第 30 卷第 8 期 中国有色金属学报 2020 年 8 月
Volume 30 Number 8 The Chinese Journal of Nonferrous Metals August 2020
DOI: 10.11817/j.ysxb.1004.0609.2020-35838

# 基于 SISSO 和机器学习方法的钙钛矿结构的稳定性预测：新型容许因子建立与验证

胡红青[1]，吴邵刚[1]，郭治廷[1]，周高锋[1]，戴东波[1]，魏 晓[1,2]，张惠然[1,2]

(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)

**摘 要**：由于钙钛矿型材料具有广泛的应用前景，因此对其结构及物理、化学性质的研究一直是材料研究领域的热点之一。其中，利用容许因子(Tolerance factor)来预测钙钛矿型材料的结构稳定性可以帮助研究者发现更多的新型功能材料，而传统的基于离子半径定义的容许因子 $t_{IR}$ 存在一定的局限性。本文基于 SISSO(Sure independence screening and sparsifying operator)方法和键价模型提出一种新型的容许因子 $\tau_{BV}$，其可以有效地避免由离子半径带来的局限性。本工作使用机器学习中的决策树算法建立容许因子验证模型，实验结果表明，新型容许因子 $\tau_{BV}$ 可以很好地预测 $ABO_3$ 型化合物是否具有钙钛矿结构，并大大提高了预测精度。

**关键词**：钙钛矿；结构稳定性；SISSO；新型容许因子
**文章编号**：1004-0609(2020)-08-1887-08 **中图分类号**：TB34 **文献标志码**：A

钙钛矿结构是最常见的材料结构之一，由于具有该结构的材料其物理和化学性质极具多样性，其在材料化学领域、生产生活中被广泛地研究和应用[1−5]。$ABO_3$ 钙钛矿结构定义为任何包含一个共享角的 $BO_6$ 八面体的三维框架的 $ABO_3$ 型化合物[6]。自然界中存在着大量的 $ABO_3$ 型化合物，但其中只有部分具有钙钛矿结构。对 $ABO_3$ 型化合物进行钙钛矿结构分类已研究了多年[7−10]，预测 $ABO_3$ 化合物能否形成钙钛矿结构一直是一个长期的研究点[11]。传统的方法是使用结构图(Structure map)[12]对化合物进行钙钛矿结构分类。但这些结构图仅包含了化合物的某两个特征，若要将这些样本点进行分类还需要人为地画出分界线，其分类效率低下且效果并不理想。

自 GOLDSCHMIDT[13]于 1926 年首次提出容许因子($t_{IR}$)以来，利用容许因子来判断材料的结构成了一种重要的方法。容许因子能够描述离子半径与结构稳定性的关系，就钙钛矿结构而言，$t_{IR}$ 的表达式如下：

$$t_{IR}=\frac{(R_A+R_O)}{\sqrt{2}(R_B+R_O)} \tag{1}$$

式中：$R_A$、$R_B$ 和 $R_O$ 分别表示 A、B 和 O 离子的半径。在理想的钙钛矿立方结构中，$t_{IR}$ 的理论值为 1。然而，对于很多已知具有钙钛矿结构的晶体，A 和 B 的离子半径会在一定范围内波动，其 $t_{IR}$ 的值分布在 0.78~1.05 之间[14−15]，此时 $ABO_3$ 型化合物能保持稳定的钙钛矿立方结构。当 $t_{IR}<0.78$ 时，晶体为铁钛矿结构；当 $t_{IR}>1.05$ 时，晶体则为方解石或者文石结构[14, 16−17]。由此来看，对于一种未知的 $ABO_3$ 化合物，其 $t_{IR}$ 的值偏离 1 越多，其越不具备立方体结构。因此，将 $t_{IR}$ 作为评估钙钛矿结构稳定性的标准已经较为常用。

最近越来越多的研究表明，利用容许因子 $t_{IR}$ 来预测化合物是否为钙钛矿结构时，它的准确性并不高[6]。近年来，随着机器学习的发展，机器学习在预测钙钛矿结构问题中得到了广泛应用[18−19]。在 2018 年，BARTEL 等[6]借助于 SISSO 方法[20]得到了一个新的容许因子 $\tau$，预测的准确性获得了较大提高(提高了 18%)。该新容许因子 $\tau$ 具有以下的形式：

$$\tau=\frac{R_X}{R_B}-n_A\left[\frac{\frac{R_A}{R_B}}{\ln\left(\frac{R_A}{R_B}\right)}\right] \tag{2}$$

式中：$n_A$ 表示 A 离子的价态，$R_i$ 表示 $i$ 离子的半径。然而，无论是 $t_{IR}$ 还是 $\tau$，都没有考虑配位数对离子半

**基金项目**：国家重点研发计划项目(2018YFB0704400)
**收稿日期**：2019-08-19；**修订日期**：2019-12-06
**通信作者**：张惠然，讲师，博士；电话：13621855760；E-mail：hrzhangsh@shu.edu.cn

