文章编号: 1000-324X(2022)12-1311-10 **DOI:** 10.15541/jim20220149

融合材料领域知识的数据准确性检测方法

施思齐[1,2,5], 孙拾雨[1], 马舒畅[3], 邹欣欣[3], 钱 权[3,4,5], 刘 悦[3,4,5]

(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; 3. 上海大学 计算机工程与科学学院, 上海 200444; 4. 上海大学 上海市智能计算系统工程技术研究中心, 上海 200444; 5. 之江实验室, 杭州 311100)

摘 要: 材料数据由于小样本、高维度、噪音大等特性, 用于机器学习建模时常常会产生与领域专家认知不一致的结果。面向机器学习全流程, 开发材料领域知识嵌入的机器学习模型是解决这一问题的有效途径。材料数据的准确性直接影响了数据驱动的材料性能预测的可靠性。本研究针对机器学习应用过程中的数据预处理阶段, 提出了融合材料领域知识的数据准确性检测方法。该方法首先结合材料专家认知构建了材料领域知识库。然后, 将其与数据驱动的数据准确性检测方法结合, 从数据和领域知识两个角度对材料数据集进行基于描述符取值规则的单维度数据正确性检测、基于描述符相关性规则的多维度数据相关性检测以及基于多维相似样本识别策略的全维度数据可靠性检测。对于每一阶段识别出的异常数据, 结合材料领域知识进行修正, 并将领域知识融入到数据准确性检测方法的全过程以确保数据集从初始阶段就具有较高准确性。最后该方法在 NASICON 型固态电解质激活能预测数据集上的实验结果表明: 本研究提出的方法可以有效识别异常数据并进行合理修正。与原始数据集相比, 基于修正数据集的 6 种机器学习模型的预测精度都有不同程度的提升。其中, 在最优模型上 R^2 提升了 33%。

关 键 词: 机器学习; 材料科学; 数据质量; 领域知识

中图分类号: TP181; O646; TB30 **文献标志码:** A

Detection Method on Data Accuracy Incorporating Materials Domain Knowledge

SHI Siqi[1,2,5], SUN Shiyu[1], MA Shuchang[3], ZOU Xinxin[3], QIAN Quan[3,4,5], LIU Yue[3,4,5]

(1. Materials Genome Institute, Shanghai University, Shanghai 200444, China; 2. School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China; 3. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; 4. Shanghai Engineering Research Center of Intelligent Computing System, Shanghai University, Shanghai 200444, China; 5. Zhejiang Laboratory, Hangzhou 311100, China)

Abstract: Due to the characteristics of small samples, high dimensions, and much noise, materials data often produce inconsistent results with those obtained from domain experts when used for machine learning modeling. For the whole process of machine learning, developing machine learning models embedding materials domain knowledge is a solution to this problem. The accuracy of materials data directly affects the reliability of data-driven materials performance prediction. Here, a data accuracy detection method incorporating materials domain knowledge is proposed by focusing on the data preprocessing stage in the machine learning application process. Firstly, a

收稿日期: 2022-03-21; **收到修改稿日期:** 2022-05-06; **网络出版日期:** 2022-05-27

基金项目: 国家重点研发计划(2021YFB3802101); 国家自然科学基金(52073169); 之江实验室科研攻关项目(2021PE0AC02)
National Key Research and Development Program of China (2021YFB3802101); National Natural Science Foundation of China (52073169); Key Research Project of Zhejiang Laboratory (2021PE0AC02)

作者简介: 施思齐(1978–), 男, 博士, 教授. E-mail: sqshi@shu.edu.cn
SHI Siqi (1978–), male, PhD, professor. E-mail: sqshi@shu.edu.cn

通信作者: 刘 悦, 教授. E-mail: yueliu@shu.edu.cn
LIU Yue, professor. E-mail: yueliu@shu.edu.cn

materials domain knowledge database is constructed based on the knowledge from materials experts. Secondly, it is coordinated with the data-driven data accuracy detection method to perform single-dimensional data accuracy detection based on the rule for value of descriptors, multi-dimensional data correlation detection based on the rule for correlation of descriptors, and full-dimensional data reliable detection based on multi-dimensional similar sample identification strategy from both data and domain knowledge perspectives. For the anomalous data identified at each stage, they are corrected by incorporating the materials domain knowledge. Furthermore, domain knowledge is incorporated into the whole process of the data accuracy detection method to ensure high accuracy of the dataset from the initial stage. Finally, experiments on the NASICON-type solid electrolyte activation energy prediction dataset demonstrate that this method can effectively identify anomalous data and make reasonable corrections. Compared with the original dataset, the prediction accuracy of all six machine learning models based on the revised dataset is improved to different degrees, among which R^2 achieves a 33% improvement on the optimal model.

Key words: machine learning; materials science; data quality; domain knowledge

机器学习(Machine Learning, ML)作为人工智能研究的一个分支, 是研究计算机如何模拟人类学习行为以自动获取知识或技能, 从而不断改善自身性能的一门学科[1]。近年来, 数据驱动的机器学习以其卓越的预测精度、效率和适用性, 在材料研究中得到了广泛应用[2-8]。然而, 数据驱动的机器学习方法的性能严重依赖于数据的质量[3]。目前, 大部分材料数据来自实验测量、理论计算以及工业生产[6-8], 因测量手段和实验方案不同、实验环境差异、理论计算误差以及数据录入错误等因素, 收集到的材料数据与真实数据之间往往存在一定误差, 使得材料数据的准确性降低, 进而严重影响机器学习模型预测结果的准确性和合理性[9]。因此, 在机器学习建模前对材料数据进行有效的准确性检测和提升至关重要。

近年来, 随着机器学习在材料领域的应用更加广泛, 研究人员开始意识到数据准确性的重要性, 并采用人工检测、简单统计方法以及分类算法来对数据准确性进行评估[10-11]。例如, Beal 等[10]在采用人工神经网络(Artificial Neural Network, ANN)对钙钛矿的离子电导率进行预测时, 利用基于距离的异常点检测算法识别并删除了远离样本中心的样本点, 藉此构建了拟合优度 R^2 高达 0.92 的机器学习模型。Gharagheizi 等[11]在使用最小二乘支持向量机(Least square SVM, LSSVM)对离子液体的电子电导率进行预测时, 对实验测量的电子电导率数据进行人工校验, 发现并剔除了 100 个可疑数据, 最终构建的机器学习模型的平均预测偏差仅为 1.9%。Hemmati-Sarapardeh 等[12]在利用 LSSVM 算法对离子液体的密度进行预测时, 联合杠杆检测法、William 图、Hat 矩阵来识别异常数据, 成功检测到了 6 条由于文献资料错误报道或实验测量误差导致的异常样本, 最终 LSSVM 算法的预测精度达到 98.86%。类似地, Hosseinzadeh 等[13]也采用杠杆检测法对文献中离子液体电子电导率实验数据进行了准确性评估, 共发现了 7 个异常样本点, 机器学习模型的预测精度达到了 99.98%。上述实验结果表明, 提高数据的准确性能够显著提升机器学习模型的预测精度。然而, 上述研究均采用传统的单一数据准确性检测方法对学习样本的准确性进行检测, 下列两方面问题仍然需要进一步探索和研究。一方面, 现有的数据准确性检测方法解决问题的角度各有不同, 利用单一的方法难以全面评估数据准确性。例如, 常见的基于统计分析的 3σ 探测法虽然快速易实现, 但是它假设数据服从正态分布, 且仅适用于单维数据的准确性检测。基于距离的 K 最近邻方法虽然实现简单且无须估计样本分布, 但结果容易受参数影响。另一方面, 材料数据本身具有很强的专业性, 现有的数据准确性检测方法仅从数据分布的角度去衡量数据正确与否, 往往会忽略领域知识在数据准确性检测中的重要性, 从而难以发现有悖于领域知识的异常数据。例如, 在用于机器学习建模的材料数据集中, 描述符之间往往存在某种关联关系, 仅仅从数据分布特性出发对数据准确性进行检测难以发现有悖于描述符之间相关关系的异常数据。如果能在专家经验或领域知识的协助下, 通过集成不同的数据准确性检测方法, 对材料领域的学习样本进行多角度的准确性检测, 将有助于提升数据准确性检测的全面性。

