

基于机器学习与多尺度特征的 金属氧化物性能预测

李 想¹ 曹 卓¹ 但雅波¹ 牛程程² 胡建军^{1,3*}

(1.贵州大学机械工程学院,贵阳 550025;2.贵州大学现代制造技术教育部重点实验室,贵阳 550025;
3.美国南卡罗来纳大学计算机科学与工程系,哥伦比亚 29208,美国)

摘 要 材料信息学是信息技术在材料学中的应用,通过材料信息数据库和集成材料设计平台对材料的数据进行分析和预测。通过应用不同的机器学习(回归分析)方法和不同的特征选择算法,从众多的多尺度特征集中选择最优的特征子集可以预测金属氧化物的物理特性,归纳出适合材料不同特性的机器学习模型。分析结果表明,特征选择方法可以提升机器学习模型的性能,为进一步开发更有效的材料性能预测方法提供参考。

关键词 材料信息学,机器学习,多尺度特征,特征选择,材料性能预测

Prediction of metal oxide performance based on machine learning and multi-scale feature

Li Xiang¹ Cao Zhuo¹ Dan Yabo¹ Niu Chengcheng² Hu Jianjun^{1,3}

(1.School of Mechanical Engineering,Guizhou University,Guiyang 550025;2.Key Laboratory of Modern Manufacturing Technology,Ministry of Education,Guizhou University, Guiyang 550025;3.Department of Computer Science and Engineering,University of South Carolina,Columbia 29208,USA)

Abstract Material informatics is the application of information technology in material science,and the material data is analyzed and predicted by material information database and integrated material design platform.The main research was to predict the metal oxide physical properties by using different machine learning (regression analysis) methods and different feature selection algorithms to select the optimal feature subset from a large number of multi-scale feature sets.The most effective machine learning algorithm model was concluded for every material characteristic. At the same time,the performance improvement effect of feature selection algorithm was got for different machine learning models.It provided a reference for further development of more effective material performance prediction methods.

Key words material informatics,machine learning,multi-scale feature,feature selection,material performance prediction

近年来,实验和模拟方法是新材料开发和材料制造技术研究的主要途径,这些方法不仅工作量而且十分耗时。目前,材料科学研究者的首要任务是提高材料研发效率并缩短研发周期。2011年,美国提出了材料基因组计划^[1](MGI),其目标是“先进材料的发现、开发、制造和使用的速度提高一倍”。实现这个宏伟的目标需要满足以下3个条件^[2]:

(1)从理论层面理解物理机制与材料结构和性能的关系;(2)计算机的计算能力;(3)数据库及有效的筛选方法。

要实现材料基因组计划,计算机技术的支持必不可少。目前在生命科学和化学研究等领域利用计算机对材料信息进行采集、存储、分类和搜索已经取得了显著效果^[2]。随着材料数据库的不断累积,近

基金项目:国家自然科学基金项目(51741101)

作者简介:李想(1994-),男,硕士研究生,研究方向为材料信息学。

联系人:胡建军(1973-),男,博士,教授,从事机器学习、深度学习及材料信息学研究。

年来材料信息学得到飞速发展^[3]。利用机器学习和材料大数据以及基于第一性原理密度泛函理论(DFT)的仿真方法^[4-6]、各种材料的特性预测方法不断涌现。Meredig等^[7]利用现有数据库,通过DFT训练得到正向模型并用于预测材料性能。Takahashi等^[8]利用DFT中的投影仪增强波(GPAW)方法创建了材料数据库并预测金属化合物的性能,通过预测得到的点阵常数与实验数据基本相同。Fischer课题组^[9-10]将第一性原理与信息学技术相结合来预测晶体结构;Calfa等^[11]利用机器学习和回归模型对材料的电子特性、弹性性能进行预测,并提出了材料反向设计的详尽枚举算法(EE);Hart^[12]开发了一种预测晶体结构的基因算法。这些方法已被证明可以有效预测材料的某些性能,预测材料性能时这些方法大多以指定的特征进行预测且使用的机器学习模型较为单一。

金属氧化物的物理特性(密度、原子平均能量、费米能、带隙等^[11])与原子的大小、原子质量、原子数量等相关。这些多尺度的材料特征与所要预测的材料物理特性之间的关系错综复杂。本研究提出利用机器学习与多尺度的特征对金属氧化物的物理特性进行预测,对比了不同的特征选择方法在3种机器学习模型中的表现,归纳出最优组合模型。

