基于高通量计算及机器学习的新材料带隙预测

徐永林¹, 王香蒙², 李鑫², 席丽丽^{2*}, 倪剑樾¹, 朱文浩¹, 张武¹, 杨炯^{2*}

1. 上海大学计算机工程与科学学院, 上海 200444;

2. 上海大学材料基因组工程研究院, 上海 200444

* E-mail: lilyxi@t.shu.edu.cn; jiongy@t.shu.edu.cn

收稿日期: 2018-06-19; 接受日期: 2018-08-06; 网络版发表日期: 2018-12-05

国家重点研发计划(编号: 2017YFB0701501)和国家自然科学基金重大研究计划重点项目(编号: 91630206)资助

摘要 在功能材料应用中, 带隙往往起着重要的作用, 如光电材料一般为宽带隙半导体, 而热电材料为窄带隙半导体, 因此对指定类别的材料体系带隙进行快速而准确的预测对于功能材料应用具有非常重要的科学意义. 然而, 通过基于第一性原理的高通量计算获取高精度带隙的方法耗时长, 效率低, 而实验上系统测量大量材料体系带隙也不现实, 所以基于统计学的机器学习预测方法就成了一种有前景的可能性替代方案. 本文设计了一种集成学习模型用于有效而准确地预测带隙值. 在已计算过带隙值的热电材料类金刚石化合物的基础上, 一方面利用单组元组分替换策略产生大批量相似化合物, 并用查重技术过滤掉重复体系, 得到356个相似材料体系. 另一方面结合机器学习技术, 构建高效的带隙预测模型, 预测并验证了50个相似材料体系的带隙值. 通过实验证明, 该预测模型具有77.73%的准确率, 且足够健壮稳定, 可以广泛应用于需要进行大批量带隙预测的热电材料的研究情景中.

关键词 类金刚石结构, 带隙, 组分替换, 机器学习, 集成学习

1 引言

热电材料通过赛贝克(Seebeck)效应和帕尔贴(Peltier)效应实现热能和电能之间的直接转换^[1-4], 在热电发电和制冷方面具有重要应用. 热电技术已经在工业废热发电利用、空间/深海特殊电源、半导体芯片降温/控温、高端制冷等工业、国防、民用领域取得应用^[1-3], 以热电材料为核心的热电发电器件和系统兼具全固态、无运动部件、无排放、结构简单、体积小、重量轻等诸多优点, 已成为一种具有广泛应用前景的绿色能源技术, 在环境污染和能源危机日益严重的今

天, 进行新型热电材料的研究具有很强的现实意义. 性能优异的热电材料一般是窄带隙半导体材料, 然而半导体材料种类多, 现有的带隙实验数据有限, 大规模数据测量也很难实现, 因此, 带隙的准确预测与计算对新型高性能热电材料的进一步筛选具有重要的意义. 近年来, 在新材料筛选方面, 高通量计算手段显著地加速了材料设计和发现过程^[5-7]. 也随之产生了大量的可开放访问利用的计算/预测材料属性数据库(包括电子结构、热力学和结构性质等)^[8,9]. 然而, 高通量计算往往需要消耗大量的计算资源及时间, 材料科学家们已不满足于通过高通量计算在庞大的结构空间中进

引用格式: 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: 44-54

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 (in Chinese). Sci Sin Tech, 2019, 49: 44-54, doi: 10.1360/N092018-00202

行盲目搜索的低效率方式. 从已有的材料大数据中进行数据挖掘, 提取有用的信息并将这些信息用于未知材料的性能预测, 成为一些材料科学家实现新材料预测的新手段^[10].