本课题组近期研究也对材料数据中存在的数据集质量问题进行了探讨, 特别指出了材料领域知识嵌入机器学习全流程的必要性和有效性[14-15]。基于此, 本研究提出融合材料领域知识的数据准确性检测方法(Data Accuracy Detection Method Incor-

porating Materials Domain Knowledge, $DADM_{mdk}$)。该方法高度重视维度间的相关性和专家经验在数据准确性检测中的作用, 综合考虑了目前已有的数据准确性检测方法的优缺点, 对材料数据的准确性进行全面且高效的检测, 从而为后续的机器学习建模提供高质量的学习样本。首先, 结合材料专家经验构建包括描述符取值规则、描述符相关性规则以及多维相似样本识别策略在内的专家经验知识库。然后, 联合专家经验知识库与数据驱动的数据准确性检测方法, 依次对数据集进行正确性、相关性和可靠性检测, 从数据和领域知识两个角度检测和提升材料数据的准确性。最后, 以 NASICON 型固态电解质激活能预测数据集为例来说明所提方法的有效性。

本研究首先介绍了描述符知识的抽取和符号化表示, 并提出了融入描述符知识的数据准确性检测算法; 然后在 NASICON 型固态电解质激活能预测数据集上进行验证, 证明了 $DADM_{mdk}$ 的有效性和可行性。

1　描述符知识的抽取和符号化

在材料领域, 表征材料结构、成分等物理/化学性质的物理/化学因素被称为描述符。描述符是用于机器学习建模的材料数据的关键组成部分, 决定了机器学习模型学习能力的上限。目前, 已有大量工作为不同材料体系的广泛性能提供了描述符的选择和量化方法[16-17], 并积累了大量描述符相关领域的知识。基于这些材料领域知识, 本研究提出了基于描述符取值规则、描述符相关性规则以及多维相似样本识别策略的数据准确性检测方法, 以全面评估材料数据的准确性。

1.1　描述符取值规则

每个材料描述符都具有特定的物理含义、数据类型、取值范围、数据来源和量化计算方法(补充材料 S1)。对于某一特定样本, 需要判断其描述符的数据类型、取值范围是否与领域知识一致。此外, 某些描述符还可由其他描述符根据计算公式计算得到。因此, 为了避免计算错误造成数据异常, 还需要根据计算公式对其进行检测。定义描述符取值规则如下:

规则 1: 给定三元组 $<D, T, R>$, 其中 $D=\{\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_n\}$, n 为描述符总数且 $\vec{d}_i(i=1,\cdots,n)$ 是 D 中第 i 个描述符的取值向量; $T=\{t_1,t_2,\cdots,t_n\}$ 且 $t_i(i=1,\cdots,n)$ 是第 i 个描述符的数据类型; $R=\{r_1,r_2,\cdots,r_n\}$ 且 $r_i=\left\langle r_i^{\min},r_i^{\max}\right\rangle$ 是第 i 个描述符的经验取值范围, $r_i^{\min}$ 和 $r_i^{\max}$ 分别为经验取值的最小值和最大值。令 m 为样本总数, 给定任意描述符取值 $d_i^j(j=1,\cdots,m)$, 可根据公式(1)判断其是否为潜在异常数据点:

$$\mathrm{Ind1}(d_i^j)=\begin{cases}0, \text{If } \mathrm{type}(d_i^j)=t_i \text{ and } r_i^{\min}\leqslant d_i^j\leqslant r_i^{\max}\\ 1, \text{Otherwise}\end{cases} \tag{1}$$

其中, type(·) 表示获取任意数据点的数据类型, 1 和 0 分别表示 d_i^j 是或不是潜在异常数据点。若 d_i^j 的数据类型 $\mathrm{type}(d_i^j)$ 与所属描述符的经验取值类型 t_i 不匹配或 d_i^j 超出所属描述符的经验取值范围 $[r_i^{\min}, r_i^{\max}]$, 则判断其为潜在的异常数据点。

规则 2: 给定二元组 $<D,F>$, 其中 $D=\{\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_n\}$, n 为描述符总数且 $\vec{d}_i(i=1,\cdots,n)$ 是 D 中第 i 个描述符的取值向量; $F=\{f_1, f_2,\cdots,f_n\}$ 且 $f_i(i=1,\cdots,n)$, 若描述符是由特定公式计算得到的, 那么 f_i 为相应的计算公式, 若描述符不是通过公式计算得到的, 则 f_i 为常数 0。令 m 为样本总数, 给定任意描述符取值 $d_i^j(j=1,\cdots,m)$, 可根据公式(2)判断其是否为潜在异常数据点:

$$\mathrm{Ind2}(d_i^j)=\begin{cases}0, \text{If } (f_i=0) \text{ or } (f_i\neq 0 \text{ but } d_i^j=f_i(d_i^j))\\ 1, \text{Otherwise}\end{cases} \tag{2}$$

其中, 1 和 0 分别表示 d_i^j 是或不是潜在异常数据点, $f_i(\cdot)$ 表示根据计算公式 f_i 重新计算得到描述符值。若通过公式 f_i 得到 d_i^j, 但 $f_i(d_i^j)$ 与 d_i^j 不相等, 则判断其为潜在的异常数据点。

1.2　描述符相关性规则

在表征材料性质的描述符中, 某些描述符具有相似或相同的物理意义, 可能对材料的性能具有相似或相同的影响机理, 即这些描述符之间具有一定的相关关系, 这种关系称为“描述符相关性规则”例如, 在 NASICON 型固态电解质材料中, 来自材料领域知识的描述符相关规则如下:

在空间群为 R-3c 的 NASICON 型固态电解质中, 晶格常数(a, b, c)和晶胞体积 V_{cell} 均可以表征晶胞大小。晶格常数 a 等于晶格常数 b, 且与单元胞体积 V_{cell} 的关系如公式(3~5)所示。

$$V_{\mathrm{cell}}=\frac{\sqrt{3}}{2}a^2c \tag{3}$$

$$a\propto V_{\mathrm{cell}} \tag{4}$$

$$c\propto V_{\mathrm{cell}} \tag{5}$$

其中, $\propto$ 表示正相关关系。

文献[18-19]研究了迁移离子浓度对激活能的影响, 实验结果表明迁移离子浓度增大会降低激活能, 如公式(6)所示:

$$C\_\mathrm{Na} \propto -E_\mathrm{a} \tag{6}$$

其中, C_Na 是钠的浓度。E_a 是激活能, 在 NASICON 型固态电解质材料中通常作为材料性能(其余相关性规则见补充材料 S2)。

将以上从文献中得到的描述符与描述符或描述符与激活能之间的相关关系规则作为 $\mathrm{DADM_{mdk}}$ 方法中维度间相关性检测结果的判断准则, 可从领域知识的角度保证维度间相关关系的准确性。上述描述符相关性规则形式化表示如下:

规则 3: 给定二元组 $<D,C>$, 其中, $D=\{\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_n\}$ n 为描述符总数且 $\vec{d}_i(i=1,\cdots,n)$ 是 D 中第 i 个描述符的取值向量; $C=\{c_1,c_2,\cdots,c_k\}$, $c_p=\langle d_i,d_i,\mathrm{cor}(d_i,d_i)\rangle(p=1,\cdots,k)$ 表示从材料领域知识获取的描述符间相关性规则, $\mathrm{cor}(d_i,d_j)$ 表示描述符 d_i 和 d_j 间的相关性(正相关或负相关), 其形式化表达如公式(7~8)。

$$\mathrm{cor}(d_i,d_j)=k,k=\pm1 \tag{7}$$

$$\mathrm{Ind3}(d_i,d_j)=\begin{cases}0, \dfrac{R(d_i,d_j)}{|R(d_i,d_j)|}=\mathrm{cor}(d_i,d_j)\\ 1, \text{Otherwise}\end{cases} \tag{8}$$

其中, 1 和 0 分别表示 d_i 或 d_j 是或不是潜在异常描述符, $R(\cdot)$ 表示通过数据驱动的相关性分析方法得到的任意两个描述符 d_i 和 d_j 间的相关系数。k 等于 1 或–1 分别表示描述符 d_i 和 d_j 之间存在正相关或负相关关系。当 k 等于 1 (–1)但 $R(\cdot)$ 小于(大于) 0 时, 说明通过数据驱动的相关性分析方法获得的描述符间的相关性与领域知识不一致, 需要重新确定描述符的获取和计算方式以对数据集进行修正。

