第 29 卷第 4 期 中国有色金属学报 2019 年 4 月
Volume 29 Number 4 The Chinese Journal of Nonferrous Metals April 2019
DOI：10.19476/j.ysxb.1004.0609.2019.04.17

基于不同机器学习算法的钙钛矿材料性能预测

郑伟达[1,3]，张惠然[1,2]，胡红青[2]，刘 尧[2]，李盛洲[2]，丁广太[1,2]，张金仓[1,3]

(1. 上海大学 材料基因组工程研究院，上海 200444；
2. 上海大学 计算工程与科学学院，上海 200444；
3. 上海大学 理学院，上海 200444)

摘 要：钙钛矿材料由于在各领域具有广泛的应用前景而备受材料学家的关注，对其各种物理化学性能的研究一直是材料领域研究的热点。本文建立随机森林(Random forest，RF)、岭回归(Ridge regression，RR)、以及基于径向基核函数和线性核函数的支持向量回归(Support vector regression，SVR)等 4 种机器学习算法的预测模型，对钙钛矿材料数据集中的密度、形成能、带隙、晶体体积等 4 种性能参数进行预测。结果表明：RF 方法可以对钙钛矿材料的密度、带隙性能进行有效预测；RR 方法可以实现对密度性能的预测；线性核函数的 SVR 方法可以实现对形成能性能的预测。该研究表明，不同的机器学习算法对数据样本分布的敏感程度不同，因此针对不同的性能参数预测需要选择不同方法。

关键词：钙钛矿材料；机器学习；性能预测；算法选择
文章编号：1004-0609(2019)-04-0803-07 **中图分类号**：TB34 **文献标志码**：A

钙钛矿材料是指与钛酸钙($CaTiO_3$)具有相似晶体结构的一大类化合物，常用通式为 ABX_3。该类材料具有许多特殊的物理化学性能，使得其一直以来在材料学研究领域备受关注。如铁电材料($BaTiO_3$)[1]、热电材料($LaCoO_3$)[2]、介电材料($PbZrO_3$)[3]、超导材料($YBa_2Cu_3O_7$)[4]、庞磁电阻材料($SrMnO_3$)[5]以及近几年获得巨大关注与发展的有机无机杂化钙钛矿太阳能电池材料($CH_3NH_3PbI_3$)[6]，这些材料都被称为钙钛矿材料。其中由于钙钛矿太阳能电池制备方法简单，转换效率优势较为明显，大家公认其最终将取代硅基太阳能电池的统治地位，这也使得钙钛矿材料迅速成为新能源领域的一颗巨星。

材料学领域的研究一般是基于成功制备实验样品的基础上，对样品进行各种物性测量从而了解它的各种物理性质，并通过不同的性能参数对材料进行分析和分类应用。但是，实验研究对实验样品具有很大的依赖性，实验过程往往需要大量重复的繁杂工作，甚至做很多无用功。随着理论的积累和计算机技术的发展，第一性原理计算成为获得材料理论性能参数的新途径[7−9]，但其耗时过长。因此无论采用哪种方法，对材料性能的发现和总结都会耗费大量的人力、物力和财力。随着人工智能的发展，不少研究者提出将机器学习方法逐渐在材料学科上推广应用[10−11]。对目前所产生的大量材料数据集使用不同的机器学习方法，进行材料性能参数的预测[12−13]，可以有效提高材料性能预测准确率，从而筛选出合理性能的材料再进行实验研究。利用已有数据对性能参数进行预测，不仅能扩充材料数据的数据量，同时还能为材料实验和应用提供指导。同时，对于机器学习方法而言，不同算法在不同范围的数据集上对材料数据的敏感度不同[14−17]，需要根据特定的材料数据样本对算法进行有针对性的选择，然后通过相应的性能评估手段来评价一个算法的优劣。

选取 JAIN 等[18]基于第一性原理和密度泛函理论(Density functional theory，DFT)计算得到的钙钛矿材料数据集为范本，加入 A、B 位原子的电负性(Electronegativity of atom A; Electronegativity of atom B)和 A、B、X 位原子的有效原子半径(Effective Radius of atom A; Effective radius of atom B; Effective radius of atom X)等 5 个特征性能参数，采用不同的机器学