最近一些利用统计学技术预测材料属性的探索取得了一定的进展, 包括分子的预测^[11,12], 周期系统的特性^[13~16], 熔化温度^[17], 离子电导率^[18,19], 相稳定性^[20,21], 分子原子化能^[11,22], 介电特性^[22], 过渡态^[23], 结构分类^[21,23], 势能^[24,25], 量子力学的自洽解^[26], 带隙预测^[27~31], 电池监控和筛选新型固态电解质材料^[32]等. 目前, 关于带隙的预测, 已经报道了一些回归模型的应用^[27~31]. Dey等人^[27]基于28个已获得实验带隙的化合物, 用最小二乘回归(OLSR)、稀疏偏最小二乘回归(SPLSR)和套索算法(Least absolute shrinkage and selection operator, 以下简称为Lasso)预测了200种三元黄铜矿化合物的直接带隙, 其中使用了化合物每个元素的价态、原子序数、熔点、电负性和赝势半径等特征描述符. Lee等人^[28]用OLSR、Lasso和支持向量回归算法(Support Vector Regression, 以下简称为SVR)建立了156个AX二元化合物的 G_0W_0 带隙预测模型. Pilania等人^[29]评估了超过120万个属性, 最终使用原子种类的最低占据Kohn-Sham水平和元素电负性作为统计学习特征描述符, 利用KRR模型开发了双钙钛矿带隙预测模型. Setwayan等人^[30]从AFLOWLIB^[9]数据库中选出大约100个化合物作为训练数据集, 用普通OLSR方法估计了PBE KS-带隙与实验带隙之间的关系. Gu等人^[31]应用SVR和人工神经网络预测了25种二元化合物和31种三元化合物的实验带隙. 尽管已有这些开创性的工作, 然而Liu等人^[33]认为在构建材料学习样本、训练泛化能力更好以及学习效率更高的材料统计学习模型上, 仍然有很大的改进空间.

本文的目标是建立一整套基于材料结构查重技术、组分替换策略以及机器学习技术的有效且稳定的新材料带隙预测模型. 该模型将以类金刚石材料为研究样本, 主要是因为类金刚石化合物种类繁多, 从立方闪锌矿结构衍变, 通过不同阳离子等价态多元替换可获得多种类金刚石结构, 且其带隙值变化较大, 从零点几eV到3 eV不等, 一般具有较好的电输运性能. 同时多元阳离子使得材料热导率较低, 是一类有

发展前景的潜在新型热电材料体系^[34]. 到目前为止, 从实验方面对类金刚石化合物体系热电性能的研究较少, 仅为有限的几种, 最优化的功率因子不超过 15 mW/cmK^2 ^[35]. 通过结合机器学习技术与高通量计算手段研究这类材料体系的电子结构, 并进一步学习预测新材料带隙值, 对这类材料体系热电性能初步筛选具有重要的意义.

本文开发了一种有效且稳定的带隙预测模型, 用于建立在预定义的系列结构下材料带隙值和相关特征空间的一个数学映射关系. 在此, 一个材料样本的特征空间由多个不同的材料相关属性组成, 并唯一地描述了该材料样本实例. 同时, 本文提出了利用组分替换策略, 在已有的基于第一性原理高通量计算的材料体系数据的基础上, 粗放式地产生大批量相似新型材料体系, 结合已建立的材料结构查重规则, 借助机器学习技术最终成功获取新材料带隙预测结果的一整套方法, 并展示了特征选择和模型学习的过程, 以寻找最佳的特征子空间来建立带隙和特征空间的映射, 有效且准确地预测了类金刚石组分替换产物的带隙值.

2 研究方法

本文通过材料结构查重技术和组分替换策略产生材料数据集, 包括训练集和测试集, 利用基于第一性原理的VASP (Vienna Ab initio Simulation Package)程序包^[36,37]计算了训练集中所有材料体系的带隙数值, 并在训练集上学习获得相关机器学习模型, 进而通过系列模型预测测试集中新材料体系的未知带隙数值, 而后使用第一性原理高通量计算进行验证, 最终得到预测带隙相对准确的模型. 在带隙数值预测方面, 本文对比分析了Lasso^[38]、SVR^[39]和梯度提升树(Gradient Boosting Decision Tree, 以下简称为GBDT)^[40]三种不同类型的机器学习模型的预测结果, 并结合这三种基本模型设计了一种稳定有效的新型集成学习方法(以下简称为Ensemble)^[41], 以此获得最优带隙预测模型. 如图1所示, 尽管在训练集上GBDT模型的预测精度最高(88.11%), 然而, 在测试集上显示, Ensemble模型是对未知材料体系带隙预测性能最优的模型(预测精度为77.73%). 下文简述了以上各种相关技术以及各种技术之间的相互作用.