径的影响，直接用离子半径之和表示两个离子之间的距离存在着一定的缺陷[19]。为解决上述问题中离子半径的不足，通过基于键价模型(Bond-valence model)计算得到的 A—O 和 B—O 的键长(用 $d_{AO}$ 和 $d_{BO}$ 表示)，用 $d_{AO}$ 代替 $R_A+R_O$，用 $d_{BO}$ 代替 $R_B+R_O$，可以得到一个新的容许因子 $t_{BV}$，其表达式如下：

$$t_{BV}=\frac{d_{AO}}{\sqrt{2}d_{BO}} \tag{3}$$

在本文中，借助于 SISSO 方法，利用机器学习中的决策树方法[21]分别构建了基于 $t_{BV}$ 和 $\tau_{BV}$ 的两个钙钛矿结构预测模型，并对模型的性能进行评估和比较。实验结果表明，利用 $d_{AO}$ 和 $d_{BO}$ 构建的新型容许因子 $\tau_{BV}$，在样本上表现出比 $t_{BV}$ 更高的分类准确率。

# 1 方法

## 1.1 数据集

该数据集包含 376 种 $ABO_3$ 型化合物，其中 232 种化合物具有钙钛矿结构，另外 144 种为非钙钛矿结构化合物。根据 A、B 离子的价态，可将化合物分为三大类别：$A^{1+}B^{5+}O_3$、$A^{2+}B^{4+}O_3$ 和 $A^{3+}B^{3+}O_3$。对每一种化合物，选取以下特征：A、B 离子的价态，A—O、B—O 的键长以及由键价模型得到的容许因子 $t_{BV}$，具体如表 1 所示。

## 1.2 SISSO

SISSO(Sure independence screening sparsifying operator)是一种基于压缩感知的数据分析方法，它可以在大量的特征中找到一个最佳的低维度的特征描述子[20]。对于材料数据而言，SISSO 可以构造出一个公式，该公式可以描述材料的某个指定的性质，且公式中包含若干个与该性质相关的特征。

SISSO 算法从构造特征空间开始。所有与目标性质可能相关的特征组成初始的特征空间 $\Phi_0$，$\Phi_0$ 中包含已知的材料特征，如原子半径、电离能等等。定义运算符集合为：

$$\hat{H}^{(m)}=\{I,+,-,\times,/,\exp,\log,|-|,\sqrt{},^{-1},^{-2},^{-3}\}[\phi_1,\phi_2] \tag{4}$$

式中：$\phi_1$、$\phi_2$ 表示在特征空间 $\Phi$ 中的两个元素。从 $\Phi_0$ 开始迭代，每次迭代将 $\hat{H}^{(m)}$ 作用于特征空间，在迭代 $n$ 次时，特征空间 $\Phi_n$ 可表示如下：

$$\Phi_n=\bigcup_{i=1}^{n}\hat{H}^{(m)}[\phi_1,\phi_2],\ \forall\phi_1,\phi_2\in\Phi_{i-1} \tag{5}$$

显然，$\Phi_n$ 中的元素数目会随着 $n$ 的变大急剧增多。SISSO 过程可以分为 SIS 和 SO 两部分。其中 SIS 使用相关幅度(目标属性和特征之间的内积的绝对值)对每个度量化后的特征进行评分，并且仅保持分数排名最高的；SO 用于精确定位最佳的 $n$ 维描述符[19]。图 1 表示 SISSO 的总体框架：将那些与 SIS 产生的残差 $\Delta$(或 $P$)具有最大相关性的子空间进行集合并运算，得到子空间的并集，然后利用 SO 进一步提取最佳描述子。

## 1.3 ROC 曲线

在机器学习领域，ROC(Receiver operating characteristic)曲线[22]被广泛地应用于对模型性能的评估。根据模型的预测结果对样本进行排序，然后按照排序结果逐个把样本作为正例进行预测，每次计算出两个值，并以它们为横、纵坐标作图，即可得到 ROC 曲线。

ROC 曲线的纵坐标为“真正例率”，即 TPR(True positive rate)，横坐标为“假正例率”，即 FPR(False positive rate)。两者定义如下：

**表 1** 实验数据集特征细节

**Table 1** Details of experimental data set

| Parameter | Minimum | Maximum | Average | Description |
|---|---|---|---|---|
| $n_A$ | 1 | 3 | 2 | Valence state of A-site ion (1,2,3) |
| $n_B$ | 3 | 5 | 4 | Valence state of B-site ion (3,4,5) |
| $d_{AO}$/Å | 2.133 | 3.336 | 2.631 | A—O bond length |
| $d_{BO}$/Å | 1.499 | 2.428 | 1.970 | B—O bond length |
| $t_{BV}$ | 0.730 | 1.501 | 0.730 | Tolerance factor calculated from bond valence model |