基金项目：国家重点研发计划资助项目(2016YFB0700502)
收稿日期：2018-03-19；**修订日期**：2018-05-31
通信作者：张惠然，讲师，博士；电话：13621855760；E-mail：hrzhangsh@shu.edu.cn

习方法，对原始材料数据集中部分缺失的形成能(Formation energy)、密度(Density)、带隙(Band gap)和晶体体积(Volume)这 4 个重要的特征性能参数进行预测。如图 1 所示，将原始数据集分为训练集和测试集，通过训练集对多算法模型进行训练学习，然后将利用训练后的多算法模型对测试集进行预测，得到回归拟合的结果。分别采用随机森林(RF)[19]、岭回归(RR)[20−21]、以及径向基核函数(SVR-RBF)和线性核函数(SVR-Linear)的支持向量回归[22−23] 4 种机器学习算法，采用交叉验证的方式建立针对钙钛矿体系材料多种性能参数的多算法预测模型，同时对模型的预测精度进行比较和评估。实验结果对进一步研究机器学习方法在钙钛矿体系材料性能预测乃至新材料体系发现方面具有重要的参考价值和实用意义。

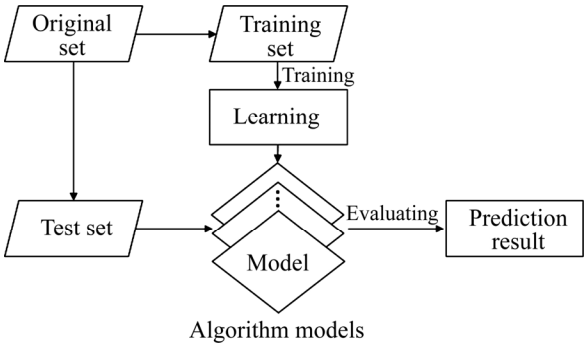

图 1 机器学习的性能预测模型工作流程图

Fig. 1 Workflow of performance prediction model via machine learning

1 原理与方法

1.1 随机森林(RF)

RF 是由 BREIMAN[19]于 2001 年提出的由决策树组合成的算法，顾名思义就是用随机的方式建立一个由很多决策树组成的森林，每棵决策树之间没有关联。RF 是指在变量和数据的使用上进行随机化，产生很多决策树再将之汇总的结果，主要应用于回归和分类。以下是 RF 的构造过程：

步骤(1) 设有 N 个样本，有放回的随机选择 n 个样本(每次随机选择一个样本，然后返回继续选择)。选择好的 N 个样本用来训练一个决策树，作为决策树节点处样本。

步骤(2) 当每个样本有 M 个性能时，决策树的每个节点需要分裂，随机从这 M 个性能中选出 m 个性能，满足 $m \ll M$。然后从这 m 个性能中采用某种策略(例如信息增益)来选择 1 个性能作为该节点的分裂性能。

步骤(3) 决策树形成过程中每个节点都按照步骤(2)来分裂，一直到不能再分裂为止。整个决策树形成过程中没有进行剪枝。

步骤(4) 按照步骤(1)~(3)建立大量的决策树，即构成 RF。

1.2 岭回归(RR)

在监督学习中，对于连续性的目标变量的分析预测，一般采用回归模型(Regression model)，简单的回归模型为简单线性回归：通过描述特征(特征变量 x_i)与连续输出(目标变量 y)之间的关系。线性模型的函数定义为：

$$y = w_0 + \sum w_i x_i \tag{1}$$

式中：w_0 为权值；w_i 为特征变量的系数。然后，通过最小二乘法(Ordinary least squares, OLS)估计回归曲线的参数，使得回归曲线到样本点垂直距离(残差或误差)的平方最小。然而，对于普通最小二乘的系数估计问题，其依赖于模型各项的相互独立性。当各项是相关的，会导致最小二乘估计对于随机误差非常敏感，产生很大的方差。所以，通过对最小二乘法正则化，使回归的误差尽量减小并有效地缓解过拟合的问题，本文选用了 L2 正则方法的 RR。