图1 类金刚石组分替换产物的带隙预测模型框架

2.1 数据集与特征描述符

2.1.1 数据集

本文利用Yang等人¹⁾计算的87个结构不重复的类金刚石化合物带隙数据,按照2.3节组分替换策略描述,通过与MIP (Materials Informatics Platform)数据库²⁾中的已有数据进行材料结构查重,获得了356个新的不重复结构的材料体系,并随机选择了32个新体系进行高通量计算,与原始的87个类金刚石化合物数据共同构成了用于本次研究中构建带隙预测模型的119个训练样本集,将剩余的324种新体系作为测试集。在进行材料第一性原理计算时,采用了基于密度泛函理论的VASP程序包。所有计算均采用缀加平面波(PAW)的方法来处理电子波函数,采用最新研发的SCAN meta-GGA进行高通量计算^[42,43]。该程序研发者认为此方法在晶格参数、能量计算方面更接近实验值,对电子结构、带隙结果也有所改进。平面波截断能取520 eV,电子自洽能量收敛标准为 10^{-4} eV,所有原子受力均小于0.01 eV。KPOINTS数采用 $K=60/L$ (L 为对应晶格基矢)。本次研究中预测了测试集中所有材料体系的带隙值(见补充材料中模型预测结果“prediction_results.csv”),并随机挑选出50个材料体系进行高通量计算验证,以此获取类金刚石化合物带隙预测的最优模型。

2.1.2 特征描述符

训练样本集中每个材料体系数据样本共有42个特

征描述符和1个目标变量(即带隙)。其中,这些特征描述符来源于类金刚石化合物中原子层面上的多个不同的属性,包括6个晶格结构参数(A, B, C, α (AL), β (BE), γ (GA)),类金刚石化合物中阴离子与所有阳离子的相对电负性差(Electronegativity Difference, ED),阴离子的最近邻配位数(Nearest Neighbor Coordination Numbers, NNCN),方差(Variance, V),阴离子平均键长(Average Shortest Bond Length, ASBL),阴离子原子质量(Anion Atomic Mass, AAM),所有阳离子平均原子质量(Cation Average Atomic Mass, CAAM),平均价电子数(Average Valence Electron Numbers, AVEN),元素种类数(Element Type Numbers, ETN)以及化合物中每个阳离子的电离能(Ionization Energy, IE),单原子质量(Atomic Mass, M),电负性值(Electronegativity, E)和原子数量(Atomic Numbers, N),具体特征数值见补充材料中数据集“training_dataset.csv”和“test_dataset.csv”。例如,类金刚石化合物的带隙值可以使用基于电负性差(ED),1A族元素的单原子质量(1_AAM),1A族元素的电离能(1_AIE)等多个特征描述符的回归模型来预测获得。补充材料中元素以及相关属性值“element_attributes.csv”列出了这些特征描述符中可能涉及到的元素以及相关属性值,包括相对原子质量、电负性以及电离能。

2.2 材料结构查重

鉴于材料结构数据庞大,同一类化合物不同人测

1) Yang J, Xi L L, Pan S S, et al. Discovery of high performance thermoelectric chalcogenides through reliable high throughput material screening. J Am Chem Soc, 2018, 140: 10785–10793

2) Materials Informatics Platform (MIP). <http://mip.shu.edu.cn>.

量会产生多个数据结构, 这些数据结构均存储在已知的数据库中, 如无机晶体学数据库(ICSD)等. 不同实验测得的数据有微量差别, 但属于同一种材料. 通过查重技术, 对具有重复结构的材料体系进行筛选, 将重复的结构去掉, 避免高通量材料计算中的重复计算, 从而可以节省资源、节约时间. 本文在筛选结构不重复材料体系过程中, 利用Phonopy工具³⁾获取误差量在0.1以下的最小原胞^[44], 以化学式、空间群、原子映射关系以及位置对称性值是否相同为判断依据, 获得候选材料体系集合中的不重复条目. 其中, 原子映射关系将体系中互相等价的原子映射到同一个等价群中, 通过比较两个体系中等价群数以及等价群中的等价原子个数和元素类型是否相等来判定二者是否重复. 位置对称性查重通过比较等价原子的Wyckoff符号是否一致来判定两个体系是否重复, 具体查重流程图如图2所示.

2.3 组分替换

新材料体系的发现手段有多种, 常见的方法是使用化学性质相似的元素取代已知化合物晶体结构中的对应对象来产生新的材料体系, 然后进行下一步制备表征、计算模拟、性质研究等^[45]. 例如, 当第一个超导磷族元素氧化物 $\text{LaFeAsO}_{1-x}\text{F}_x$ 被发现时, 晶体化学家开始用其他稀土元素(如钐代替镧)来合成许多其他同构化的新化合物^[46], 与之类似的组分替换方法对新型无机化合物的大规模搜索具有一定的实用价值. 本文中, 通过对每个已有化合物中的单个同族元素做简单的单组元置换来产生新的材料体系候选项, 通过2.2节提到的查重技术过滤掉具有重复结构的体系, 得到356个新的材料体系, 并按照2.1.2节中的方式提取出化合物中蕴含的相关特征数据. 其组分替换流程图如图3所示.