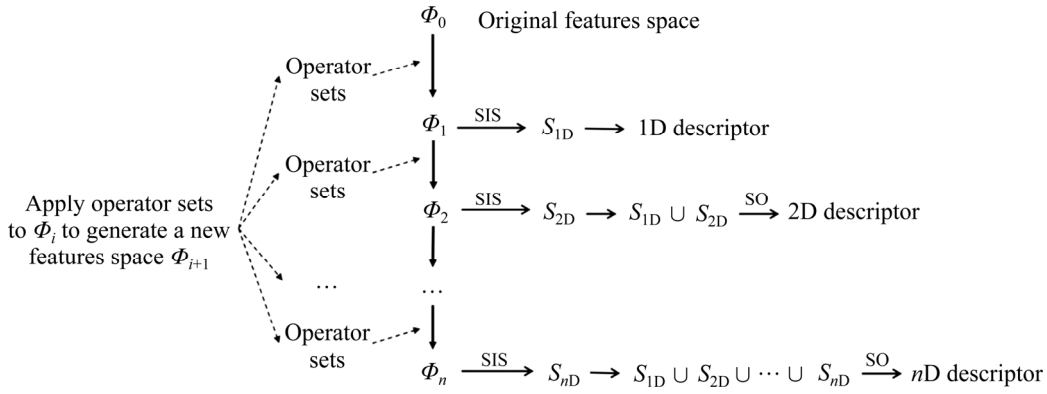


**图 1**　SISSO 算法的总体框架示意图

**Fig. 1**　Schematic diagram of SISSO algorithm (Apply the operator sets to $\Phi_i$ to generate a new features space $\Phi_{i+1}$. Algorithm uses SIS to score new features space to select feature with highest score, and use SO to further extract best descriptor)

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{6}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN+FP}} \tag{7}$$

式中：TP 为真正例(True positive)样本数，FP 为假正例(False positive)样本数，TN 为真反例(True negative)样本数，FN 为假反例(False negative)样本数，具体含义如表 2 所示。

**表 2**　混淆矩阵

**Table 2**　Confusion matrix

| Real | Predict | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

# 2　模型的建立

## 2.1　实验流程的建立

实验的具体流程如图 2 所示。从原始数据集中进行特征选择，挑选出部分特征，再额外加入新的描述子组成数据集 $A$，对 $A$ 数据集运用 SISSO 算法从而得到新的容许因子 $\tau_{\mathrm{BV}}$。再将新得到的 $\tau_{\mathrm{BV}}$ 通过机器学习中的决策树算法建立预测钙钛矿材料稳定性的模型，同时作为对比也将原 $t_{\mathrm{BV}}$ 作为描述子建立决策树预测模型，从而得到新的 $\tau_{\mathrm{BV}}$ 的预测效果。

## 2.2　数据集 $A$ 的生成

该数据集包含上述 376 种 $ABO_3$ 型化合物及它们的特征。将特征分为三类：离子的价态($n_{\mathrm{A}}$，$n_{\mathrm{B}}$)，键长($d_{\mathrm{AO}}$，$d_{\mathrm{BO}}$)和键长的比值($d_{\mathrm{AO}}/d_{\mathrm{BO}}$)。该数据集包含 376 行数据，每一行数据代表一种化合物并包含了这 5 个特征，前 232 行数据为钙钛矿结构，后 144 行数据为非钙钛矿结构。

## 2.3　新型容许因子 $\tau_{\mathrm{BV}}$ 的生成

SISSO 模型以 SISSO.in 和数据集 $A$ 作为输入数据，其中 SISSO.in 中包含该模型的参数，如迭代的次数，构造特征空间的梯度等；数据集 $A$，即 2.2 节中生成的数据集，则作为模型构建时的训练数据。调整好 SISSO.in 文件中的参数和数据集 $A$ 后，将二者作为输入数据，通过 SISSO 模型可得到一个新的容许因子 $\tau_{\mathrm{BV}}$。$\tau_{\mathrm{BV}}$ 的表达式中应包含数据集 $A$ 中的若干个特征。通过 SISSO 的方法，可以将数据集 $A$ 中的 5 个特征进行代数运算或函数运算(如四则运算，开方和取对数等)的结合，得到一个由这 5 个特征(或不足 5 个特征)组合成的新的表达式。其具有以下形式：

$$\tau_{\mathrm{BV}}=\frac{n_{\mathrm{B}}\times d\times \mathrm{e}^{d}}{\ln d_{\mathrm{BO}}\times \ln d} \tag{8}$$

式中：$d=d_{\mathrm{AO}}/d_{\mathrm{BO}}$。

在式(3)和(8)中的键长($d_{\mathrm{AO}}$ 和 $d_{\mathrm{BO}}$)比离子半径之和($R_{\mathrm{A}}+R_{\mathrm{O}}$ 和 $R_{\mathrm{B}}+R_{\mathrm{O}}$)更能精确地表示两个离子之间的距离。除此之外，通过键价模型计算得到的键长，可弥补某些钙钛矿材料因离子半径难以测定而造成的数据缺失问题。