1.3 支持向量回归(SVR)

SVR 是运用支持向量机(Support vector machine，SVM)来解决回归问题的方法。SVM 方法是建立在统计学习理论基础上的，根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力(或称泛化能力)。在解决小样本、非线性及高维模式识别中支持向量回归表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

SVR 的基本思想是通过一个非线性映射 ϕ，将数据 x 映射到高维特征空间 F，并在这个空间进行线性回归。假设一个样本集 $\{(x_i, y_i)\}_i^N$，其中输入数据 $x_i \in R^n$，$y_i \in R$，在高维空间中构造的最优线性模型函数为：

$$f(x) = \omega^{\mathrm{T}} \phi(x) + b \tag{2}$$

式中：ω 是权重；b 为偏置项。这样，在高维特征空

间的线性回归便对应于低维输入空间的非线性回归。

利用 SVM 解决回归问题时，需要根据求解问题的特性，通过使用恰当的核函数来代替内积，以便隐式地把高维特征空间的点积运算转化为低维原始空间的核函数运算，巧妙地解决在高维特征空间中计算带来的“维数灾难”，从而解决计算上的技术问题。这个核函数不仅要在理论上要满足 Mercer 条件，而且在实际应用中要能够反映训练样本数据的分布特性。因此，在使用支持向量机解决某一特定的回归问题时，选择适当的核函数是一个关键因素。核函数的种类很多，常用的有 RBF、Linear 等核函数。

此外，SVR 引入一个以 ε 为参数的不敏感损失函数作为损失函数，通过采用损失函数在高维特征空间完成线性回归，同时通过最小化 $\|\omega\|^2$ 来减少模型的复杂度。常用的损失函数有：线性 ε 不敏感损失函数；二次 ε 不敏感损失函数；Huber 损失函数。而参数 ε 用来度量 ε −不敏感带外的训练样本的偏离程度，ε 取值大小影响支持向量的数目。最终，支持向量回归的优化的目标函数为

$$\min_{\omega,b,\xi,\xi'} J(\omega,\xi,\xi') = \frac{1}{2}\omega^{\mathrm{T}}\omega + C\sum_{i=1}^{n}(\xi+\xi') \tag{3}$$

这里引入非负的松驰变量 ξ 和 ξ'；C 为正则化参数，控制对超出误差的样本的惩罚程度。

使用支持向量回归构建模型，首先我们明确模型的输入与输出，然后选择合适的核函数，损失函数以及调整参数如核参数 σ，正则项 C，不敏感损失参数 ε，根据样本训练集得到训练模型。

1.4　性能评估

平均绝对误差 E_{MAE}(Mean absolute error，MAE)、均方根误差 E_{RMSE}(Root mean squared error，RMSE)和 R^2 值被用来作为泛化性能评估。它们的公式如下：

$$E_{\mathrm{MAE}} = \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - y_j\right| \tag{4}$$

$$E_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{y}_j - y_j)^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{j=0}^{n-1}(y_j - \hat{y}_j)^2}{\sum_{j=0}^{n-1}(y_j - \bar{y}_j)^2} \tag{6}$$

式中：n 为样本数量；$\hat{y}_j$ 为真实值；y_j 为预测值。

2　多算法模型的建立

材料的特征性能参数之间有着千丝万缕的内在联系，同时这些性能也会相互影响，这种内在联系使得部分缺失的特征性能能够通过其他完整的特征性能得到预测。本文选取来自 Materials Project 数据库的 102 条钙钛矿材料的数据集，对原始材料数据集中的形成能、密度、带隙和晶体体积 4 个特征性能参数进行预测。为了进一步提高钙钛矿材料性能预测的精确度，除了原始数据集的晶胞原子数(Nsites)以及原子核外能(Energy above hull)外，加入 A、B 位原子的电负性和 A、B、X 位原子的有效原子半径 5 个特征性能，从而获得标准数据集，数据集的特征性能参数如表 1 所列。本次实验采用的 11 个性能参数为钙钛矿材料本身的特征性能，其中 1 个预测性能为因变量，其他 10 个性能则为自变量。

