

多保真度数据学习算法的定量噪声评价

刘晓彤^{1,2}, 王滋明², 欧阳嘉华³, 杨 涛^{1,2}

(1. 北京信息科技大学, 北京市材料基因工程高精尖创新中心, 北京 100101;

2. 北京信息科技大学计算机学院, 北京 100101; 3. 暨南大学信息科学技术学院, 广州 511442)

摘 要: 多保真度数据是当前材料领域数据的主要存在形式。在数据生产端, 不同量化方法在材料同种属性的计算上存在较大差距。对于数据消费端的机器学习算法, 研究人员为最大化提取数据中知识设计了各种方法。采用定量噪声添加的方法, 评价不同噪声强度、类型对不同多保真度数据学习方法的影响, 通过迭代降噪验证数据修正方法的适用场景。结果表明: 多保真度数据的利用方式至关重要, 需对各子数据集中数据量及含噪情况进行综合考量。在使用不同噪声类型与强度构造出的多种数据集上, 得益于数据间的协同效应, 逐步删除低保真度数据的“Onion”训练方式明显优于按数据集所含噪声减小方向逐个进行的训练方式。在多保真度数据训练中, 无论何种噪声强度及训练方式, 线性噪声对模型的影响更小。对于采样噪声来说, 在各环节更好地模拟了真实多保真度数据, 建议被后续研究采用。此外, 复杂噪声难以让少量真值数据发挥“纠偏”作用, 更适合进行迭代降噪处理。

关键词: 多保真度; 属性预测; 机器学习; 定量噪声

中图分类号: TP181 文献标志码: A 文章编号: 0454-5648(2023)02-0405-06

网络出版时间: 2023-01-17



A Quantitative Noise Method to Evaluate Machine Learning Algorithm on Multi-Fidelity Data

LIU Xiaotong^{1,2}, WANG Ziming², OUYANG Jiahua³, YANG Tao^{1,2}

(1. Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, China; 2. School of Computer, Beijing Information Science and Technology University, Beijing 100101, China;

3. School of Information Science and Technology, Jinan University, Guangzhou 511442, China)

Abstract: Most data in material science are multi-fidelity data. From the viewpoint of data producer, there is a system error for any quantum method. For machine learning algorithm, as a data consumer, various methods have been designed to maximize the number of knowledges extracted from the multi-fidelity data. In this paper, a quantitative method of noise addition was used to evaluate the influence of different noise types and intensities on some multi-fidelity data learning methods. And the effective scope of the data correction method was verified *via* iterative noise reduction. The results show that the ways to exploit the multi-fidelity data are crucial. It is necessary to consider comprehensively both the size and the noise level of the datasets. On a variety of datasets constructed with different noise types and intensities, the "Onion" training method that gradually deletes lower fidelity data is better than the "one by one" training method in the direction of noise reduction due to the synergistic effect of different multi-fidelity data. No matter what kind of noise intensity and training method, linear noise has less impact on the final performance of model. However, the data with sampled noise added, which the final testing results are similar to the real multi-fidelity data, were recommended to be adopted in a future research. Also, the complex noise in data is difficult to be corrected by a small amount of true data, thus being more suitable for the iterative noise reduction processing.

Keywords: multi-fidelity; property prediction; machine learning; quantitative noise

收稿日期: 2022-09-29。 修订日期: 2022-11-05。

基金项目: 国家自然科学基金项目(22203008, 22272009)。

第一作者: 刘晓彤(1987—), 男, 博士。

通信作者: 杨 涛(1980—), 女, 博士。

Received date: 2022-09-29. Revised date: 2022-11-05.

First author: LIU Xiaotong (1987—), male, Ph.D.

E-mail: liuxiaotong@bistu.edu.cn

Correspondent author: YANG Tao (1980—), female, Ph.D.