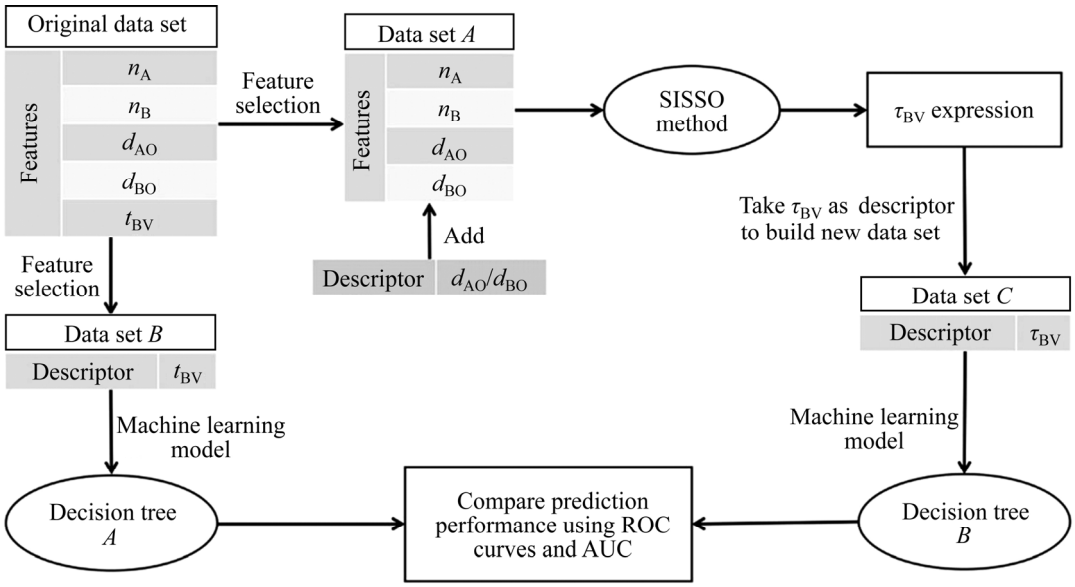


**图 2** 实验整体学习流程

**Fig. 2** Overall learning process of experiment (Data set *A* is used to generate a new tolerance factor $\tau_{BV}$ based on SISSO method, and data set *C* is used to establish a perovskite stability prediction model taking $\tau_{BV}$ as descriptor. For comparison purposes, data set *B* uses $t_{BV}$ as descriptor to build a prediction model)

为了使用机器学习中的分类器对 $ABO_3$ 型化合物进行分类，需要找到一个描述子作为特征，输入到该分类器中，而分类器能输出预测的结果，即化合物是否具有钙钛矿结构。在这整个过程中，描述子至关重要，找到的描述子必须能精准地表示化合物的结构，因此，描述子本身也应该由那些与结构相关的特征组成。无论是 $t_{IR}$ 还是 $\tau$，它们的表达式中都有离子半径这一属性。为了确定某个离子在化合物中的半径，需要知道其配位数，但对于有些离子，相应配位数下的半径则无法获取[19]。这就导致了无法得到这些 $ABO_3$ 型化合物中 A、B 的离子半径，也就无法准确地表示 A—O 和 B—O 的长度。因此这种表示方法没有考虑配位数对离子半径的影响。而利用由键价模型(Bond-valence model)计算得到的 A—O 和 B—O 的键长($d_{AO}$ 和 $d_{BO}$)所定义的新型容许因子 $t_{BV}=\dfrac{d_{AO}}{\sqrt{2}d_{BO}}$，它被认为是判断 $ABO_3$ 型钙钛矿结构稳定性的新标准。

以上的事实提供了一个启发：用($d_{AO}$，$d_{BO}$)替代($R_A+R_O$，$R_B+R_O$)得到的 $t_{BV}$，比 $t_{IR}$ 更能准确地表示化合物的立方结构。因此，在实验中将($d_{AO}$，$d_{BO}$)作为基本特征，输入到以上的模型中，从而得到一个基于离子的价态($n_A$，$n_B$)，键长($d_{AO}$，$d_{BO}$)和键长之比($d_{AO}/d_{BO}$)的全新的描述子 $\tau_{BV}$。

## 3 结果验证与讨论

为了验证新型容许因子，对 $t_{BV}$ 和 $\tau_{BV}$ 的分类性能进行比较。如图 2 所示，从原始数据集中选择 $t_{BV}$ 这一特征，用于构建数据集 *B*。类似地，对原始数据集中的每一种化合物，将 $\tau_{BV}$ 作为新的特征，可构建得到新的数据集 *C*。因此，最终得到的数据集 *B* 和数据集 *C* 各包含了原始数据集中的 376 种化合物。

利用式(8)计算出 376 种 $ABO_3$ 型化合物的 $\tau_{BV}$ 值，可得到样本在该维度上的分布情况，其与 $t_{BV}$ 的对比如图 3 所示。从图 3(b)中可明显看出，钙钛矿和非钙钛矿重叠更少，说明 $\tau_{BV}$ 可以更好地将钙钛矿与非钙钛矿分离开来，使二者在 $\tau_{BV}$ 这一维度上有更明显的边界；相比之下，图 3(a)中钙钛矿和非钙钛矿重叠较多，说明 $t_{BV}$ 难以对样本进行明显地类别分离。

图 3 在 $\tau_{BV}$ 样本的分布给了非常直观的了解，为了更好地说明 $\tau_{BV}$ 在预测钙钛矿结构稳定性的分类问题上的效果，利用机器学习的方法，构造一个深度为 2 的决策树分类器，分别对数据集 *B* 和数据集 *C* 中的数据进行分类。