表 1　钙钛矿数据集的特征性能参数

Table 1　Feature property parameters of perovskite data sets

No.	Prediction property	Adding property
1	Formation energy	Electronegativity of atom A
2	Density	Electronegativity of atom B
3	Band gap	Effective radius of atom A
4	Volume	Effective radius of atom B
5	–	Effective radius of atom X
6	–	Energy above hull
7	–	Nsites

根据获得的数据集和预测变量，基于机器学习的钙钛矿材料性能预测模型构建方法如下。

1) 数据准备：将 102 条数据集随机分成 93 条训练集和 9 条测试集。

2) 模型训练：设置 10 折交叉检验，分别建立 RF、SVR-RBF、SVR-Linear、RR 算法的模型。利用上述模型分别针对训练集的形成能、带隙、密度、晶体体积这 4 个性能进行独立训练。

3) 模型效果评估：使用 E_{MAE}、E_{RMSE}、R^2 评价指标对模型效果进行评估。

4) 模型应用：利用训练后的多个算法模型对测试集的 4 个性能分别进行独立预测，并作简单评估。

为了提高模型预测精度，在模型训练之前通常要对算法进行参数寻优。这里以形成能属性为例，进行参数寻优。对于 RF 算法，通过迭代选择最佳最大决

策树个数为 40。由于整体的样本集数量不大，因此使用默认的决策树参数设置。对于 SVR-RBF 与 SVR-Linear 算法，通过迭代分别选择最佳的惩罚项参数为 1.5 和 0.5。对于 RR 算法，由于样本集数量不大，因此设置默认的参数。同理，带隙、密度和晶体体积属性的参数也能通过迭代过程进行寻优。

3 结果与分析

仍然以形成能为例，表 2 列出了 RF、SVR-RBF、SVR-Linear、RR 算法的交叉验证结果。从表 2 中可以看出 RF 算法在交叉验证的 10 次测试中，其结果相对稳定性较高，这也说明该算法在应对形成能数据集时稳定效果最好。同时，针对带隙、密度和晶体体积属性的交叉验证结果也能说明不同算法应对不同属性数据的稳定效果不同，需要通过其它的模型训练和应用进行进一步的评估。

表 2 形成能的交叉验证结果

Table 2 Cross validation results of formation energy

Test No.	Formation energy/eV			
	RF	SVR-RBF	SVR-Linear	RR
1	0.2379	−0.5595	0.5520	0.3701
2	0.9469	−0.0390	0.8792	0.8129
3	0.7715	0.3404	0.7476	0.6781
4	0.9551	−1.2600	0.1070	0.2523
5	0.8405	0.2031	0.7485	0.6712
6	0.7249	0.1315	0.8191	0.7694
7	0.3566	0.2561	0.5340	0.4132
8	0.9222	0.0214	0.9305	0.8447
9	0.2225	0.0387	0.3515	0.3385
10	0.5823	−0.2729	0.4831	0.3041

表 3 列出了利用不同算法模型的训练性能评估结果， 由于晶体体积性能的取值范围较大(0~1000 Å^3)，表 3 的数据表明他的 E_{MAE}、E_{RMSE} 值明显较其他 3 个更大。RF 算法对 4 个性能的训练效果都非常好，可以解释 90%以上的方差变化，这是因为在训练样本量不大的情况下，RF 算法的集成算法对数据集的拟合效果更好。而 SVR 算法对核函数的选择比较敏感，因此对于不同的性能参数，SVR-RBF 和 SVR-Linear 算法的拟合效果会有较大的差别。