1 实验方法

通过应用不同的机器学习(回归分析)方法和不同的特征选择方法,从多尺度的特征集中选择最优的特征子集来预测金属氧化物的物理特性,如密度、原子平均能量、费米能、带隙。

1.1 数据集的准备

实验使用的数据来源于材料工程(MP)数据库^[13],MP数据库是材料基因组的相关数据库,旨在通过高通量计算方法发现新材料。使用基于表述性状态转移(REST)^[14]的材料应用程序接口(MA-PI)获取数据,得到746个二元金属氧化物;然后再经过处理,剔除分子式相同而空间结构不同的二元金属氧化物,最终得到330个二元金属氧化物。Magpie是一个可扩展的平台,用于计算材料属性。本研究使用一组145个属性。这些属性分为4个类别^[15],即化学计量属性、元素属性统计、电子结构属性和离子化合物属性。

1.2 特征选择方法

机器学习算法中,特征选择对预测结果具有重

要作用,选择的特征不同对预测结果的影响也不同。将数据集按一定的比例划分为训练集和测试集,结合数据集数量较少这一特点,将数据集随机分配。从不同侧重点选择,能够找到更适合的特征选择方法,得到更好的实验结果。采用去除低方差特征选择(VT)、单变量特征选择(UFS)及嵌入式特征选择与L1正则化(LASSO)这3种特征选择方法进行实验。

1.2.1 VT方法

VT是较简单的特征选择方法,它删除所有差异不符合某个阈值的特征。默认条件下,它将删除所有零方差特征,即所有样本中具有相同值的特征。假设有1个具有布尔特征的数据集,并且希望在80%以上的样本中删除1个或零个(0或1)的所有特征。布尔特征是伯努利随机变量,这些变量方差的计算方法见式(1):

$$\text{Var}[x] = p(1-p) \quad (1)$$

式中, $\text{Var}[x]$ 为方差; p 为布尔系数。为此实验选择使用阈值 $8 \times (1-0.8)$ 。

1.2.2 UFS方法

UFS主要通过计算单个特征与目标值之间的相关性进行特征选择^[16],对每个特征进行测试,衡量该特征与响应变量之间的关系,剔除不好的特征。UFS方法在数据清洗和特征提取之后使用,且计算简单、速度快、较易实现。UFS方法有助于对数据的理解,但对特征优化、提高泛化能力不一定有效。

1.2.3 LASSO方法^[17]

嵌入式特征选择结合了特征选择过程与学习器训练过程,同时对两方面进行优化。若给定的数据集为 $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_m, y_m)\}$,其中 $x \in R^d$ (x 为向量, R^d 为域的维度), $y \in R$ (y 为输出目标值, R 为域),则优化目标为:

$$\min_w \sum_{i=1}^m (y_i - w^T x_i)^2 \quad (2)$$

式中, m 为样本数量; w 表示向量; w^T 为 w 的转置。

当样本特征很多,而样本数量相对较少时,式(2)很容易陷入过拟合。为解决这个问题,对式(2)引入正则化项,使用 L_1 范数正则化,则有:

$$\min_w \sum_{i=1}^m (y_i - w^T x_i)^2 + \lambda \|w\|_1 \quad (3)$$

式中, λ 为正则化参数;当 $\lambda > 0$,式(3)称为套索算法^[18]。

1.3 回归模型

1.3.1 线性回归模型

在机器学习中虽然线性回归(LR)模型建模容易、模型简单^[19],但却有着重要的地位。例如数据集为 $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_m, y_m)\}$, 其中 $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{id})$, $y_i \in R$, 其中 m 为样本数量, d 为特征维数。通过线性回归试图学得到公式(4):

$$f(x_i) = wx_i + b \quad (4)$$

$$f(x_i) \cong y_i \quad (5)$$

式中, b 为偏置向量; $w = (w_1, w_2, \dots, w_d)$ 。由式(4)可以看出 w_i 每个特征在预测中的权重。对于 w, b 的求解, 通过使均方误差最小化及使用最小二乘法得到:

$$(w^*, b^*) = \min_{(w, b)} \sum_{i=1}^m (y_i - wx_i - b) \quad (6)$$

式中, w^*, b^* 分别为 w, b 的解。由此可得 w 和 b 的最优解。

1.3.2 随机森林回归模型

随机森林回归(RFR)模型最早由 Breiman^[20]提出, 该模型通过与随机向量 θ 有关的回归树生长构成。其基本思想源于统计学理论, 利用 bootstrap 在训练集中随机抽取 N 个训练样本, 每个训练样本假设能构造决策树, 将全部决策树预测值加起来取平均值作为最终预测结果。若训练集从随机向量 (X, Y) 中抽取, 设 X 为输入向量, Y 为输出向量, 那么预测值 $h(X)$ 的均方泛化误差为:

$$E_{X,Y}[Y - h(X)]^2 \quad (7)$$

式中, $E_{X,Y}$ 为均方泛化误差。

RFR 预测值是对 k 颗决策树取平均值, 步骤如下^[21-22]:

步骤 1: 通过 bootstrap 方法采样, 生成 k 个训练集 $(\theta_1, \theta_2, \theta_3, \dots, \theta_k)$; 然后生成对应的决策树。

步骤 2: 假设特征有 M 维, 随机抽取 m 个特征作为当前节点的分裂特征集, 并对该节点进行分裂 (m 值不变); 每棵决策树都以最大限度生长, 并且没有剪枝过程。

步骤 3: 单棵决策树的预测值由因变量 Y_i ($i = 1, 2, 3, \dots, n$) 的加权平均计算得到, 计算公式如下:

$$I_\epsilon = \begin{cases} 0, & |f(x) - y| \leq \epsilon \\ |f(x) - y| - \epsilon, & \text{其他} \end{cases} \quad (8)$$

式中, I_ϵ 为损失函数; ϵ 为特定值。

步骤 4: 通过式(9)对决策树权重 $w_i(x, \theta_i)$ 取平均值得到每个观测值 $Y_i \in (1, 2, \dots, n)$ 的权重

$w_i(x)$:

$$w_i(x) = \frac{1}{k} \sum_{i=1}^n w_i(x, \theta_i) Y_i \quad (9)$$

因此 RFR 的预测值为:

$$\hat{u}(x) = \sum_{i=1}^n w_i(x) Y_i \quad (10)$$

式中, $\hat{u}(x)$ 为预测值。

1.3.3 支持向量机回归模型

支持向量机回归(SVR)模型的优势是解决小样本数据集, 其基本思路是寻找一个超平面, 把不同类别的样本分离。假设目前考虑解决回归的问题, 给出的训练样本为 $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_m, y_m)\}$, $y_i \in R$ 。期望得到一个形如 $f(x) = w^T x + b$ 的回归模型, 使得 $f(x)$ 与 y 尽可能接近。当且仅当 $f(x)$ 与 y 完全相同时, 损失才为零。假设能容忍 $f(x)$ 与 y 之间最多有 ϵ 的偏差, 即偏差绝对值大于 ϵ 时才计算损失。此时, 以 $f(x)$ 为中心, 构建一个宽度为 2ϵ 的间隔带。则 SVR 问题可化为:

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m [f(x_i) - y_i] \quad (11)$$

$$I_\epsilon = \begin{cases} 0, & |f(x) - y| \leq \epsilon \\ |f(x) - y| - \epsilon, & \text{其他} \end{cases} \quad (12)$$

式中, C 为正则化常数。

2 结果与讨论

2.1 预测性能指标

实验采用 K 折交叉验证法, 它是一种常用的数据建模和检验分析技术, 可以打乱并重复利用原始样本, 利用有限的样本来减少预测偏差, 并兼顾训练误差和泛化误差。对于小样本数据, 此方法的检验结果比将原始样本分为训练集和测试集的结果更为可靠, 因此取 $K = 10$ (把数据分成 K 份, 每次拿出一份作为验证集, 剩下 $k - 1$ 份作为训练集)。采用平均绝对百分比误差(MAPE)、均方根误差(RMSE)、平均绝对误差(MAE)作为学习器性能评价指标, 以相关系数 R^2 作为预测值与实际值吻合程度的衡量指标。

MAPE 计算公式如下:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \bar{y}_i}{y_i} \right| \quad (13)$$

式中, n 表示样本数量。

RMSE 计算公式如下:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (14)$$

MAE 计算公式如下:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}_i| \quad (15)$$

相关系数 R^2 计算公式如下:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (16)$$

式中, $\hat{y}_i$ 表示 y_i 的预测值。相关系数 R^2 越接近 1, 说明吻合程度越高。

2.2 模型预测结果

为了选择适宜的机器学习算法, 将 3 种机器学习模型与 3 种特征选择方法进行组合, 通过对比筛选出最佳组合模型。将 VT 方法与 3 种机器学习模型(LR, SVR, RFR)组合, 对二元金属氧化物的物理性能进行预测, 预测结果见表 1。