决策树模型作为一种白盒模型，易于理解和实现，且可解释性极强。利用决策树模型，可以在 $t_{BV}$ 和 $\tau_{BV}$

的特征维度上，进行数值区间的划分，找到最有可能具有钙钛矿结构的区间。分别以 $t_{BV}$ 和 $\tau_{BV}$ 为特征，构造成功后的决策树模型如图 4 所示，矩形内为判断条件，圆形和三角形为叶子结点(预测结果)，圆形代表钙钛矿结构，三角形代表非钙钛矿结构。

当化合物的 $t_{BV}$ 值在 0.831~1.062 的范围内时，分类器判断其为钙钛矿，否则为非钙钛矿(见图 4(a))。对于图 4(b)中的 $\tau_{BV}$，则以 99.56 为分界线：低于该值则判断化合物为钙钛矿，否则为非钙钛矿。决策树分类器得到的结果，这与图 3 中所表达的信息相符，即 $\tau_{BV}$ 能把化合物的钙钛矿与非钙钛矿结构在该特征维度上变得更加线性可分。

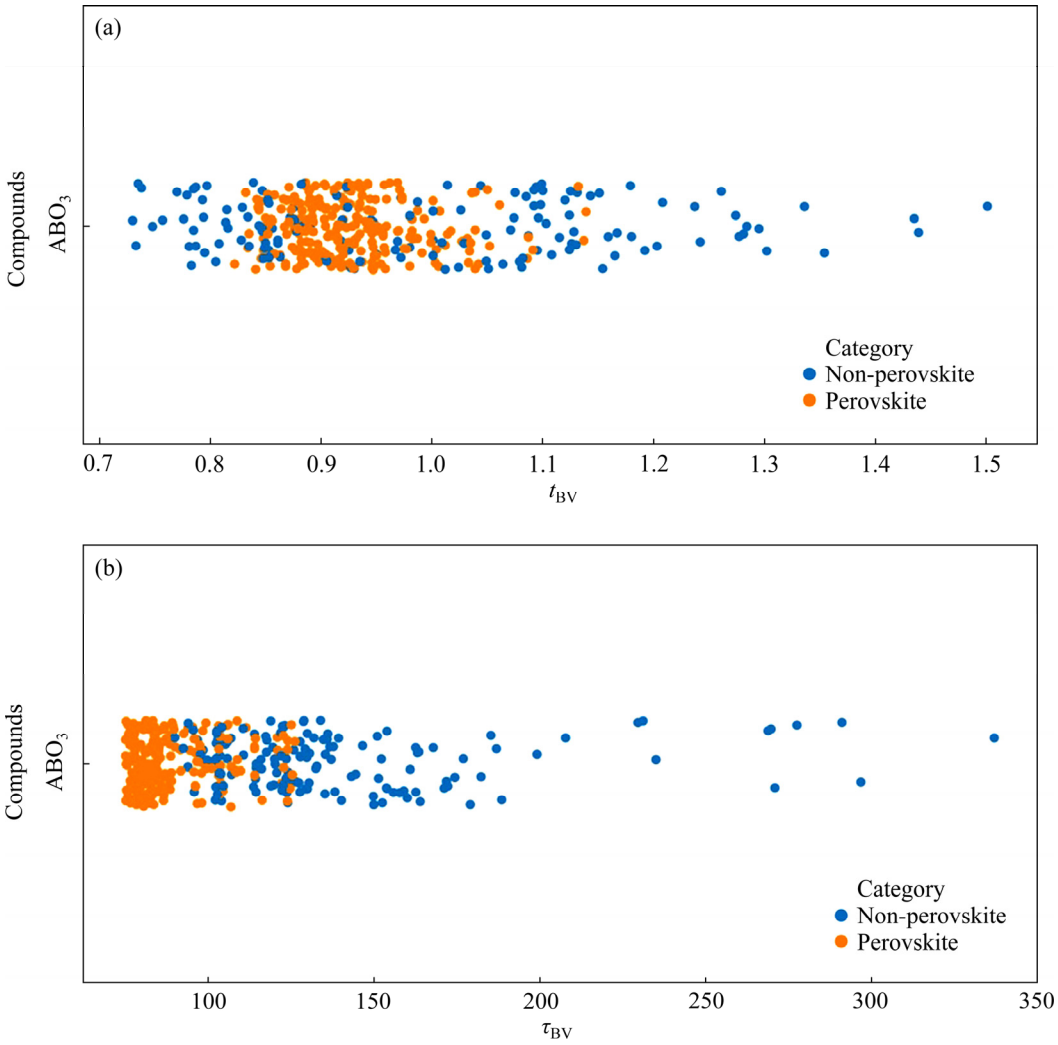


**图 3**　$ABO_3$ 型化合物在 $t_{BV}$ 和 $\tau_{BV}$ 上的分布情况

**Fig.3**　Distributions of $ABO_3$ type compounds on $t_{BV}$(a) and $\tau_{BV}$(b)

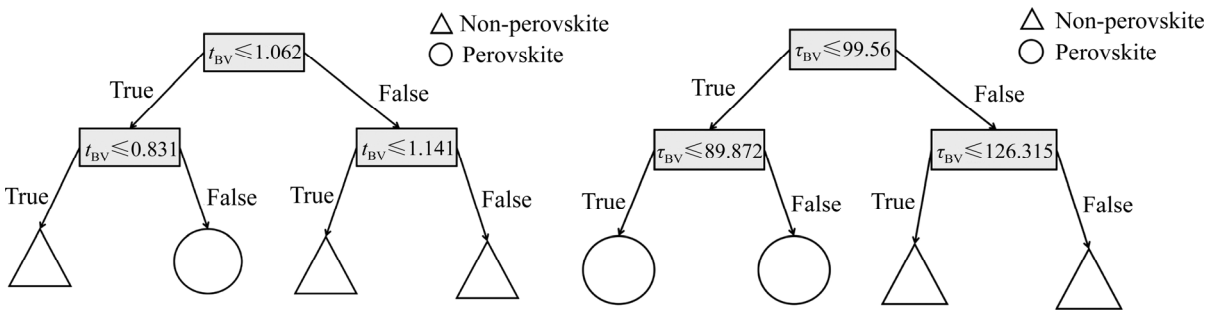


**图 4**　基于 $t_{BV}$ 和 $\tau_{BV}$ 的决策树验证模型

**Fig. 4**　Decision tree models based on $t_{BV}$(a) and $\tau_{BV}$(b)

为了对以上两个决策树模型的性能进行评估，可借助于 ROC 曲线和 AUC。由于数据集较小，模型预测效果受训练集和测试集划分影响较大。为了减轻这种影响，采用 10 折交叉验证法，将包含 376 条化合物的数据集划分为 10 个大小相似的子集。这样就获得了 10 组训练/测试模型，即产生了 10 个分类器，从而进行了 10 次训练和测试。对于每一次训练得到的分类器，使用 ROC 曲线和 AUC 来评估该分类器的性能。其中，AUC 指的是由 ROC 曲线与坐标轴围起来的面积，其值在 0~1 之间，值越接近 1，说明该分类器越完美。