分别利用训练后的多个算法模型对测试集的 4 个性能进行独立测试。表 4 为评估结果，从表 4 中可以

表 3 训练集拟合结果比较

Table 3 Fitting results performance comparison of training set

Property	Method	Evaluation index		
		E_{MAE}	E_{RMSE}	R^2
Formation energy	RF	0.0700	0.0270	0.9454
	SVR-RBF	0.1249	0.0503	0.8983
	SVR-Linear	0.2154	0.1274	0.7420
	RR	0.2331	0.1217	0.7537
Density	RF	0.2348	0.1029	0.9315
	SVR-RBF	0.2808	0.2376	0.8420
	SVR-Linear	0.9094	1.3173	0.1238
	RR	0.9111	1.2737	0.1529
Band gap	RF	0.2041	0.0858	0.9607
	SVR-RBF	0.4232	0.6134	0.7188
	SVR-Linear	0.8208	1.4811	0.3209
	RR	0.8705	1.3664	0.3735
Volume	RF	11.1	435.1	0.9847
	SVR-RBF	101.7	29410.4	−0.0373
	SVR-Linear	27.1	2335.4	0.9176
	RR	18.4	790.0	0.9721

表 4 测试集测试结果比较

Table 4 Testing results comparison of training set

Property	R^2 value			
	SVR-RBF	SVR-Linear	RF	RR
Formation energy	0.4202	0.7964	0.6622	0.7761
Density	0.3260	0.0596	0.7246	−0.0864
Band gap	0.3512	0.2495	0.5783	0.4155
Volume	0.1765	0.9680	0.9394	0.9742

看出：SVR-Linear、RF、RR 算法模型对形成能的解释性更好；对于密度参数，SVR-REF、RF 算法模型的解释性更好；对于带隙性能，RF、RR 算法模型的解释性更好；对于体积性能 SVR-Linear、RF、RR 算法模型的解释性更好。

根据以上结论，结合多个算法模型与真实值的对比图(见图 2)可以看出，SVR-Linear 算法模型对形成能性能的测试结果最好，RF 算法模型对密度性能的测试结果最好，RF 算法模型对带隙性能的测试结果最好；RR 算法模型对体积性能的测试结果最好。SVR-RBF 对 4 个性能的测试结果都不算最好，这是因为 SVR 算法对于核函数的高维(11 维)映射解释力不强，尤其是 RBF 核函数的 SVR 算法。对于当前的钙钛矿材料数据集，RF 算法的表现优于其他算法的表现。