表 1 基于 VT 方法的组合模型预测的金属氧化物物理性能

密度	RMSE	MAE	MAPE
LR	0.80	0.56	0.09
SVR	0.85	0.57	0.09
RFR	0.75	0.55	0.09
平均原子能量	RMSE	MAE	MAPE
LR	0.51	0.34	0.32
SVR	0.54	0.33	0.46
RFR	0.47	0.30	0.31
带隙	RMSE	MAE	MAPE
LR	1.02	0.70	—
SVR	0.88	0.59	—
RFR	0.62	0.52	—
费米能	RMSE	MAE	MAPE
LR	1.67	1.23	1.28
SVR	1.53	1.14	0.58
RFR	1.51	1.08	0.57

注: “—”表示数据集中至少 1 个属性的响应数据点为零, 因此不计算 MAPE。

由表 1 可知: 无论是 RMSE、MAE 还是 MAPE 误差, 金属氧化物物理性能衡量指标在测试集上的大小是按模型排序的。密度的排序结果如下: RMSE 中 $SVR > LR > RFR$; MAE 中 $SVR > LR > RFR$; MAPE 中 $SVR = LR = RFR$ 。平均原子能量的排序结果如下: RMSE 中 $SVR > LR > RFR$; MAE 中 $LR > SVR > RFR$; MAPE 中 $SVR > LR > RFR$ 。带隙的排序结果如下: RMSE 中 $LR > SVR > RFR$; MAE 中 $LR > SVR > RFR$ 。费米能的

排序结果如下: RMSE 中 $LR > SVR > RFR$; MAE 中 $LR > SVR > RFR$; MAPE 中 $LR > SVR > RFR$ 。综合分析表明, RFR 模型表现优异, 而 LR 和 SVR 在不同方面交替领先。

另外, 单独使用 RMSE、MAE、MAPE 对模型进行评估不够全面, 因此将 3 种机器学习模型(LR, SVR, RFR)与 3 种特征选择方法(VT, UFS, LASSO)两两组合, 共组合成 9 种模型。利用 9 种组合模型对金属氧化物的相关系数 R^2 进行预测, 结果见表 2。

表 2 利用 9 种组合模型预测的金属氧化物的相关系数 R^2

密度	VT	UFS	LASSO
LR	0.92	0.91	0.78
SVR	0.91	0.91	0.86
RFR	0.91	0.92	0.92
平均原子能量	VT	UFS	LASSO
LR	0.79	0.78	0.61
SVR	0.77	0.80	0.86
RFR	0.82	0.88	0.88
费米能	VT	UFS	LASSO
LR	0.67	0.71	0.52
SVR	0.72	0.70	0.71
RFR	0.73	0.80	0.82

由表 2 可知, RFR 模型与 LASSO 方法是最佳组合, 该组合模型预测的金属氧化物物理性能(密度、平均原子能量、费米能)的相关系数 R^2 分别为 0.92, 0.88, 0.82, 预测效果较好。

基于 LASSO 方法, 利用 LR、SVR、RFR 模型对金属氧化物的费米能进行预测, 预测结果见图 1。由图可以看出, 吻合程度最高的是 RFR 模型, 其次是 SVR 模型, 较差的是 LR 模型。

3 结论

使用 3 种特征选择方法(VT、UFS 和 LASSO)和 3 种机器学习模型(LR、RFR 和 SVR)对 330 种金属氧化物的性能进行了预测。通过对比同一特征选择方法与不同回归模型组合、不同特征选择方法与不同回归模型组合的预测结果, 发现基于同一特征选择方法的不同回归模型预测的金属氧化物的物理性能差异较大, 而基于同一回归模型的不同特征选择方法差异较小, 其中 SVR 模型预测的金属氧化物相关系数 R^2 在 ± 0.05 范围波动, 较稳定; RFR