2.4 回归技术

本文使用了三种基本的机器学习回归预测算法, 即Lasso^[38]、SVR^[39]和GBDT^[40], 以及基于这三种基本算法组合成的一个更加强大稳定的Ensemble算法^[41]. 这四种算法具有不同的模型特性, 其中, Lasso具有较好的线性表达能力, SVR可以利用核函数将非线性数

图2 材料结构查重流程. 本文利用Phonopy工具获取最小原胞中的化学式、空间群、原子映射关系和位置对称性, 并依次判断是否与MIP数据库中的材料体系结构重复, 进而筛选出有效数据

图3 组分替换流程图. 通过对MIP类金刚石体系化合物中的每个元素依次执行同族元素单组元置换产生相似化合物, 并与MIP数据库中的已有体系进行查重, 生成新数据集

据转换为线性数据进行处理, GBDT的预测结果可以获得较小的偏差, 而Ensemble算法融合了不同弱学习模型的能力, 具有更好的泛化能力, 详细介绍参见附录B. 通过对原始数据集进行预处理, 获得最优特征子集, 然后利用以上几种预测模型来探索带隙与特征之间的关系, 训练最优带隙预测模型, 同时找出有助于带隙预测的关键特征描述符.

2.5 评价指标

对于2.4节中不同的回归模型产生的预测结果, 本

3) Phonopy. <https://atztogo.github.io/phonopy>.

文采用平均绝对误差(MAE)、均方误差(MSE)和决定系数(R^2)来观察并测量每个模型在测试集上的预测精度, 比较不同模型之间性能差异. 简而言之, MAE与MSE越小且越接近于0, R^2 越大且越接近于1, 说明模型的性能越好, 预测准确率越高. 以下是几种类型误差求解计算模型.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (1)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (3)$$

公式(1)~(3)中, y_i 是从测试集中随机选择的50个类金刚石化合物带隙的高通量计算值; $\bar{y}$ 是其平均值; $\hat{y}_i$ 是与 y_i 对应的回归模型预测值, $i=1, 2, \dots, N$, $N=50$.

为了获取带隙真实值与模型预测结果的MAE, MSE和 R^2 三种指标值, 避免出现随机误差, 使用了10折交叉检验^[47]来综合判断2.4节中各个回归模型的预测能力. 在 k 折(本文中 k 值为10)交叉检验中, 原始数据集被随机分割成 k 个大小相等的子样本集. 在 k 个子样本集中, 保留1个子样本集作为测试模型的验证数据, 剩余的 $k-1$ 个子样本集用来训练模型. 然后将该交叉验证过程重复迭代 k 次, 其中, 每个子样本集仅用作验证数据一次. 最后, 将来自每次交叉验证预测结果的值进行平均, 可以得到一个总体估计的性能指标值.

3 结果分析

本文的目标是利用计算过的已知类金刚石化合物带隙数据预测出经过组分替换后的新材料体系的带隙值, 并通过第一性原理高通量计算对预测结果进行验证, 获取最优带隙预测模型. 该部分3.1节介绍了用于带隙回归预测的特征选择过程; 3.2节描述了几种不同的回归模型对测试集中50个随机选择的化合物的带隙值预测结果, 并挑选出了最优带隙预测模型; 3.3节通过三种不同的技术找出影响带隙值预测的关键特征描述符.

3.1 特征选择

在训练回归模型之前, 通过特征选择, 构建对目标变量带隙具有有效预测能力的输入特征空间, 可以显著地提高模型预测的准确性, 增强模型的泛化能力. 本文使用缺失值处理、方差评估以及特征相关性分析三种技术对具有119个样本的训练数据集集中的42个特征进行了分析与筛选.

3.1.1 缺失值处理

训练数据集集中的化合物元素涉及到“1_A, 2_A, 3_A, 4_A, 6_A, 1_B, 2_B”共7个族中元素的信息, 然而并非每个化合物均包含所有族的元素, 这导致从化合物中提取的训练集特征中出现了大量的缺失数据. 表1描述了训练集的42个特征中的数值缺失比例.