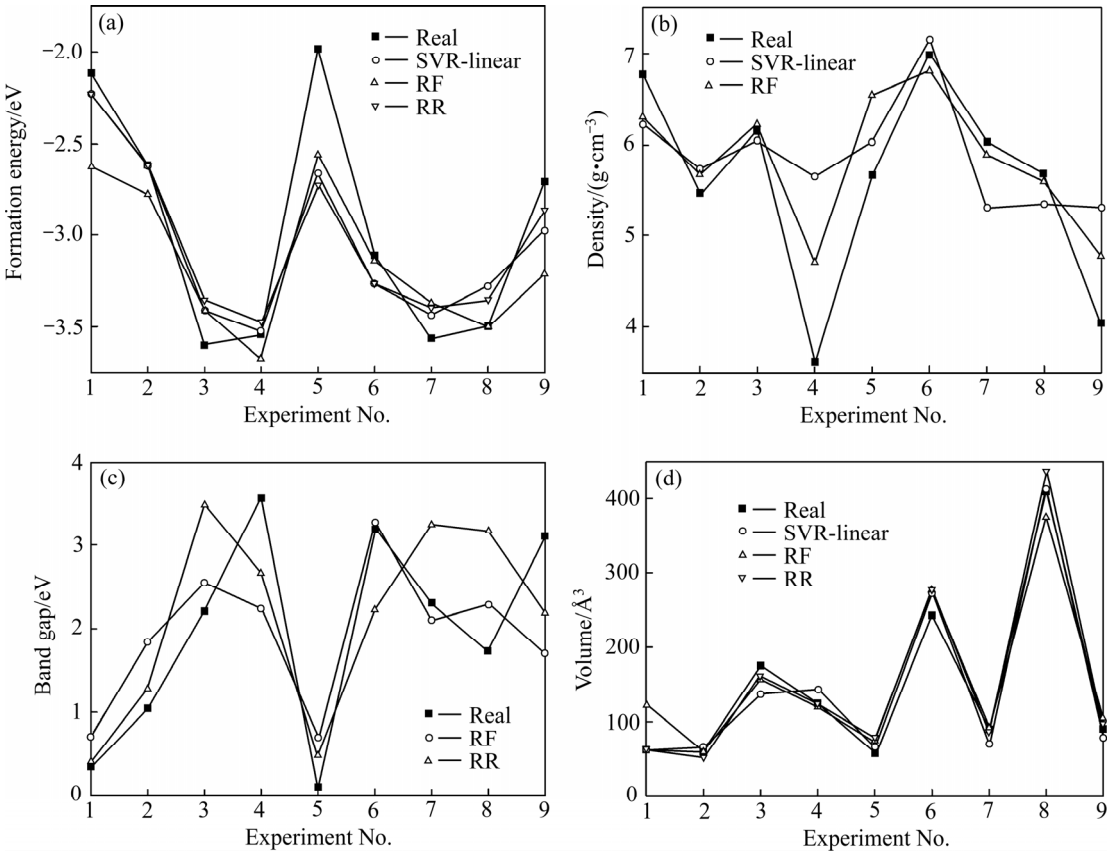

图 2　测试结果比较

Fig. 2　Testing results comparison

但对于有着不同性能取值范围的材料数据集，性能取值划分较多时会对 RF 算法模型的预测效果产生较大的影响，会出现训练集与测试集拟合效果不同的结果。因此，对于形成能和体积性能而言，训练拟合效果最好的 RF 测试效果并非最好。在图 2 中，形成能属性 5 号位的预测与实际值相差较大，这是因为钴酸锶($SrCoO_3$)作为一种钙钛矿型纳米复合氧化物，与其他典型钙钛矿材料相比，其结构和理化性质差异较大。

以上结果表明，不同算法对材料数据不同性能的预测效果并不一样，RF 方法可以对钙钛矿材料的密度、带隙性能进行有效预测；RR 方法可以实现对密度性能的预测；SVR-linear 可以实现对形成能性能的预测。总的来说，在钙钛矿材料体系中，使用传统方法很难得到确定的性能参数，采用机器学习方法解决了这种局限性问题。

4　结论

1) 从钙钛矿型材料自身的特征性能参数出发，分别采用 RF、RR、SVR-RBF 和 SVR-linear 4 种机器学习预测方法，对钙钛矿材料部分性能参数数据进行回归拟合，构建了多算法预测模型，并结合交叉检验的方法找到最优的回归方法。

2) RF 方法可以对钙钛矿材料的密度、带隙性能进行有效预测；RR 方法可以实现对密度性能的预测；SVR-linear 方法可以实现对形成能性能的预测。进一步证明，由于不同的机器学习算法对数据样本分布的敏感程度不同，针对不同的性能参数进行预测需要选择不同方法。将机器学习方法应用在钙钛矿材料数据的性能预测中，可以不经传统实验和第一性原理计算获得相对可靠的完整性能，大大提高了预测效率和接下来的性能预测效果。其结果对研究机器学习方法在钙钛矿体系材料性能预测乃至新材料体系发现方面具有一定的参考价值和实用意义。

3) 在钙钛矿材料的性能预测上，只是从相关数据库和文献中获得特征性能参数，更多的结构性能参数还需进一步的添加与补充。由于钙钛矿材料的数据集数量有限，X 位元素还不能完全实现所有卤族元素钙钛矿材料的性能缺失值补充。因此，在今后的研究中

X 位元素包含 F、Cl、Br、I 等卤族元素的钙钛矿数据集的数量和比例还需进一步扩大，使性能预测效果更具普适性。新的以不同人工智能机器学习方法在材料科学中的应用，具有很大的探索性和挑战性，还需要进行不断的完善、改进和发展，以提高对钙钛矿材料性能的预测精度和更广泛的实用性。

REFERENCES

[1] GRINBERG I, WEST D V, TORREST M, GOU G, STEIN D M, WU LY, CHEN GN, GALLO E M, AKBASHEV A R, DAVIES P K, SPANIER J E, RAPPE A M. Perovskite oxides for visible-light-absorbing ferroelectric and photovoltaic materials[J]. Nature, 2013, 503(7477): 509−512.