对于图 4 中的两个决策树模型分别进行 10 次交叉验证，相应地，分别产生 10 个不同的 ROC 曲线和 AUC，对其取均值后得到的 ROC 曲线和 AUC 则可用来度量模型的性能。基于 $t_{BV}$ 构造的决策树模型的 ROC 曲线和 AUC 的值如图 5(a)所示，10 次交叉验证的结果中，AUC 的最小值为 0.61，最大值为 0.88，均值为 0.76；基于 $\tau_{BV}$ 构造的决策树模型的 ROC 曲线和 AUC 的值如图 5(b)所示，AUC 的最小值为 0.71，最大值达到 0.99，均值达到 0.91。这说明基于 $\tau_{BV}$ 构造的决策树模型在本论文数据集上表现优异，比基于 $t_{BV}$ 构造的决策树模型预测效果准确许多。

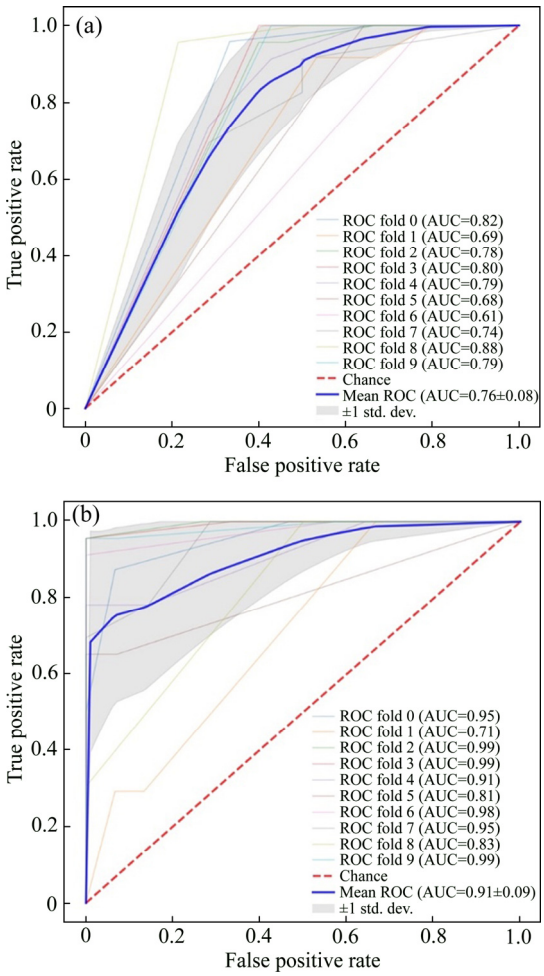


**图 5** 基于 $t_{BV}$ 和基于 $\tau_{BV}$ 的决策树模型 ROC 曲线和 AUC 值

**Fig.5** ROC curves and AUC values of decision tree model based on $t_{BV}$(a) and $\tau_{BV}$(b)

以上结果表明，通过使用 SISSO 算法，结合离子价态和键长信息得到的新容许因子 $\tau_{BV}$ 是有效的。无论是直观的 $ABO_3$–$\tau_{BV}$ 分布图，还是基于决策树的分类模型都显示出新的容许因子 $\tau_{BV}$ 对钙钛矿的预测能力要优于其他文献中提及的容许因子 $t_{BV}$。此外，基于 ROC 曲线的模型分析也表明，基于 $\tau_{BV}$ 构建的机器学习模型有着更好的鲁棒性和精确性。另一方面，通过 SISSO 算法来寻找材料描述子的方法也被证明是可行的，对已有特征的非线性组合进行探索，进而找到能够准确描述目标属性(如本文中的钙钛矿稳定性)的新描述子，该方法甚至还可以用于其他描述子的发现。