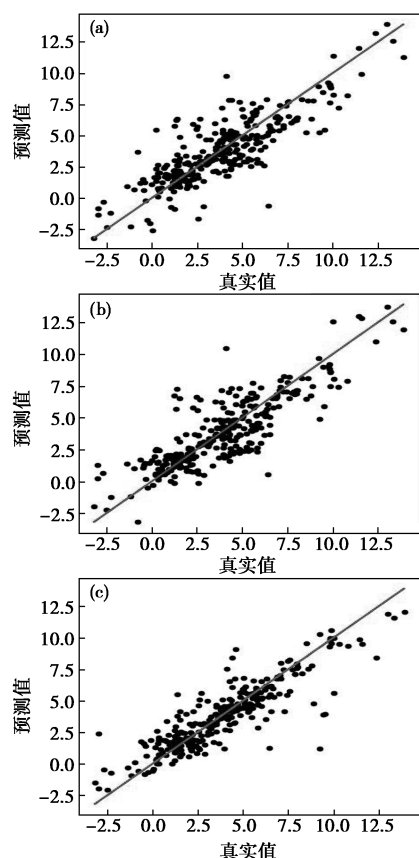


图 1 基于 LASSO 方法的 3 种回归模型预测的
费米能的真实值-预测值
[(a)LR;(b)SVR;(c)RFR]

模型与不同特征选择方法组合的模型在 RSME、MAE、MAPE 和 R^2 预测方面都取得较好结果。

参考文献

- [1] 赵继成.材料基因组计划简介[J].自然杂志,2014,36(2):89-104.
- [2] 王卓,王礞,雍歧龙,等.材料信息学及其在材料研究中的应用[J].中国材料进展,2017,36(2):132-140.
- [3] 宋庆功,姜恩永,韩志勇,等.材料信息学平台的研究现状发展趋势与作用[J].材料导报,2008,22(1):70-73.
- [4] Majzoub E H, Ozoliņš V. Prototype electrostatic ground state approach to predicting crystal structures of ionic compounds: application to hydrogen storage materials[J]. Physical Review B, 2008, 77(10):104115.
- [5] Wolverton C, Ozoliņš V, Asta M. Hydrogen in aluminum: first-principles calculations of structure and thermodynamics[J]. Physical Review B, 2016, 69(14):1124-1133.
- [6] Wolverton C, Ozoliņš V. Hydrogen storage in calcium alanate: first-principles thermodynamics and crystal structures[J]. Physical Review B, 2007, 75(6):519-542.
- [7] Meredig B, Agrawal A, Kirkin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning[J]. Physical Review B, 2014, 89(9):82-84.
- [8] Takahashi K, Tanaka Y. Material synthesis and design from first principle calculations and machine learning[J]. Computational Materials Science, 2016, 112:364-367.
- [9] Fischer C C, Tibbetts K J, Morgan D, et al. Predicting crystal structure by merging data mining with quantum mechanics[J]. Nature Materials, 2006, 5(8):641-646.
- [10] Ceder G, Morgan D, Fischer C, et al. Data-mining-driven quantum mechanics for the prediction of structure[J]. Mrs Bulletin, 2006, 31(12):981-985.
- [11] Calfa B A, Kitchin J R. Property prediction of crystalline solids from composition and crystal structure[J]. Aiche Journal, 2016, 62(8):2605-2613.
- [12] Hart G L W. Evolutionary approach for determining first-principles model Hamiltonians[J]. Nature Materials, 2005, 4(5):391-394.
- [13] Jain A, Ong S P, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation[J]. APL Materials, 2013, 1(1):1049.
- [14] Ong S P, Cholia S, Jain A, et al. The materials application programming interface (API): a simple, flexible and efficient API for materials data based on representational state transfer (REST) principles[J]. Computational Materials Science, 2015, 97(5):209-215.
- [15] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials[J]. 2016, 2:16028.
- [16] 王俊.基于原型方法的特征选择算法的研究[D].哈尔滨:哈尔滨工业大学,2009.
- [17] Xiong Kai, Han Junwei. A review of regularization based feature selection algorithms[EB/OL]. Beijing: Sciencepaper Online[2017-04-27]. <http://www.paper.edu.cn/releasepaper/content/201704-704>.
- [18] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of the Royal Statistical Society, 2011, 73(3):273-282.
- [19] Explained P M D. Introduction to linear regression analysis[J]. Technometrics, 2013, 49(2):232-233.
- [20] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.
- [21] 方匡南.随机森林组合预测理论及其在金融中的应用[M].厦门:厦门大学出版社,2012.
- [22] 周志华.机器学习[M].北京:清华大学出版社,2016:178-181.

收稿日期:2018-05-04

修稿日期:2019-05-19