1.3 多维相似样本识别策略

材料性能受成分、结构、实验条件和环境等多种因素影响。实验条件和环境的影响有时表现为材料结构不同, 进而影响性能。在理想情况下, 成分和结构相似的材料具有相似的性能。分别对表征材料结构和成分的描述符(即特征)同材料性能进行聚类, 以识别潜在的异常样本。当以特征为对象的聚类结果与以材料性能为对象的聚类结果属于同一类别或相邻类别时, 说明相似样本具有相似材料性能, 这些样本被判别为正确; 当以特征为对象的聚类结果不属于与材料性能相似的聚类时, 该样本被判别为异常。基于此, 将成分和结构相似但材料性能相差较大的样本点视为潜在的异常样本。将这种异常样本检测方法称为“相似样本识别策略”, 形式化表达如规则 4 所示:

规则 4: 给定四元组 $<F,T,C_F,C_T>$, F 为所有样本在特定描述符上的取值集合, T 为所有样本的材料性能数据, $C_F=\{c_f^1,c_f^2,\cdots,c_f^k\}$ 和 $C_T=\{c_t^1,c_t^2,\cdots,c_t^k\}$ 分别表示特征数据 F 和材料性能数据 T 上的聚类结果, k 为聚类个数。对于给定样本 $S_j(j=1,\cdots,m)$, 可根据公式(9)判断其是否为潜在的异常样本:

$$\mathrm{Ind4}(S_j)=\begin{cases}0, |c_f^k\cap c_t^k|<3\\ 1, \text{Otherwise}\end{cases} \tag{9}$$

其中, $|\cdot|$ 表示集合中元素个数, c_f^k 和 c_t^k 分别表示任意样本 S_j 在 F 和 T 上聚类结果所属集合。对于 S_j 所在集合 c_f^k 和集合 c_t^k, 如果其交集 $c_f^k\cap c_t^k$ 中样本个数小于 3, 说明和 S_j 在特征上聚为一类的样本在材料性能上并未聚为一类甚至相差较远。因此, S_j 是一个潜在的异常样本。

2 融入描述符知识的数据准确性检测算法

融合材料领域知识的数据准确性检测方法 $\mathrm{DADM_{mdk}}$ 的总体框架如图 1 所示, 基于第 1 节抽取的符号化描述符专家经验知识, 协同数据驱动的基于统计分析的单维度数据准确性检测、基于统计分析和信息论的维度间相关性分析以及基于机器学习的全维度数据可靠性检测, 在材料领域知识和数据的双向驱动下, 开展全面、合理且高效的准确性检测与提升(补充材料 S3)。整个流程包括基于描述符取值规则的单维度数据正确性检测、基于描述符相关性规则的多维度数据相关性检测和基于多维相似样本识别策略的全维度数据可靠性检测三个阶段:

2.1 第一阶段: 基于描述符取值规则的单维度数据正确性检测

在单维数据中, 规定某些描述符在维度上具有不合理的数据类型和取值范围的数据为异常数据。单维数据正确性检测对单维数据是否合理进行衡量。基于统计分析的数据准确性检测方法是最早适用于单维数据的异常点检测方法, 其基本思想是根据数据的特性给定其服从某个分布的概率模型, 并

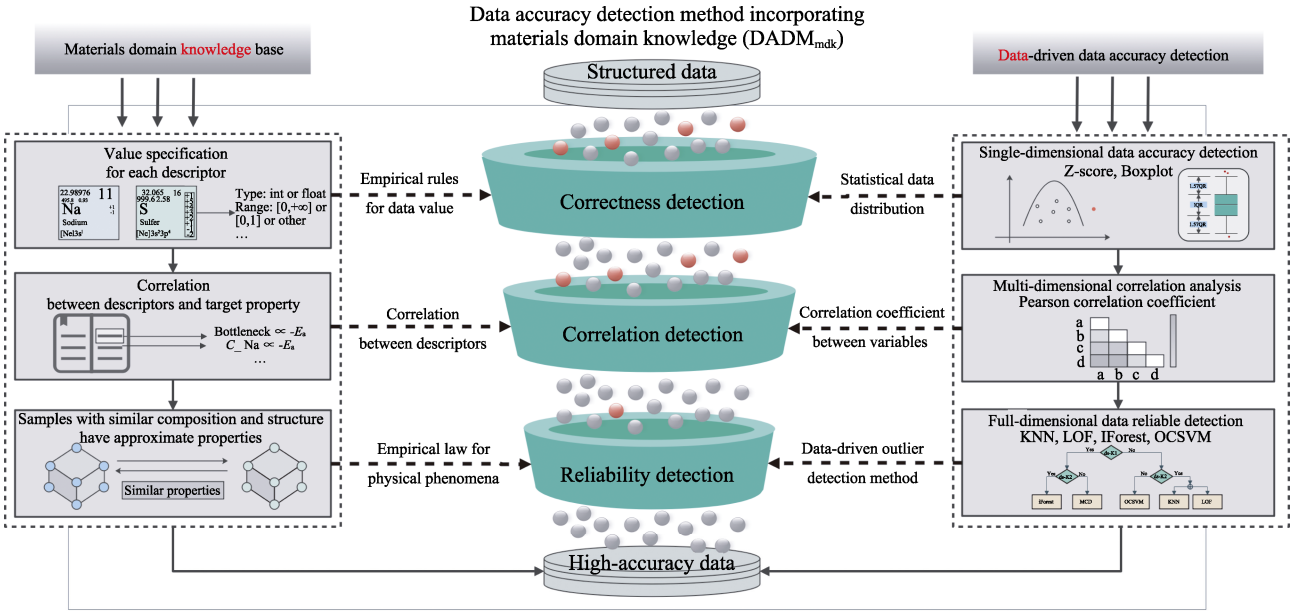

图 1　$DADM_{mdk}$ 框架

Fig. 1　Framework of $DADM_{mdk}$

根据该概率模型对每个数据点的置信程度进行评估，从而确定可能异常的数据点。对于单维度数据，现有的基于统计分析的异常点检测方法有 3σ 探测法，Z-score[20]，箱线图[21]。这些方法仅从数据分布的角度对数据的准确性进行衡量，忽略了领域知识在数据准确性检测中的重要性，可能难以发现与领域知识不一致的异常数据。因此，考虑到材料科学数据本身的一致性，将 1.1 节得到的取值规则融入到数据驱动的准确性检测过程中，基于描述符取值规则对单维度数据进行正确性检测，并根据材料领域知识对异常点进行修正，再选择合适的统计分析方法对单维度数据进行二次检测。

2.2　第二阶段：基于描述符相关性规则的多维度数据相关性检测

单维数据的正确性检测在一定程度上保证了单维数据的准确性。然而，无法保证维度之间的准确性。维度间相关性的准确性会影响后续特征选择的结果，进而影响模型预测精度。因此，提升材料数据的准确性依赖于维度间正确的相关关系。对于多维数据间的相关性，现有的检测方法有很多。例如，对于线性关系，通常使用皮尔逊相关系数[22] (Pearson Correlation Coefficient, PCC)，斯皮尔曼相关系数[23] (Spearman Correlation Coefficient, SCC)进行维度间相关性的检测。与 PCC 相比，SCC 对于数据错误和极端值不敏感，使用不频繁。对于非线性关系，通常使用距离相关系数[24] (Distance Correlation Coefficient, DCC)，最大互信息系数 (Maximal Information Coefficient, MIC)评估维度间的相关关系。为了识别维度间的异常数据，将描述符相关性规则作为评估多维数据相关性的判断标准，对相关性检测方法中识别出的具有一定相关性(如正相关和负相关)的二元描述符组合进行检测，将相关性与描述符相关规则相悖的描述符标记为可能的异常数据，并提交给专家进一步分析。

2.3　第三阶段：基于多维相似样本识别策略的全维度数据可靠性检测

经过上述单维度数据正确性检测和多维度数据相关性检测之后，材料数据在每个维度上的准确性已经达到了较高水平。但是，这并不能保证每一条包含多个维度样本的准确性。现有多维数据异常点检测方法可用于解决这一问题，包括 K 最近邻 (K-Nearest neighbors, KNN)、孤立森林[25] (Isolation Forest, IForest)、局部异常因子[26] (Local Outlier Factor, LOF)、单类别支持向量机(One Class Support Vector Machine, OCSVM)和协方差最小行列式[27] (Minimum Covariance Determinant, MCD)等方法。本节根据这些异常点检测方法解决问题的不同角度、应用范围和优缺点(补充材料表 S4)，得到了如图 2 所示的数据驱动的全维度数据可靠性检测方法选择策略。