E-mail: ytyangmei@bistu.edu.cn

“精度-计算资源权衡”广泛存在于机器学习中数据使用与生产的各个环节,多保真度的概念也由此涌现^[1-2]。在数据使用环节,相同的数据在不同设备(如边缘计算设备,超级计算机等)被不同精度、不同资源消耗的多保真度模型使用^[3-4],根据实际需求更快或更准地产出结果。在数据生产环节,高精度的计算方式往往更耗时,长期数据积累较少;而对理论模型使用了更多近似的快速方法往往在数据积累量上占优,但精度不足。这一现象在材料科学计算领域尤其常见,各种材料数据库中大量不同等级数据,即多保真度数据。与多保真度模型处理相同数据相对应的问题,是如何从多保真度数据中最大化地提取信息。传统机器学习在训练集与测试集之间存在一些基本假设:训练集与测试集中样本彼此相互独立同分布^[5]。在多保真度数据集的利用上,由于不满足这类基础假设,模型的训练面临着诸多陷阱。对于大多场景,材料领域机器学习的目的是预测材料在现实世界中的真实属性值,因此测试集会选用由实验测得的数据结果。而材料数据库(MP)^[6]、开放量子材料数据库(OQMD)^[7]等数据库中存在的多数据,往往是由密度泛函理论(对应一系列泛函,存在不同程度误差^[8])或其他理论方法计算得到,实验测得真值数据相对较少。更有甚者,对于某些材料属性,存在多种实验测量方法,它们彼此结果之间也存在出入^[9],这进一步增加了多保真度数据的出现可能。

研究人员针对不同数据,使用了多种方法处理上述问题,包括对传统模型改进^[10]、信息融合算法^[11]、Bayes 优化^[12]、信息传递神经网络^[13]等。在本团队之前研究中^[14],化学数据的生产端与消费端被区分对待,提出从数据噪声的角度看待这一问题。数据生产者角度认为的系统误差,从数据消费者角度去理解,便是机器学习训练集与真值之间存在的噪声,即数据的保真度。以上观点恰巧与施思齐教授团队最近发表的综述^[15]4.1 节(数据预处理阶段的基础与研究)观点相呼应——“伴随着机器学习在材料领域的广泛应用,决定机器学习模型上限的材料数据质量越来越被重视。材料领域的研究者们分别围绕数据的质和量展开了研究,期望通过提升数据的品质和数量来进一步提升机器学习结果的可信性。”因此,本工作继续尝试从数据角度出发来解决多保真度数据的机器学习问题,不再去设计更复杂的模型,而是去尝试更高效地利用数据。

具体来说,本工作尝试定量地评价多种多保真度数据利用方法,以确定不同方法较优的应用场景。在文献[14]中,所采用的测试数据为 MP 中原生的多保真度数据,预测属性为材料的禁带宽度值。其中,实验测得数据被视为真值数据,使用不同泛函计算得到的数据被视为含有不同程度噪声。该工作除对降噪方法进行综述外,讨论了不同数据的含噪情况及在该特定多保真度数据集条件下,不同数据利用及降噪方法的优劣,建议感兴趣读者优先阅读该文。由于原生多保真度数据中噪声强度、类型及数据集大小均不可控,采用人工引入定量噪声的办法生成模拟多保真度数据集,并以此为基础展开模型训练,评测模型预测性能。相关降噪工作在不同属性、不同数据集及不同机器学习模型上均有复现(如使用 Schnet 模型^[16],在 QM9 数据集^[17-18]上针对形成能的测试)。为延续前文思路,仍选用禁带宽度值对方法进行评估与讨论。未来利用此方法,在为机器学习模型训练准备数据时,可以提前使用相似属性确定较优的多保真度数据配比方案,最大化地节省数据生产所需时间,帮助研究人员将超算机时更合理地分配在不同精度、不同耗时的数据生产算法上。

1 实验

1.1 数据

在之前工作^[10, 14]中,所用多保真度数据集包括了 MP 中使用泛函 Perdew-Burke-Ernzerhof (PBE)^[19], Heyd-Scuseria-Ernzerhof (HSE)^[20-21], Strongly constrained and appropriately normed (SCAN)^[22]及 Gritsenko-Leeuwen-Lenthe-Baerends (GLLB)^[23-24]计算的材料禁带宽度相关数据。各数据集元素种类、目标精度及相互覆盖关系等原始数据分析结果在相关参考文献[14]中已详细给出,此处不再赘述。表 1 为上述各数据集数据量及不同类型材料的分布情况。使用其中数据量最大、种类分布最均匀的 PBE 数据集作为真值数据集,将 52 348^[6]条数据随机划分为 5 个数据集,其中训练集(简称 A、B、C、D 集)分别包含 A: 30 000, B: 10 000, C: 5 000, D: 2 348 条数据,测试集包含 5 000 条数据。有文献报道, DFT 计算所得禁带宽度数据一般较真实值低估 30%~100%^[8],为此,实验设计 3 个噪声等级,将以上各数据集处理成多保真度数据集,具体噪声系数如表 2 所示。

