2019-0060)