当数据量较大(n 大于 K_1)且维度较高(特征数 d 大于 K_2)时，选择 MCD；否则，选择 IForest。当数据量较小(n 小于或等于 K_1)而数据维度较高(d 大于 K_2)时，采用 OCSVM；否则，使用 KNN 或 LOF。因此，

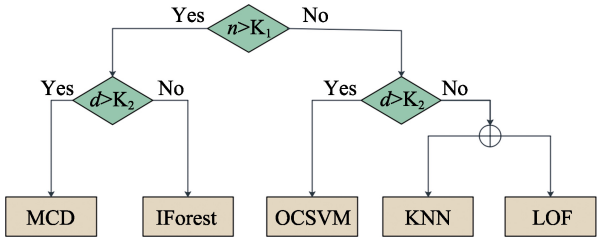

图 2 全维度可靠性检测

Fig. 2 Full-dimensional reliability detection

n: Number of learning samples; *d*: Number of features

为了保证更深层次的数据准确性, 进一步基于多维相似样本识别策略对全维度数据进行可靠性检测, 根据聚类结果将不符合预先定义条件的样本识别为异常样本, 并根据材料领域知识对异常样本进行分析修正, 再根据数据集规模选择合适的异常点检测方法对全维度数据进行二次检测。

上述三阶段的详细过程如算法 S1 所示。经过上述单维度数据正确性检测、多维度数据相关性检测和全维度数据可靠性检测三个阶段, 可以实现对异常数据点或样本点的有效剔除或合理修正以及对违反材料领域知识的描述符的重新获取和计算, 得到单维度有较高正确性、维度间相关性正确且无异常样本的高质量数据集。

3 实验结果与分析

3.1 实验数据

实验数据集是来自无机晶体结构数据库(Inorganic Crystal Structure Database, ICSD)的 90 个 NASICON 型固态电解质激活能构效关系数据, 包含 45 个描述符(即特征)和 1 个材料性能(即激活能)。激活能是利用固态电解质高通量筛选平台[28]上的键价位能方法计算得到的。

3.2 实验设置

考虑到实验数据集小样本、高维度的特点, 在单维度数据正确性检测中采用 Z-score 和箱线图法进行异常点识别。在箱线图中, 上下边缘的数据区间 IQR 设置为 1.5, 即超过 5%和 95%的数据区间被识别为异常点。本研究重点关注线性关系, 在多维度数据相关性检测中采用 PCC 评估维度间的相关性。在全维度数据可靠性检测中, 采用 K-Means 聚类算法[29]分别对特征和激活能数据进行聚类, K-Means 聚类个数设置为 8, 样本大小和维度阈值 K_1 和 K_2 分别设置为 250 和 25。

本研究选取套索回归(Least Absolute Shrinkage and Selection Operator, LASSO)、高斯过程回归(Gaussian Process Regression, GPR)、岭回归(Ridge Regression, Ridge)、支持向量回归(Support Vector Regression, SVR)、K-近邻回归(K-Nearest Neighbor Regression, KNN)和随机森林(Random Forest, RF)等 6 种材料领域广泛使用的机器学习模型为候选模型, 分别在原始数据集和修正数据集上建立激活能预测模型, 并使用十折交叉验证来评估模型性能。模型预测精度的评估指标主要采用均方根误差(Root Mean Square Error, RMSE)、平均绝对百分误差(Mean Absolute Percentage Error, MAPE)和拟合优度(*R*-square, R^2)。

3.3 三阶段数据准确性检测和修正过程

本节根据第 2 节提出的 $DADM_{mdk}$ 框架依次对数据集进行单维度数据正确性检测、多维度数据相关性检测和全维度数据可靠性检测, 并将未融入领域知识的纯数据驱动的数据准确性检测方法和 $DADM_{mdk}$ 方法进行对比, 实验结果如表 1 所示。

由表 1 可知, 与未融入领域知识的纯数据驱动的准确性检测方法相比, 本研究提出的 $DADM_{mdk}$ 方法在大多数准确性检测阶段都能够有效识别异常数据并进行合理修正。这说明领域知识和数据驱动相结合能够对数据集进行全面而高效的准确性检测与提升。

3.3.1 第一阶段实验结果与分析

本节基于 1.1 节给出的描述符取值规则和所有

表 1 数据驱动的数据准确性检测方法与 $DADM_{mdk}$ 结果比较

Table 1 Result comparison of data-driven data accuracy detection methods with $DADM_{mdk}$

Stage	*N*	Data-driven data accuracy detection				*N*1	$DADM_{mdk}$						*N*1
		Number of anomalous		Number of removed			Number of anomalous		Number of removed		Number of corrected		
		D	*S*	*D*	*S*		*D*	*S*	*D*	*S*	*D*	*S*	
1	90	18	0	0	5	85	18	2	0	5	0	0	85
2	85	0	0	0	0	85	0	0	0	0	0	0	85
3	85	0	9	0	0	85	0	9	0	0	0	3	85

N: Original sample size; *N*1: Retained sample size; *D*: Dimension; *S*: Sample

描述符的取值范围、数据类型等信息(补充材料表 S1)对单维度数据进行判断, 发现有 2 个样本存在异常点(如表 2 所示)。一般来说, Valence_M1 为整型, 因此第 17 和 72 条样本中 Valence_M1 被识别为异常数据。进一步分析它们的化学结构, 发现样本本身没有问题。在这两个样本中, 数值存在小数的原因是材料中 M1 位置的原子存在不同的价态, 因此, 可以保留这两个样本。

随后, 采用基于正态分布假设的Z-score方法进行异常点判别。在输出结果中, P 大于显著性水平 0.05, 说明样本的总体服从正态分布。在此基础上, 再使用箱线图法进行异常点识别。如图 3(a)所示, 共有 Radius_X1、E_a、Occu_X1 等在内的 18 个维度被识别出了异常点(补充材料 S4)。

表 2　单维度数据正确性检测中的异常点

Table 2　Anomalous points in single-dimensional data correctness detection

No.	ICSD	Formula	Valence_M1
17	182793	$Na_{16.74}Cr_{12}P_{18}O_{72}$	3.105
72	71326	$Na_3Nb_{12}P_{18}O_{72}$	4.25

根据相关材料领域知识, 只有少数异常点为真正异常, 其他异常点是由于原始数据集规模较小而多样性较高而产生的误识别。例如, 在图 3(b)所示的 Radius_X1 上, ICSD 编号为 35770 的样本被识别为异常点, 这是因为该样本 X1 位点上的原子半径比其他样本都大。在图 3(c)所示的 Occu_X1 上,