## 4 结论

1) 提出了一种新型容许因子表示 $\tau_{BV}$，其由离子的价态($n_A$，$n_B$)、键长($d_{AO}$，$d_{BO}$)和键长之比($d_{AO}/d_{BO}$)构成。基于键价模型得到的键长能够在一定程度上避免由于配位数的存在得到不精确离子半径的影响，并且能够利用机器学习方法(决策树)以极小的计算成本实现对钙钛矿稳定性的良好预测。

2) 原容许因子 $t_{BV}$ 对于钙钛矿稳定性的预测 AUC 平均值在 0.76，而新型容许因子 $\tau_{BV}$ 对钙钛矿稳定性的预测 AUC 平均值在 0.91，准确度上有了较大提升。

3) 由于 $\tau_{BV}$ 的简单性和预测准确性，期待其将能被用于指导探索钙钛矿型新材料的发现和设计。同时也表明了，在一定条件下机器学习方法能够快速、有效地验证理论计算结果。

## REFERENCES

[1] LU Shuai-hua, ZHOU Qiong-hua, OUYANG Yi-xin, GUO Yi-lv, LI Qiang, WANG Jin-lan. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning[J]. Nature communications, 2018, 9(1): 3405.

[2] 郑伟达, 张惠然, 胡红青, 刘 尧, 李盛洲, 丁广太, 张金仓. 基于不同机器学习算法的钙钛矿材料性能预测[J]. 中国有色金属学报, 2019, 29(4): 803−809.

ZHENG Wei-da, ZHANG Hui-ran, HU Hong-qing, LIU Yao, LI Sheng-zhou, DING Guang-tai, ZHANG Jin-cang. Performance prediction of perovskite materials based on different machine learning algorithms[J]. The Chinese Journal of Nonferrous Metals, 2019, 29(4): 803−809.

[3] 史　册, 邵光杰, 胡　婕, 赵北龙, 吕彦玲. 钙钛矿型复合氧化物纳米薄膜的研究进展[J]. 中国有色金属学报, 2008, 18(10): 1893−1902.
SHI Ce, SHAO Guang-jie, HU Jie, ZHAO Bei-long, LÜ Yan-ling. Progress in nano-thin films of perovskite-type complex oxides[J]. The Chinese Journal of Nonferrous Metals, 2018, 18(10): 1893−1902.

[4] 刘红莎, 郇昌梦, 肖秀娣, 毕卓能, 陆　源, 齐　帅, 詹勇军, 徐雪青, 徐　刚. 无机钙钛矿太阳能电池研究进展[J]. 新能源进展, 2019, 7(2): 142−148.
LIU Hong-sha, HUAN Chang-meng, XIAO Xiu-di, BI Zhuo-neng, LU Yuan, QI Shuai, ZHAN Yong-jun, XU Xue-qing, XU Gang. Research progress of all-inorganic perovskite solar cells[J]. Advances in New and Renewable Energy, 2019, 7(2): 142−148.

[5] 梁叔全, 程一兵, 方国赵, 曹鑫鑫, 沈文剑, 钟　杰, 潘安强, 周　江. 能源光电转换与大规模储能二次电池关键材料的研究进展[J]. 中国有色金属学报, 2019, 29(9): 2064−2114.
LIANG Shu-quan, CHENG Yi-bing, FANG Guo-zhao, CAO Xin-xin, SHEN Wen-jian, ZHONG Jie, PAN An-qiang, ZHOU Jiang. Research progress of key materials for energy photoelectric conversion and large-scale energy storage secondary batteries[J]. The Chinese Journal of Nonferrous Metals, 2019, 29(9): 2064−2114.

[6] BARTEL C J, SUTTON C, GOLDSMITH B R, OUYANG R H, MUSGRAVE C B, GHIRINGHELLI L M, SCHEFFLER M. New tolerance factor to predict the stability of perovskite oxides and halides[J]. Science Advances, 2019, 5(2): eeav0693.

[7] LI Chong-he, SOH K C K, WU Ping. Formability of $ABO_3$ perovskites[J]. Journal of Alloys and Compounds, 2004, 372(1/2): 40−48.

[8] ZHANG Huan, LI Na, LI Ke-yan, XUE Dong-feng. Structural stability and formability of $ABO_3$-type perovskite compounds[J]. Acta Crystallographica Section B: Structural Science, 2007, 63(6): 812−818.