较多缺失值的特征会严重影响模型的判断能力, 因此, 删除了缺失值大于50%的20个特征“1_AN,

表1 训练数据集中特征值缺失率

特征	缺失率	特征	缺失率
1_AN	0.9425	1_BAM	0.3448
1_AIE	0.9425	1_BE	0.3448
1_AAM	0.9425	1_BN	0.3448
1_AE	0.9425	A	0
2_AAM	0.9195	V	0
2_AN	0.9195	C	0
2_AIE	0.9195	B	0
2_AE	0.9195	AL	0
3_AAM	0.7126	BE	0
3_AN	0.7126	GA	0
3_AIE	0.7126	ED	0
3_AE	0.7126	NNCN	0
2_BN	0.6437	ASBL	0
2_BE	0.6437	AAM	0
2_BAM	0.6437	CAAM	0
2_BIE	0.6437	AVEN	0
4_AAM	0.5862	ETN	0
4_AE	0.5862	6_AAM	0
4_AN	0.5862	6_AN	0
4_AIE	0.5862	6_AIE	0
1_BIE	0.3448	6_AE	0

1_AIE, 1_AAM, 1_AE, 2_AN, 2_AIE, 2_AAM, 2_AE, 3_AN, 3_AIE, 3_AAM, 3_AE, 2_BN, 2_BIE, 2_BAM, 2_BE, 4_AN, 4_AIE, 4_AAM, 4_AE”, 对于其他存在缺失值的特征, 使用同族元素的相关元素属性均值进行填充。

3.1.2 方差评估

方差值为0的特征向量对回归模型识别带隙数值大小没有任何判断能力, 因此, 在3.1.1节结果的基础上, 删除了训练数据集中方差值为0的两个特征“AL, 6_AN”。

3.1.3 特征相关性分析

本文绘制了剩余特征之间的皮尔逊线性相关性系数图谱^[48](图4), 以分析并识别出具有高度相关的特征, 消除特征之间的多重共线性^[4]。多重共线性是指回归模

型中的特征之间存在着高度相关关系, 并导致模型估计失真或难以估计准确, 这种现象在线性模型(如Lasso)中尤其明显。因此, 对于具有强相关性(相关性值大于0.8)的多个特征, 一般只保留一个有效特征。从图4中可以明显看出, 1B族中3个属性“1_BIE, 1_BAM, 1_BE”之间具有较高的线性相关性, 可仅保留“1_BE”特征。此外, “AAM”特征与“6_AIE, 6_AAM, 6_AE”三个特征均存在着较强的共线性, 因此, 可仅保留“AAM”特征。

综上, 在经过特征选择后, 数据集中仅保留下了“五个晶格结构参数(A, B, C, BE, GA), 阴离子与所有阳离子的相对电负性差(ED), 阴离子的最近邻配位数(NNCN), 方差(V), 阴离子平均键长(ASBL), 阴离子原子质量(AAM), 所有阳离子平均原子质量(CAAM), 平均价电子数(AVEN), 元素种类数(ETN)、以及化合物中1B族的原子个数(1_BN)和电负性值(1_BE)”共15个

图4 (网络版彩图)训练样本集中剩余特征之间的皮尔逊相关性系数图谱。图中左侧部分数值代表特征之间的皮尔逊相关性系数值, 值域为[-1,1]。皮尔逊相关性系数绝对值越大, 表明两个特征之间相关度越高。其中, 正值代表正相关, 负值代表负相关。

4) <https://newonlinecourses.science.psu.edu/stat501/node/343/>.

有效特征, 用于训练每一个带隙回归预测模型.

3.2 回归结果分析

基于材料训练样本集, 对2.4节中的四种不同的回归算法进行网格调优^[49]产生各自的最优回归模型, 将其在测试集上的带隙预测结果与直接高通量计算的带隙结果进行比较, 如图5所示. 本文计算并绘制了测试集中50个新材料体系的数据, 补充材料模型预测结果 (“prediction_results.csv”)中提供了测试集中所有新体系的带隙预测值. 由图5可知, Ensemble和GBDT具有较好的预测效果, 这两种模型的预测值也更加接近真实值(标记为“HTC”的数据). 此外, 四种模型对带隙值位于0~0.2之间的新材料的带隙预测效果均表现出相对的较大偏差. 原因可能是用于训练统计学习模型的训练集中数据量较少, 导致模型在学习过程中, 对于带隙值位于0~0.2之间的材料体系, 无法获得更好的预测能力.