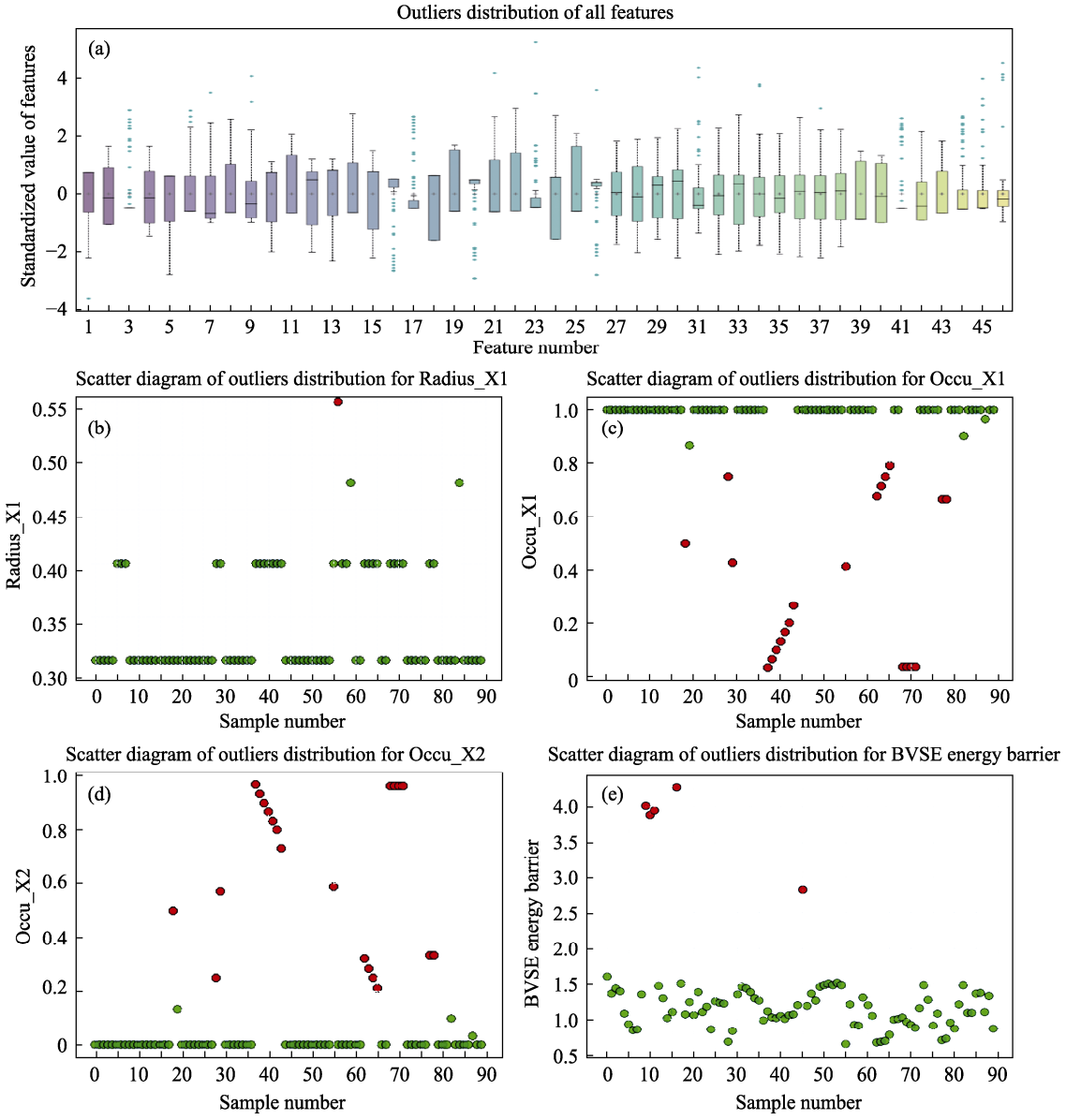

图 3　异常点检测结果

Fig. 3　Anomalous points detection results

(a) All anomalous points from the box plot; Anomalous points in (b) Radius_X1, (c) Occu_X1, (d) Occu_X2, (e) E_a

Colorful figures are available on website

Occu_X1 的值大都为 1，故少数其他值被识别为异常点。此外，由于特征 Occu_X1 和 Occu_X2 之间存在直接的关联关系，因此二者的异常点对应的样本是相同的，如图 3(c, d)所示。

对于材料性能数据 E_a，5 个样本由于值太大被识别为异常样本，如图 3(e)所示。进一步验证后发现，在计算这 5 个样本的激活能时，由于原子混占，激活能计算程序对于原子占位处理错误，导致激活能出现较大的计算误差，因此将这 5 个样本从数据集中删除。

3.3.2 第二阶段实验结果与分析

维度间的关联关系已经在第 1.2 节中做了详细介绍，通过 PCC，本节得到 NASICON 型固态电解质激活能预测数据集中各维度间的相关性，如图 4 所示。由图可知，晶格参数 a 和 c 与体积 V 呈正相关；min_BT, min_T, BT1, BT2, $V\_Na_1O_6$, C_Na 与 E_a 呈负相关。这些结果与公式(4~6)一致，本阶段未识别出异常特征。

3.3.3 第三阶段实验结果与分析

K-Means 聚类的结果如图 5 所示。其中，每个子图中彩色点表示特征聚类得到的 8 个类簇 (c_f^k)；所有数值点则表示激活能聚类的 8 个类簇 (c_t^k)。

聚类结果用来帮助专家检测可能的异常样本。以图 5(d)为例，特征和激活能均聚为同一类，验证了结构和成分相似的样本往往具有相似的激活能。在每个图中，当特征聚类的结果与其对应的激活能聚类的结果被近似归为一类时，这些样本被判别为正确。例如，在图 5(c)中没有发现异常样本。当特征聚类结果与激活能聚类结果相差较大且同一颜色的点个数小于 3 时，判定为异常样本，并交由材料专家检测。以图 5(a)为例，将第 19、23、56 条样本识别为异常样本(其余见补充材料 S5)。

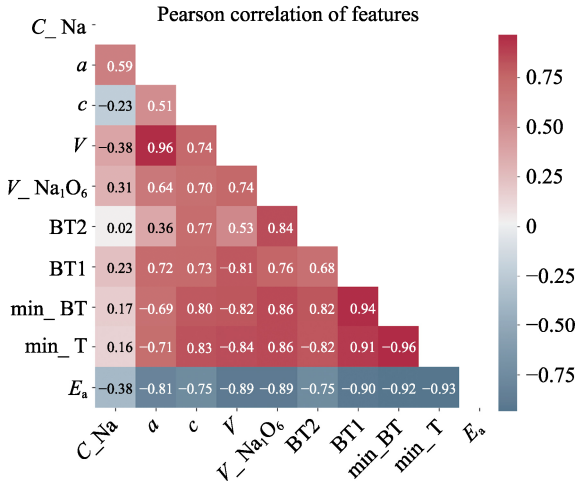

图 4 PCC 结果

Fig. 4 Result of PCC

The color bands represents the mapping of values to colors; The darker the color (red or cyan), the higher the correlation, and *vice versa*.

在标记为异常的样本中，发现了 3 个存在错误的样本(表 3)。经专家检测发现这 3 个样本在输入数据库的过程中，发生了输入错误。结果表明，CIF 文件(ICSD_15545, ICSD_15546, ICSD_15547)中的晶格常数错误，导致单元体积、多面体体积、瓶颈和激活能数据计算错误。对这 3 条样本进行修改并更新数据集。

由于学习样本较少(85 条)，而维数较多(45 维)，所以采用 OCSVM 方法进行异常点检测(补充材料 S6)。共检测出 9 条样本存在异常，将这些异常样本的 CIF 文件提交给材料专家进行核实，未发现异常。进一步检查发现，这几个样本异常均是数据分布不均匀导致的，因此不进行处理。最后，得

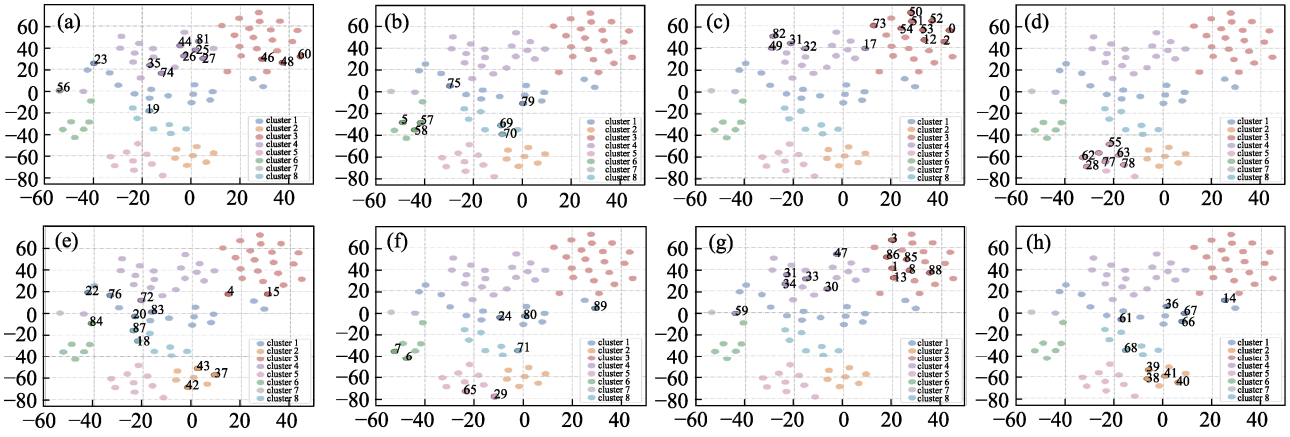

图 5 特征和激活能的聚类结果

Fig. 5 Clustering results of feature and activation energy

(a-h) Clustering overlaps of features and activation energy for each of 8 clusters; Horizontal and vertical coordinates represent two dimensions after dimension reduction by *t*-SNE (*t*-distributed Stochastic Neighbor Embedding)