[9] LI Chong-hea, LU Xiong-gang, DING Wei-zhong, FENG Li-ming, GAO Yong-hui, GUO Zi-ming. Formability of ABX3 (X=F, Cl, Br, I) Halide Perovskites[J]. Acta Crystallographica Section B: Structural Science, 2008, 64(6): 702−707.

[10] KUMAR A, VERMA A S, BHARDWAJ S R. Prediction of formability in perovskite-type oxides[J]. The Open Applied Physics Journal, 2008, 1: 11−19.

[11] EMERY A A, WOLVERTON C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of $ABO_3$ perovskites[J]. Scientific Data, 2017, 4: 170153.

[12] MOOSER E, PEARSON W B. On the crystal chemistry of normal valence compounds[J]. Acta Crystallographica, 1959, 12: 1015−1022.

[13] GOLDSCHMIDT V M. Die Laws of Crystal Chemistry[J]. The Science of Nature, 1926, 14(21): 477−485. (in German)

[14] 曾　佳, 汪　浩, 朱满康, 严　辉. 钙钛矿氧化物的化学结构及其催化性能的研究进展[J]. 材料导报, 2007(1): 33−36.
ZENG Jia, WANG Hao, ZHU Man-kang, YAN Hui. Progress of study on chemical structure and catalytic properties of the perovskite-type oxides[J]. Materials Reports, 2007(1): 33−36.

[15] 刘文兵, 李　亮, 刘桂成, 王新东. 钙钛矿太阳能电池稳定性研究进展[J]. 有色金属科学与工程, 2017, 8(2): 31−42.
LIU Wen-bing, LI Liang, LIU Gui-cheng, WANG Xin-dong. Research progress on stability of perovskite solar cells[J]. Nonferrous Metals Science and Engineering, 2017, 8(2): 31−42.

[16] 向　勇, 谢道华. $ABO_3$ 型氧化物的结构与性能及其应用[J]. 材料工程, 2000(9): 15−18.
XIANG Yong, XIE Dao-hua. Structure and characteristic of $ABO_3$-type oxide and its application[J]. Journal of Materials Engineering, 2000(9): 15−18.

[17] 姚　鑫, 丁艳丽, 张晓丹, 赵　颖. 钙钛矿太阳电池综述[J]. 物理学报, 2015, 64(3): 038805.
YAOXin, DING Yan-li, ZHANG Xiao-dan, ZHAO Ying. A review of the perovskite solar cells[J]. Acta Physica Sinica, 2015, 64(3): 038805.

[18] BALACHANDRAN P, EMERY A, GUBERNATIS J, LOOKMAN T, WOLVERTON C, ZUNGER A. Predictions of new $ABO_3$ perovskite compounds by combining machine learning and density functional theory[J]. Physical Review Materials, 2018, 2(4): 043802.

[19] PILANIA G, BALACHANDRAN P V, GUBERNATIS J E, LOOKMAN T. Classification of ABO3 perovskite solids: A machine learning study[J]. Acta Crystallographica Section B:

Structural Science, Crystal Engineering and Materials, 2015, 71(5): 507−513.

[20] OUYANG R H, CURTAROL S, AHMETCIK E, SCHEFFLER M, GHIRINGHELLI L M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates[J]. Physical Review Materials, 2018, 2(8): 083802.

[21] BREIMAN L I, FRIEDMAN J H, OLSHEN R A, STONE C J. Classification and regression trees[J]. Encyclopedia of Ecology, 2015, 57(3):582−588.

[22] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861−874.

# New tolerance factor based on SISSO and machine learning for predicting stability of perovskite structure

HU Hong-qing[1], WU Shao-gang[1], GUO Zhi-ting[1], ZHOU Gao-feng[1], DAI Dong-bo[1], WEI Xiao[1, 2], ZHANG Hui-ran[1, 2]

(1. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
2. Materials Genome Institute, Shanghai University, Shanghai 200444, China)

**Abstract:** Due to the wide application prospects of perovskite materials, research on their structures and physical and chemical properties of perovskite materials has been one of the hot topics in the field of materials research. Among them, predicting the stability of perovskite structure with the help of tolerance factor can help researchers discover more new functional materials. The conventional tolerance factor $t_{IR}$ for determining the stability of the perovskite structure based on ion radius has certain shortcomings and limitation. In view of this, this work proposes a new type of tolerance factor $\tau_{BV}$ based on the bond valence model using the SISSO (sure independence screening and sparsifying operator) method which can effectively avoid the defect limitation caused by the ionic radius. This work uses the decision tree algorithm in machine learning to establish the new tolerance factor verification model and the results show that the new tolerance factor $\tau_{BV}$ can excellently predict whether the $ABO_3$ compound is perovskite or non-perovskite, which greatly improves the prediction accuracy.

**Key words:** perovskites; structural stability; sure independence screening sparsifying operator (SISSO); new tolerance factor

**Foundation item:** Projects(2018YFB0704400) supported by the National Key Research and Development Program, China

**Received date:** 2019-08-19; **Accepted date:** 2019-12-06

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

**(编辑　何学锋)**