表2结合三种不同的评价指标描述了四种最优回归模型在测试集上的带隙预测性能表现. 从表2可以得出, Ensemble模型得到了最高的 R^2 值为0.7773, 且均方误差值最低为0.2237, 这得益于Ensemble算法融合了Lasso、SVR和GBDT三种基本模型各自的优势, 即对应的分别间接获得了对材料特征的线性表达能力、核映射空间变化能力和非线性模型表达能力, 对于未知的新数据集具有更强的拟合能力和更好的泛化能力, 而SVR与Lasso的 R^2 值相对较小. Lasso为典型的广义线性回归模型, 而其余模型均为非线性的复杂模型. 由此也可以看出, 该类金刚石化合物数据集的属性特征与目标变量带隙之间有着复杂的非线性关系的特点.

其中, 对于Lasso模型, 参数alpha与max_iter分别代表模型的惩罚项大小和模型训练的迭代次数. 对于SVR模型, 参数kernel、C、gamma与epsilon分别代表核函数类型、惩罚项大小、核函数参数和距离误差.

图5 (网络版彩图)模型预测结果散点图. 该图显示了在测试集上不同回归模型预测得到的带隙值以及利用高通量手段计算得到的带隙值的散点分布, 横坐标为计算值, 纵坐标为预测值

对于GBDT模型, 参数estimators和max_features分别代表了使用到的基本树模型的数量和深度, subsample和learning_rate分别代表了学习率和数据采样比例. 对于Ensemble模型, 使用了基于Stacked generalization的两层模型学习策略^[50], 第一层使用了前面介绍的三种基本模型, 第二层使用了性能较好的GBDT模型.

3.3 特征重要性评价

进行特征重要性评价, 有助于分析出影响类金刚石化合物带隙值预测的关键材料属性. 本文通过斯皮尔曼方法计算了最终用于模型训练的特征与目标变量带隙之间的相关性指数^[48], 用来判断特征对目标变量的重要性. 斯皮尔曼方法适用于非线性变量之间的关系检测, 不论两个变量的总体分布形态、样本容量的大小如何, 都可以通过求解斯皮尔曼相关性系数来进

表2 四种最优回归模型在测试集上的带隙预测结果

模型	参数	绝对值误差 (MAE)	均方根误差 (MSE)	决定系数 (R^2)
Lasso	alpha=0.06, max_iter=100	0.4786	0.4183	0.5615
SVR	kernel=rbf, C=15, gamma=0.108, epsilon=0.08	0.4647	0.3902	0.5910
GBDT	estimators=1500, max_features=2 subsample=0.6, learning_rate=0.03	0.3500	0.2503	0.7377
Ensemble	1rd Layer: Lasso+SVR+GBDT 2nd Layer: GBDT	0.3171	0.2237	0.7773

图 6 (网络版彩图)特征重要性评价. 本文利用斯皮尔曼相关性系数、Lasso和GBDT模型对特征进行打分,并将结果分别归一化到[0,1]区间内进行综合比较,发现“ED”与“CAAM”两个特征在不同的评分算法下均得到了相对较高的分数

行分析.

$$\rho = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}. \quad (4)$$

公式(4)为斯皮尔曼系数公式. 其中, x_i 为训练集中第*i*个样本的特征 x 的值; $\bar{x}$ 是在所有样本上的该特征平均值; y_i 为第*i*个样本的特征 y 的值; $\bar{y}$ 是其平均值, $i = 1, 2, \dots, N$, N 为训练集的样本总数.

此外,在模型训练过程中, Lasso和GBDT两种模型自带特征重要性评分功能. 在图6中, 本文绘制了基于斯皮尔曼系数、Lasso与GBDT模型特征打分功能进行了特征重要性评价. 由图6可知, 电负性差“ED”与化合物中阳离子平均质量“CAAM”两个特征在不同的评分算法下均得到了相对较高的分数, 体现为对训练模型的性能好坏具有较高的影响力.