Colorful figures are available on website

表 3　异常点检测及修正

Table 3　Anomalous data detection and correction

No.	ICSD	Formula	a	c	V_{cell}	Revised a	Revised c	Revised V_{cell}
5	15545	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	9.198	22.210	1627.29
6	15546	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	9.199	22.470	1646.70
7	15547	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	9.199	22.706	1663.99

表 4　机器学习模型实验结果

Table 4　Experimental results of ML models

Model	Origin data			Revised data		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
LASSO	0.621	0.218	0.118	0.058	0.035	0.943
GPR	0.637	0.231	0.073	0.051	0.032	0.957
Ridge	0.548	0.244	0.313	0.051	0.033	0.956
SVR	0.638	0.102	0.072	0.071	0.057	0.916
KNN	0.624	0.226	0.108	0.079	0.051	0.894
RF	0.400	0.088	0.629	0.051	0.035	0.956

到包含 85 条样本的修正数据集。

3.4　基于数据准确性检测后数据的激活能预测实验结果与分析

为了验证 $DADM_{mdk}$ 方法的有效性, 分别采用 LASSO、GPR、Ridge、SVR、KNN 和 RF 模型对激活能进行预测。在原始和修正的数据集上, 6 个模型数据的 10 次平均 RMSE、MAPE 和 R^2 如表 4 所示。RMSE 和 MAPE 的值越小, 模型的预测性能越好。与原始数据集相比, NASICON 型固态电解质激活能预测数据集的修正数据集上每个模型的预测精度都得到了显著提高, 在修正数据集上最优模型 RF 的 R^2 提高了 33%, 达到了 0.96。这表明修正后的数据集是合理且有效的, 进一步说明使用 $DADM_{mdk}$ 能够对数据集进行全面检测以提高其准确性, 证明了 $DADM_{mdk}$ 的可行性和高效性。

4　总结与展望

现有数据准确性检测方法仅仅从数据特性出发, 往往忽略了专家经验在数据准确性检测中的重要性, 难以发现学习样本中有悖于领域知识的异常数据。针对此问题, 本研究提出了融合材料领域知识的数据准确性检测方法 $DADM_{mdk}$。该方法首先基于材料专家对固态电解质材料的输运机制研究、洞察与理解, 构建了包括描述符取值规则、描述符相关性规则以及多维相似样本识别策略在内的专家经验知识库。然后, 根据每一阶段准确性检测的具体内容, 建立数据驱动的数据准确性检测方法。最后, 在每个准确性检测阶段, 将知识库的相关规则或策略与相应的数据准确性检测方法进行结合, 从数据和专家经验两个角度出发对学习样本依次进行正确性检测、相关性检测和可靠性检测。在 NASICON 型固态电解质激活能预测数据集上进行的准确性检测实验结果表明, 该方法可以有效地识别学习样本中可能存在的异常数据, 并结合材料领域知识剔除了其中存在明显错误的 5 个样本, 修正了由于输入错误导致异常的 3 个样本。与原始数据集相比, 在经过 $DADM_{mdk}$ 方法检测和修正的高质量 NASICON 数据集上, 最优模型的 R^2 提高了 33%, 达到 0.96。

本研究综合考虑了材料数据集单个维度上数据的正确性、维度间的相关性以及样本间关系的可靠性, 通过将材料领域知识和数据驱动的统计分析方法相结合, 有效识别和合理修正了数据集中可能存在的异常数据。今后, 可将该方法推广至硫银锗矿、石榴石等其他材料体系进行准确性检测与提升, 帮助材料专家了解数据集的整体特征, 进而识别数据集中可能存在的异常数据, 以提高数据集的整体质量。

然而, 除直接来自实验测量或计算模拟或数据库的数据外, 材料领域还存在大量非结构化文本数据。随着文本挖掘(Text Mining, TM)和自然语言处理(Natural Language Processing, NLP)的发展, 已有大量研究表明 TM 和 NLP 技术能够从科学文本中大规模地提取数据[14]。但由于材料文献本身含有大量公式、化学式等符号, 可能导致 TM 和 NLP 提取的非结构化文本数据存在乱码、乱序等错误, 如果能对这些错误进行准确性检测并修正或剔除, 将大大

提高文本数据的质量, 从而为后续实体抽取、知识图谱构建提供高质量数据集, 进一步推动对蕴含更多材料领域知识的非结构化文本数据的研究。

补充材料:

本文相关补充材料可登陆 https://doi.org/10.15541/jim20220149 查看。

参考文献:

[1] MURPHY K P. Machine learning: a probabilistic perspective. Cambridge: MIT Press, 2012.

[2] LIU Y, GUO B R, ZOU X X, *et al*. Machine learning assisted materials design and discovery for rechargeable batteries. *Energy Storage Materials*, 2020, **31:** 434–450.

[3] LIU Y, ZHAO T L, WU J M, *et al*. Materials discovery and design using machine learning. *Journal of Materiomics*, 2017, **3:** 159–177.

[4] GUBERNATIS J E, LOOKMAN T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Physical Review Materials*, 2018, **2(12):** 120301.

[5] RAMPRASAD R, BATRA R, PILANIA G, *et al*. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 2017, **3:** 54.

[6] KATCHO N A, CARRETE J, REYNAUD M, *et al*. An investigation of the structural properties of Li and Na fast ion conductors using high-throughput bond-valence calculations and machine learning. *Journal of Applied Crystallography*, 2019, **52:** 148–157.

[7] NAKAYAMA M, KANAMORI K, NAKANO K, *et al*. Data-driven materials exploration for Li-ion conductive ceramics by exhaustive and informatics-aided computations. *Chemical Record*, 2019, **19:** 771–778.

[8] XU Y J, ZONG Y, HIPPALGAONKAR K. Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors. *Journal of Physics Communications*, 2020, **4:** 055015.

[9] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey. *ACM Computing Surveys*, 2009, **41(3):** 15.

[10] BEAL M S, HAYDEN B E, GALL T L, *et al*. High throughput methodology for synthesis, screening, and optimization of solid-state lithium ion electrolytes. *ACS Combinatorial Science*, 2011, **13(4):** 375–381.

[11] GHARAGHEIZI F, SATTARI M, ILANI-KASHKOULI P, *et al*. A "non-linear" quantitative structure–property relationship for the prediction of electrical conductivity of ionic liquids. *Chemical Engineering Science*, 2013, **101:** 478–885.

[12] HEMMATI-SARAPARDEH A, TASHAKKORI M, HOSSEINZADEH M, *et al*. On the evaluation of density of ionic liquid binary mixtures: modeling and data assessment. *Journal of Molecular Liquids*, 2016, **222:** 745–751.

[13] HOSSEINZADEH M, HEMMATI-SARAPARDEH A, AMELI F, *et al*. A computational intelligence scheme for estimating electrical conductivity of ternary mixtures containing ionic liquids. *Journal of Molecular Liquids*, 2016, **221:** 624–632.

[14] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习. 硅酸盐学报, 2022, **50(3):** 863–876.

[15] 施思齐, 涂章伟, 邹欣欣, 等. 数据驱动的机器学习在电化学储能材料研究中的应用. 储能科学与技术, 2022, **11(3):** 739–759.

[16] OUYANG R, CURTAROLO S, AHMETCIK E, *et al*. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2018, **2:** 083802.

[17] CHEN C, YE W K, ZUO Y X, *et al*. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 2019, **31:** 3564–3572.

[18] PARK H, JUNG K, NEZAFATI M, *et al*. Sodium ion diffusion in NASICON ($Na_3Zr_2Si_2PO_{12}$) solid electrolytes: effects of excess sodium. *ACS Applied Materials & Interfaces*, 2016, **8(41):** 27814–27824.

[19] LOSILLA E R, ARANDA M A G, BRUQUE S, *et al*. Sodium mobility in the NASICON series $Na_{1+x}Zr_{2-x}In_x(PO_4)_3$. *Chemistry of Materials*, 2000, **12(8):** 2134–2142.