[2] SNYDER G J, TOBERER E S. Complex thermoelectric materials[J]. Nature Materials, 2008, 7(2): 105−114.

[3] YAN Y, CHO K H, PRIYA S. Piezoelectric properties and temperature stability of Mn-doped $Pb(Mg_{1/3}Nb_{2/3})$-$PbZrO_3$-$PbTiO_3$ textured ceramics[J]. Applied Physics Letters, 2012, 100(13): 217−221.

[4] ALTFEDER I, KRIM J. Temperature dependence of nanoscale friction for Fe on YBCO[J]. Journal of Applied Physics, 2012, 111(9): 263−279.

[5] JUNGBAUER M, HUHN S, MICHELMANN M, GOERING E, MOSHNYAGA V. Exchange bias in $La_{0.7}Sr_{0.3}MnO_3$/$SrMnO_3$/$La_{0.7}Sr_{0.3}MnO_3$ trilayers[J]. Journal of Applied Physics, 2013, 113(17): 203−205.

[6] JENG J Y, CHIANG Y F, LEE M H, PENG S R, GUO T F, CHEN P, W T C. CH3NH3PbI3 perovskite/fullerene planar-heterojunction hybrid solar cells[J]. Advanced Materials, 2013, 25(27): 3727−32.

[7] LILIENFELD O A V. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties[J]. International Journal of Quantum Chemistry, 2013, 113(12): 1676−1689.

[8] 赵彩甜, 王景芹, 蔡亚楠, 周露露, 吴 倩. La 掺杂 $AgSnO_2$ 触头材料导电性能的第一性原理分析[J]. 中国有色金属学报, 2017, 27(12): 2552−2559.
ZHAO Cai-tian, WANG Jing-qin, CAI Ya-nan, ZHOU Lu-lu, WU Qian. First-principles analysis of conductivity of La-doped $AgSnO_2$ contact material[J]. The Chinese Journal of Nonferrous Metals, 2017, 27(12): 2552−2559.

[9] 沈 丁, 杨绍斌, 李思南, 孙 闻, 唐树伟. 基于第一性原理 Sn-Li 合金嵌锂性能和弹性性能的计算与预测[J]. 中国有色金属学报, 2017, 27(2): 282−288.
SHEN Ding, YANG Shao-bin, LI Si-nan, SUN Wen, TANG Shu-wei. Calculation and prediction of lithium insertion properties and elastic properties foe Sn-Li alloy based on first-principle[J]. The Chinese Journal of Nonferrous Metals, 2017, 27(2): 282−288.

[10] RACCUGLIA P, ELBERT K C, ADLER P D F, FALK C, WENNY M B, MOLLO A, ZELLER M, FRIEDLER S A, SCHRIER J, NORQUIST A J. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, 2016, 533(7601): 73−76.

[11] TAKAHASHI K, TANAKA Y. Material synthesis and design from first principle calculations and machine learning[J]. Computational Materials Science, 2016, 112: 364−367.

[12] MARDIROSSIAN N, HEAD-GORDON M. ωB97X-V: a 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy[J]. Physical Chemistry Chemical Physics Pccp, 2014, 16(21): 9904−9924.

[13] 温玉锋, 蔡从中, 裴军芳, 朱星键, 肖婷婷. R_2O-MO-Al_2O_3-SiO_2 玻璃配方与热膨系数关系的支持向量回归研究[J]. 功能材料, 2009, 40(1): 66−70.
WEN Yu-feng, CAI Cong-zhong, PEI Jun-fang, ZHU Xing-jian, XIAO Ting-ting. Study on the relationship between thermal expansion coefficient and oxide composition of R_2O-MO-Al_2O_3-SiO_2 system glass via support vector regression approach[J]. Journal of Functional Materials, 2009, 40(1): 66−70.