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 Goldsmid H J. Introduction to Thermoelectricity. 2nd ed. Berlin Heidelberg: Springer-Verlag, 2005. 197–220

4 结论

本文设计了一整套基于材料结构查重技术, 组分替换策略以及机器学习方法的新材料带隙预测模型, 并针对一类典型热电材料-类金刚石化合物组分替换产物的带隙值开展了预测实验. 通过特征选择、网格调优、模型组合等一系列手段, 结合Lasso、SVR和GBDT三种不同模型的优势, 获得了具有77.73%预测精度的集成学习带隙预测模型Ensemble, 并发现了类金刚石化合物的电负性差“ED”与化合物中阳离子平均原子质量“CAAM”两个特征对预测目标变量带隙值具有重要影响. 这种集成学习模型具有强大的学习能力与稳定的预测性能, 可以应用到其它类型材料体系的带隙以及其它材料属性的预测过程中, 加速材料性能的设计与优化过程, 对新型功能材料的快速筛选与高性能预测具有重要的科学意义.

- 2 Snyder G J, Toberer E S. Complex thermoelectric materials. *Nat Mater*, 2008, 7: 105–114
- 3 Bell L E. Cooling, heating, generating power, and recovering waste heat with thermoelectric systems. *Science*, 2008, 321: 1457–1461
- 4 Uher C. Semiconductors and Semimetals. San Diego: Academic Press Inc, 2001. 139–253
- 5 Curtarolo S, Hart G L W, Nardelli M B, et al. The high-throughput highway to computational materials design. *Nat Mater*, 2013, 12: 191–201
- 6 Sharma V, Wang C, Lorenzini R G, et al. Rational design of all organic polymer dielectrics. *Nat Commun*, 2014, 5: 4845
- 7 Ceder G, Hautier G, Jain A, et al. Recharging lithium battery research with first-principles methods. *MRS Bull*, 2011, 36: 185–191
- 8 Curtarolo S, Setyawan W, Wang S, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput Mater Sci*, 2012, 58: 227–235
- 9 Landis D D, Hummelshøj J S, Nestorov S, et al. The computational materials repository. *Comput Sci Eng*, 2012, 14: 51–57
- 10 Service R F. Materials scientists look to a data-intensive future. *Science*, 2012, 335: 1434–1435
- 11 Rupp M, Tkatchenko A, Müller K R, et al. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett*, 2011, 108: 058301
- 12 Huan T D, Mannodi-Kanakkithodi A, Ramprasad R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys Rev B*, 2015, 92: 014106
- 13 Schütt K T, Glawe H, Brockherde F, et al. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys Rev B*, 2014, 89: 205118
- 14 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: 094104
- 15 Faber F, Lindmaa A, von Lilienfeld O A, et al. Crystal structure representations for machine learning models of formation energies. *Int J Quantum Chem*, 2015, 115: 1094–1101
- 16 Faber F A, Lindmaa A, von Lilienfeld O A, et al. Machine learning energies of 2 million elpasolite (ABC2D6) crystals. *Phys Rev Lett*, 2016, 117: 135502
- 17 Seko A, Maekawa T, Tsuda K, et al. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids. *Phys Rev B*, 2014, 89: 054303
- 18 Kishida I. The graph-theoretic minimum energy path problem for ionic conduction. *AIP Adv*, 2015, 5: 107107
- 19 Toyoura K, Hirano D, Seko A, et al. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. *Phys Rev B*, 2016, 93: 054112
- 20 Zunger A. Systematization of the stable crystal structure of all AB-type binary compounds: A pseudopotential orbital-radii approach. *Phys Rev B*, 1980, 22: 5839–5872
- 21 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: 105503
- 22 Pilania G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. *Sci Rep*, 2013, 3: 2810
- 23 Pozun Z D, Hansen K, Sheppard D, et al. Optimizing transition states via kernel-based machine learning. *J Chem Phys*, 2012, 136: 174101
- 24 Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J Chem Phys*, 2011, 134: 074106
- 25 Botu V, Ramprasad R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int J Quantum Chem*, 2015, 115: 1074–1083
- 26 Snyder J C, Rupp M, Hansen K, et al. Finding density functionals with machine learning. *Phys Rev Lett*, 2012, 108: 253002
- 27 Dey P, Bible J, Datta S, et al. Informatics-aided bandgap engineering for solar materials. *Comput Mater Sci*, 2014, 83: 185–195
- 28 Lee J, Seko A, Shitara K, et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys Rev B*, 2016, 93: 115104
- 29 Pilania G, Mannodi-Kanakkithodi A, Uberuaga B P, et al. Machine learning bandgaps of double perovskites. *Sci Rep*, 2016, 6: 19375
- 30 Setyawan W, Gaume R M, Lam S, et al. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Comb Sci*, 2011, 13: 382–390
- 31 Gu T, Lu W, Bao X, et al. Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sci*, 2016, 8: 129–136
- 32 Shi S Q, Jian G, Yue L, et al. Multi-scale computation methods: Their applications in lithium-ion battery research and development. *Chin Phys B*, 2015, 25: 018212