[20] AGGARWAL C C. Outlier Analysis. 2nd Edition. New York: Springer, 2013.

[21] VANDERVIEREN E, HUBERT M. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 2004, **52(12):** 5186–5201.

[22] SEDGWICK P. Pearson's correlation coefficient. *The British Medical Journal*, 2012, **345:** e4483.

[23] ZHOU Y, LI S J. BP neural network modeling with sensitivity analysis on monotonicity-based Spearman coefficient. *Chemometrics and Intelligent Laboratory Systems*, 2020, **200:** 103977.

[24] LI R Z, ZHONG W, ZHU L P. Feature screening *via* distance correlation learning. *Journal of the American Statistical Association*, 2012, **107(499):** 1129–1139.

[25] LIU F T, TING K M, ZHOU Z. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 2012, **6(1):** 1–39.

[26] BREUING M M, KRIEGEL H P, NG R T, *et al*. OPTICS-OF: Identifying density-based local outliers. European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 1999.

[27] HARDIN J, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 2007, **44(4):** 625–638.

[28] HE B, CHI S T, YE A J, *et al*. High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms. *Scientific Data*, 2020, **7:** 151.

[29] TZORTZIS G, LIKAS A. The MinMax k-means clustering algorithm. *Pattern Recognition*, 2014, **47(7):** 2505–2516.

补充材料

融合材料领域知识的数据准确性检测方法

施思齐[1,2,5], 孙拾雨[1], 马舒畅[3], 邹欣欣[3], 钱 权[3,4,5], 刘 悦[3,4,5]

(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; 3. 上海大学 计算机工程与科学学院, 上海 200444; 4. 上海大学 上海市智能计算系统工程技术研究中心, 上海 200444; 5. 之江实验室, 杭州 311100)

S1　描述符取值规则

在 NASICON 型固态电解质材料数据中, 可以根据相关文献获取如表 S1 所示的描述符含义、数据类型、取值范围和数据来源。其中, 描述符 Radius_avg_M 的数据类型是浮点型, 值的范围为(0,+∞)。同时, Radius_avg_M 与元素占据率和半径有关, 且它们之间的相关计算公式如公式(1)所示。

$$\mathrm{Radius\_avg\_M} = (\mathrm{Occu\_M1})\mathrm{Radius\_M1} + (\mathrm{Occu\_M2})\mathrm{Radius\_M2} \tag{S1}$$

其中, Occu_M1, Occu_M2 表示 M1 和 M2 元素占据率, Radius_M1, Radius_M2 表示 M1 和 M2 元素半径。如果 Radius_avg_M 的数据类型不是浮点型或值的范围不在(0,+∞)或 Radius_avg_M 的值与正文公式(3)计算出来的值不一致, 那么该点为潜在的异常数据点。

表 S1 列出了 NASICON 型固态电解质材料中所有描述符的相关信息。

表 S1　NASICON 中所有描述符的含义、数据类型、取值范围和数据来源

Table S1　The meanings, types, ranges, and sources of all descriptors in NASICON

No.	Descriptors	Description	Type	Range	Source
1	Occu_6b	Occupancy of Na in 6b site	float	[0,1]	CIF
2	Occu_18e	Occupancy of Na in 18e site	float	[0,1]	CIF
3	Occu_36f	Occupancy of Na in 36f site	float	[0,1]	CIF
4	*C*_Na	Na^+ concentration	float	(0,+∞)	Formula
5	Occu_M1	Occupancy of element M1	float	[0,1]	CIF
6	Occu_M2	Occupancy of element M2	float	[0,1]	CIF
7	EN_M1	Electronegativity of element M1	float	(0,+∞)	Pauling electronegativity meter
8	EN_M2	Electronegativity of element M2	float	[0,+∞)	Pauling electronegativity meter
9	EN_avg_M	Average effective electronegativity of M site	float	(0,+∞)	Formula
10	Radius_M1	Ionic radius of element M1	float	(0,+∞)	Shannon radius table
11	Radius_M2	Ionic radius of element M2	float	[0,+∞)	Shannon radius table
12	Radius_avg_M	Average effective ionic radius of M site	float	(0,+∞)	Formula
13	Valence_M1	Valence of element M1	int	(0,+∞)	CIF
14	Valence_M2	Valence of element M2	int	[0,+∞)	CIF
15	Valence_avg_M	Average effective ionic valence of M site	float	(0,+∞)	Formula
16	Occu_X1	Occupancy of element X1	float	[0,1]	CIF
17	Occu_X2	Occupancy of element X2	float	[0,1]	CIF
18	EN_X1	Electronegativity of element X1	float	(0,+∞)	CIF
19	EN_X2	Electronegativity of element X2	float	[0,+∞)	CIF
20	EN_avg_X	Average effective electronegativity of X site	float	(0,+∞)	Formula
21	Radius_X1	Ionic radius of element X1	float	(0,+∞)	Shannon radius table

续表

No.	Descriptors	Description	Type	Range	Source
22	Radius_X2	Ionic radius of element X2	float	[0,+∞)	Shannon radius table
23	Radius_avg_X	Average effective ionic radius of X site	float	(0,+∞)	Formula
24	Valence_X1	Valence of element X1	int	(0,+∞)	CIF file
25	Valence_X2	Valence of element X2	int	[0,+∞)	CIF file
26	Valence_avg_X	Average effective ionic valence of X site	float	(0,+∞)	Formula
27	*a*	Lattice parameter	float	(0,+∞)	CIF file
28	*c*	Lattice parameter	float	(0,+∞)	CIF file
29	V_{cell}	Lattice parameter	float	(0,+∞)	Formula
30	V_MO_6	Volume of MO_6 polyhedron	float	(0,+∞)	VESTA file
31	V_XO_4	Volume of XO_4 polyhedron	float	(0,+∞)	VESTA file
32	V_Na_1O_6	Volume of Na_1O_6 polyhedron	float	(0,+∞)	VESTA file
33	V_Na_2O_8	Volume of Na_2O_8 polyhedron	float	(0,+∞)	VESTA file
34	V_Na_3O_5	Volume of Na_3O_5 polyhedron	float	(0,+∞)	VESTA file
35	BT1	Bottleneck	float	(0,+∞)	Formula
36	BT2	Bottleneck	float	(0,+∞)	Formula
37	min_BT	The minimum of BT2 and BT1	float	(0,+∞)	VESTA file
38	RT	Radius of largest sphere probe that can freely pass through the void space packed by framework ions	float	(0,+∞)	Geometry-based Ion-transport Analysis Library CAVD
39	EP_6b	Configurational entropy of Na in 6b site	float	[0,+∞)	Formula
40	EP_18e	Configurational entropy of Na in 18e site	float	[0,+∞)	Formula
41	EP_36f	Configurational entropy of Na in 36f site	float	[0,+∞)	Formula
42	EP_Na	Configurational entropy of Na	float	[0,+∞)	Formula
43	EP_M	Configurational entropy of cationic in M site	float	[0,+∞)	Formula
44	EP_X	Configurational entropy of cationic in X site	float	[0,+∞)	Formula
45	*T*′	Temperature	float	(0,+∞)	Reference

S2 描述符相关性规则

瓶颈是迁移通道中一个非常重要的概念, 描述了离子传输通道中截面积最小的区域。许多情况下瓶颈越大, 迁移离子在离子输运通道中越容易跃迁, 离子在化合物中迁移的激活能越小[1-2]。导通阈值 min_BT 作为离子迁移通道中的最小瓶颈, 也与激活能呈负相关关系, 可以表示为公式(2~4):

$$\mathrm{BT1} \propto -E_a \quad (S2)$$

$$\mathrm{BT2} \propto -E_a \quad (S3)$$

$$\mathrm{min\_BT} \propto -E_a \quad (S4)$$

其中, BT1 和 BT2 分别表示 NASICON 中的两个瓶颈; min_BT 是 BT1 和 BT2 中的最小值, E_a是激活能, 在 NASICON 型固态电解质材料中通常作为材料性能。