[14] LIU X, LU W, JIN S, LI Y, CHEN N. Support vector regression applied to materials optimization of sialon ceramics[J]. Chemometrics & Intelligent Laboratory Systems, 2006, 82(1/2): 8−14.

[15] COOTES T F, IONITA M C, LINDNER C, SAUER P. Robust and accurate shape model fitting using random forest regression voting[C]//European Conference on Computer Vision. Heidelberg: Springer, 2012: 278−291.

[16] XUE D, XUE D, YUAN R, ZHOU Y, BALACHANDRAN P V, DING X, SUN J, LOOKMAN T. An informatics approach to transformation temperatures of NiTi-based shape memory alloys[J]. Acta Materialia, 2017, 125: 532−541.

[17] ZIO M D, GUARNERA U. Semiparametric predictive mean matching[J]. Asta Advances in Statistical Analysis, 2009, 93(2): 175−186.

[18] JAIN A, ONG S P, HAUTIER G, HAUTIER G, CHEN W, RICHARDS W D, DACEK S, CHOLIA S, GUNTER D, SKINNER D, CEDER G, PERSSON K A. Commentary: The materials project: A materials genome approach to

accelerating materials innovation[J]. Apl Materials, 2013, 1(1): 1049−1060.

[19] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5−32.

[20] ZHANG Z, DAI G, XU C, MI J. Regularized discriminant analysis, ridge regression and beyond[J]. Journal of Machine Learning Research, 2013, 11(12): 2199−2228.

[21] SAUNDERS C, GAMMERMAN A, VOVK V. Ridge regression learning algorithm in dual variables[C]// International Conference on Machine Learning. Madison: Wi, 1998: 515−521.

[22] GU B, SHENG V S, WANG Z, DEREK H, SAID O, LI S. Incremental learning for ν-support vector regression[J]. Neural Networks, 2015, 67(C): 140−150.

[23] 徐 燕, 张玉凤, 高 湉, 张 研, 张惠然, 刘永生. Al 基非晶合金表征参数的支持向量回归分析[J]. 中国有色金属学报, 2016, 26(4): 836−843.
XU Yan, ZHANG Yu-feng, GAO Tian, ZHANG Yan, ZHANG Hui-ran, LIU Yong-sheng. Parameters analysis of Al-based amorphous alloys using support vector regression[J]. The Chinese Journal of Nonferrous Metals, 2016, 26(4): 836−843.

Performance prediction of perovskite materials based on different machine learning algorithms

ZHENG Wei-da[1, 3], ZHANG Hui-ran[2, 3], HU Hong-qing[2], LIU Yao[2], LI Sheng-zhou[2], DING Guang-tai[2], ZHANG Jin-cang[1, 3]

(1. College of Sciences, Shanghai University, Shanghai 200444, China;
2. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
3. School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China)

Abstract: Due to the potential application in various fields, there is great opportunity for further research into the basic physics and chemistry around perovskites. In this work, four machine learning algorithms prediction models have been built. They are random forest(RF), ridge regression(RR), and support vector regression(SVR) based on the radial basis kernel function and linear kernel function. They are used to predict the density, formation energy, band gap, and the crystal volume of the perovskite materials. The experimental results show that the RF method can effectively predict the density and band gap of perovskite materials. The RR method can realize the prediction of density performance. The SVR method of linear kernel function can realize the prediction of the performance. This study shows that different machine learning algorithms have different sensitivity to the distribution of data set samples, so different methods should be selected to predict different performance parameters.

Key words: perovskite; machine learning; performance prediction; algorithm selection

Foundation item: Project(2016YFB0700502) supported by the National Key Research and Development Program, China

Received date: 2018-03-19; **Accepted date:** 2018-05-31

Corresponding author: ZHANG Hui-ran; Tel: +86-13621855760; E-mail: hrzhangsh@shu.edu.cn

(编辑 王 超)