- 33 Liu Y, Zhao T, Ju W, et al. Materials discovery and design using machine learning. *J Materiomics*, 2017, 3: 159–177
- 34 Sklyarchuk V M, Plevachuk Y O. Electronic properties of liquid Ti_2Te , Ti_2Se , Ag_2Te , Cu_2Te , and Cu_2Se alloys. *Semiconductors*, 2002, 36: 1123–1127
- 35 Luo Y, Yang J, Jiang Q, et al. Progressive regulation of electrical and thermal transport properties to high-performance CuInTe_2 thermoelectric materials. *Adv Energy Mater*, 2016, 6: 1600007
- 36 Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys Rev B*, 1996, 54: 11169–11186
- 37 Blöchl P E. Projector augmented-wave method. *Phys Rev B*, 1994, 50: 17953–17979
- 38 Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. *J R Stat Soc Ser B Stat Methodol*, 2011, 73: 273–282
- 39 Smola A J, Schölkopf B. A tutorial on support vector regression. *Stat Comput*, 2004, 14: 199–222
- 40 Friedman J H. Greedy function approximation: A gradient boosting machine. *Ann Stat*, 2001, 29: 1189–1232
- 41 Zhang C, Ma Y. *Ensemble Machine Learning Methods and Applications*. Springer, 2012. 35–87
- 42 Perdew J P, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett*, 1996, 77: 3865–3868
- 43 Sun J, Ruzsinszky A, Perdew J P. Strongly constrained and appropriately normed semilocal density functional. *Phys Rev Lett*, 2015, 115: 036402
- 44 Moitra A. Introduction to Electronic Structure Calculations using VASP. 2010. https://icme.hpc.msstate.edu/mediawiki/images/8/8d/IntroDeckElectStrucCalcJuly28_2010.pdf
- 45 Hautier G, Fischer C, Ehrlicher V, et al. Data mined ionic substitutions for the discovery of new compounds. *Inorg Chem*, 2010, 50: 656–663
- 46 Johrendt D, Pöttgen R. Pnictide oxides: A new class of high-TC superconductors. *Angew Chem Int Edit*, 2008, 47: 4782–4784
- 47 Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, 1995, 14: 1137–1145
- 48 Hauke J, Kossowski T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 2011, 30: 87–93
- 49 Staelin C. Parameter selection for support vector machines. Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1. 2003. <http://www.hpl.hp.com/techreports/2002/HPL-2002-354R1.html>
- 50 Wolpert D H. Stacked generalization. *Neural Networks*, 1992, 5: 241–259

New materials band gap prediction based on the high-throughput calculation and the machine learning

XU YongLin¹, WANG XiangMeng², LI Xin², XI LiLi², NI JianYue¹, ZHU WenHao¹, ZHANG Wu¹ & YANG Jiong²

¹ School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;

² Materials Genome Institute of Shanghai University, Shanghai 200444, China

The bandgap often plays an important role in functional materials applications. For example, optoelectronic materials are generally wide bandgap semiconductors, while thermoelectric materials are narrow bandgap semiconductor materials. Therefore, predicting the bandgap rapidly and accurately for a given class of materials structures has great scientific importance for the functional materials applications. However, considering that the method of obtaining high-precision band gaps based on first-principles high-throughput calculations is time consuming and inefficient, and it is also not realistic to systematically measure a large number of material system band gaps. Machine learning methods based the statistics may be a promising alternative. This paper designs an ensemble learning model for effectively and accurately predicting bandgap values. Based on the calculated band gap values of diamond-like structures in thermoelectric materials, on the one hand, single component substitution strategy was used to generate large quantities of similar compounds, and the repetitive structures was filtered out by using the structural repeatability examination technique, resulting in 356 unique material structures. On the other hand, in combination with machine learning techniques, an efficient band gap prediction model was constructed, and by which the band gap values of 50 similar material systems are predicted and verified. As is the result of the experiment, this prediction model has 77.73% accuracy. It is enough robustness and stability to be widely used in thermoelectric materials application scenarios which require large band gap prediction.

diamond-like structures, bandgap, component substitution approach, machine learning, ensemble learning

doi: [10.1360/N092018-00202](https://doi.org/10.1360/N092018-00202)