在对 NASICON 型固态电解质激活能描述符的选取研究中, 文献[3]报道了迁移离子与骨架氧离子形成的 LiO_6 八面体体积与激活能之间存在很强的负相关关系, 这是因为 LiO_6 八面体体积越大, 离子输运通道越宽, 所以迁移离子从一个位点迁移到另一个位点所需的能量就越少。因此, Na_1O_6 八面体 V_Na_1O_6 的体积也可能与激活能成反比, 表示为公式(5)。

$$V\_\mathrm{Na_1O_6} \propto -E_a \quad (S5)$$

S3 融入描述符知识的数据准确性检测算法

融入描述符知识的数据准确性检测算法具体如算法 S1 所示。其中, F_k用于表示不同的数据驱动的数据正确性检测方法, F_k 的输入为待检测的数据, 输出为潜在的异常数据。

算法 S1　融入描述符知识的数据准确性检测算法

Algorithm S1　Data accuracy detection algorithm incorporating descriptor knowledge

Input	Original dataset D; Anomalous data set $O=\varnothing$; Feature set F; Material property set T; Statistical analysis driven anomalous detection method F_1; Correlation detection method F_2; Multi-dimensional data outlier detection method F_3
Output	Revised dataset $D3$
	# Single-dimensional data correctness detection
1	**For** d_i **in** D:
2	**For** d_i^j **in** d_i:
3	**If** $\mathrm{Ind1}(d_i^j)=1$ **Or** $\mathrm{Ind2}(d_i^j)=1$ **Then**
4	$O=O\bigcup\{d_i^j\}$
5	**If** $O\neq\varnothing$ **Then**
6	According to the material domain knowledge, the anomaly points in O are modified and the modified data set $\mathrm{temp\_}D1$ is obtained
7	$O=\varnothing$
8	**For** d_i **in** $\mathrm{temp\_}D1$:
9	$O=O\bigcup F_1(d_i)$
10	According to the material domain knowledge, the anomaly points in O are modified and the modified data set $D1$ is obtained
	# Multi-dimensional data correlation detection
11	**For** d_i, d_j **in** $D1$:
12	$R(d_i,d_j)=F_2(d_i,d_j)$
13	**If** $\mathrm{Ind3}(d_i,d_j)=1$ **Then**
14	$O=O\bigcup\{d_i\}$
15	According to the material domain knowledge, the anomaly points in O are modified and the modified data set $D2$ is obtained
	# Full-dimensional data reliability detection
16	$\mathrm{Cluster}(F),\mathrm{Cluster}(T)$
17	**For** S_j **in** $D2$:
18	**If** $\mathrm{Ind4}(S_j)=1$ **Then**
19	$O=O\bigcup\{S_j\}$
20	**If** $O\neq\varnothing$ **Then**
21	According to the material domain knowledge, the anomaly points in O are modified and the modified data set $\mathrm{temp\_}D2$ is obtained
22	$O=\varnothing$
23	**For** S_j **in** $\mathrm{temp\_}D2$:
24	$O=O\bigcup F_3(S_j)$
25	According to the material domain knowledge, the anomaly points in O are modified and the modified data set $D3$ is obtained

S4 第一阶段箱线图检测结果

箱线图检测结果如表 S2, 共有 18 个维度被识别出了异常点。

表 S2 异常点检测及修正

Table S2 Anomalous data detection and correction

No.	Descriptor	Description
1	Occu_6b	Occupancy of Na in 6b site
2	Occu_36f	Occupancy of Na in 36f site
3	Occu_M2	Occupancy of element M2
4	Occu_X1	Occupancy of element X1
5	Occu_X2	Occupancy of element X2
6	EN_M1	Electronegativity of element M1
7	EN_avg_M	Average effective electronegativity of M site
8	EN_avg_X	Average effective electronegativity of X site
9	Radius_X1	Ionic radius of element X1
10	Radius_avg_X	Average effective ionic radius of X site
11	Valence_avg_X	Average effective ionic valence of X site
12	V_XO_4	Volume of XO_4 polyhedron
13	V_Na_3O_5	Volume of Na_3O_5 polyhedron
14	min_BT	The minimum of BT2 and BT1
15	EP_36f	Configurational entropy of Na in 36f site
16	EP_X	Configurational entropy of Na in X site
17	T	Temperature
18	E_a	Activation energy

S5 第三阶段实验结果

(1) K-means 聚类结果

图 5(a)中, 第 19、23、56 条样本被识别为异常; 图 5(b)中, 第 69、70、75、79 条样本被识别为异常; 图 5(e)中, 第 4、15、18、72、84、87 条样本被识别为异常; 图 5(f)中, 第 6、7、29、65、71 条样本被识别为异常; 图 5(h)中, 第 68 条样本被识别为异常。将这些异常样本交由专家进一步检验, 材料专家发现有 3 个样本存在输入错误需要进行修正, 在正文中进行了修正, 其余样本均正常。

(2) OCSVM 异常点检测结果

表 S3 基于 OCSVM 的异常样本识别结果

Table S3 Anomalous samples based on OCSVM

No.	ICSD	Formula
6	15546	$Na_{24}Zr_{12}Si_{18}O_{72}$
29	202713	$Na_{14.94}Zr_{10.8}Sc_{1.2}Si_{7.74}P_{10.26}O_{72}$
30	202860	$Na_6Mo_{12}P_{18}O_{72}$
47	260210	$Na_{24}Fe_{12}P_{18}O_{72}$
56	35770	$Na_{4.0002}Co_3Mo_{22.332}O_{72}$
59	421531	$Na_6Ti1_2As_{18}O_{72}$
75	72218	$Na_6Sn_{12}P_{18}O7_2$
84	97956	$Na_6Zr_{12}As_{18}O_{72}$
89	235775	$Na_3MnZr(PO_4)_3$

S6 第三阶段异常点检测方法

表 S4 多维数据异常点检测方法对比

Table S4 Comparison of anomaly detection methods for multi-dimensional data

Method	Description	Advantage	Disadvantage	Application scope
KNN[4]	The nonparametric and distance-based outlier detection method in one-dimensional or multi-dimensional feature space, which depends on the distance measure between data points.	Simple; There is no need to estimate the distribution.	The results are susceptible to the influence of parameters; Not applicable to large data sets.	Small and medium sample data sets.
LOF[5]	An outlier detection method based on "density", which considers the outlier data points different from the surrounding data points in density.	Provide quantitative measurement of outliers.	It is difficult to select parameters; It is not suitable for detecting outliers in the whole region.	Small-scale dataset
IForest[6]	A method considering the division of sample data from each dimension, the earlier the data points are divided into separate areas, the more likely they are outliers.	It has linear time scaling.	Only sensitive to sparse global points, not ideal for dealing with locally sparse points.	Low dimensional data with a small proportion of abnormal data in the total sample size.
OCSVM[7]	A method that the normal samples are divided in the sphere, and the abnormal samples are divided outside the sphere by looking for the hypersphere.	It can be used for high-dimensional data.	The computation cost is higher.	Data distribution has no hypothesis of high dimensional data.
MCD[8]	A method for detecting outliers by location and distribution estimation algorithm.	Simple implementation and robustness.	With the increase of data dimension, the efficiency decreases.	Large-scale high dimensional data.

参考文献:

[1] QUI D T, CAPPONI J J, GONDRAND M, *et al*. Thermal expansion of the framework in NASICION-type structure and its relation to Na^+ mobility. *Solid State Ionics*, 1981, **3(4):** 219–222.

[2] LOSILLA E R, ARANDA M A, BRUQUE S, *et al*. Understanding Na mobility in NASICON materials: a Rietveld, 23Na and 31P MAS NMR, and impedance study. *Chemistry of Materials*, 1998, **10(2):** 665–673.

[3] LANG B, ZIEBARTH B, ELSÄSSER C. Lithium ion conduction in $LiTi_2(PO_4)_3$ and related compounds based on the NASICON structure: a first-principles study. *Chemistry of Materials*, 2015, **27(14):** 5040–5048.

[4] RAMASWAMY S, RASTOGI AND R, SHIM K. Efficient algorithms for mining outliers from large data sets. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000.

[5] BREUING M M, KRIEGEL H P, NG R T, *et al*. OPTICS-OF: identifying density-based local outliers. European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 1999.

[6] LIU F T, TING K M, ZHOU Z. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 2012, **6(1):** 3.

[7] MA J S, PERKINS S. Time-series novelty detection using one-class support vector machines. *Proceedings of the International Joint Conference on Neural Networks*, 2003, **3:** 1741–1745.

[8] HARDIN J, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 2007, **44:** 625–638.