确定噪声等级后,本工作设计了 2 种噪声的添

加方式:

1) 按比例缩小: 指遵循公式 $T_{\text{noise}} = T_{\text{PBE}} * (1 - \theta)$ 对目标值进行处理。此种添加方式在不同数据集中引入了线性噪声。

2) 采样噪声: 指遵循公式 $T_{\text{noise}} = T_{\text{PBE}} * \max(1 - \sigma, 0)$ 对目标值进行处理。其中 σ 为采样自正态分布 $\sigma \sim N(\theta, \theta^2)$ 的随机值。此种添加方式在不同数据集中引入了更复杂的非线性噪声。式中: θ 为各数据集在表 2 中噪声系数; T_{noise} 为添加噪声后的属性值; T_{PBE} 为 PBE 数据集中属性值。

1.2 多保真度数据利用方法及降噪方法

多保真度数据集包含了多个不同保真度及不同大小的子数据集, 如何对其加以利用是一个值得推敲的问题。在前期工作中^[14]中, 本工作列出了以下几种数据集使用方式:

1) 逐个训练("1by1"): 对应于以上数据集中

A→B→C→D 的训练顺序;

2) 洋葱训练("Onion"): 对应于以上数据集中 ABCD→BCD→CD→D 的训练顺序;

3) 集中训练("All together"): 对应于"Onion"训练方式的第一步 ABCD, 即简单混合后进行训练;

4) 单独训练("Only true"): 对应于以上数据集中 D 数据集(根据表 2, 未添加任何噪声)。

所测试的机器学习模型选用 MEGNet^[25](v1.2.3), 超参降噪阈值 $\gamma = 0.35$ 。在迭代降噪过程中, 令训练集数据为 T_{noise} , 前一个模型预测值为 P , 迭代降噪后属性值为 T_{denoised} , 则

$$T_{\text{denoised}} = \begin{cases} P, \text{abs}(P - T_{\text{noise}}) > \gamma \\ T_{\text{noise}}, \text{abs}(P - T_{\text{noise}}) \leq \gamma \end{cases}$$

采用 Early stopping 作为防止过拟合的方法, Patience 选择 10。

表 1 不同禁带宽度(E_g)数据集数据分布情况

Table 1 Distribution of the data as a function of the band gap (E_g)

Functional	$E_g=0$		$E_g<2$		$E_g\geq 2$		Total	
	Data number	Distribution fraction	Data number	Distribution fraction	Data number	Distribution fraction	Data number	Distribution fraction
PBE	19 903	38%	15 218	29%	17 227	33%	52 348	100%
HSE	2 775	46%	1 019	17%	2 236	37%	6 030	100%
SCAN	16	3%	291	63%	165	35%	472	100%
GLLB	0	0%	531	23%	1 759	77%	2 290	100%

表 2 不同噪声等级应用于不同数据集的噪声系数(θ)

Table 2 Noise coefficient (θ) on different datasets with different noise level

Dataset	Noise coefficient $\theta/\%$		
	High noise	Middle noise	Low noise
A	90	60	30
B	60	40	20
C	30	20	10
D	0	0	0

2 结果与讨论

2.1 单保真度数据集大小及噪声的影响

为了探究噪声及数据集大小对模型训练影响, 并为后续实验提供基准, 在引入噪声前后的 A、B、C、D 数据集及全集 ABCD 上分别进行单独训练测试。图 1a 和图 1b 的区别在于数据中所添加的噪声类型, 图 1a 为对数据集添加线性噪声后训练结果; 图 1b 为对数据集添加采样噪声后训练结果。显然, 从这 2 幅热力图中可以清楚看出噪声存在对模型预测性能的影响。当无噪声添加时(对应于图 1a、图 1b 中最下行), 数据量是影响最终结果的唯一因素, 数据越多, 最终训练效果越好。同时, MEGNet 模

型在本数据集上的性能极限也被确定在约 0.4 eV。

添加噪声后, 训练结果显著恶化, 不难理解, 恶化程度与添加噪声量正相关。通过比较不同噪声等级中, 相同噪声添加量, 不同数据量的训练结果, 依然可以得出结论: 相同噪声添加不会影响模型训练性能与数据量的正相关。如高噪声添加的数据集 B 与中噪声添加的数据集 A 噪声系数均为 60%, 最终性能大数据量的 A 依然领先于 B。

在图 1-低噪声”时, 各种多保真度数据集协同训练性能显著优于单独少量真值数据训练结果。这是一个典型的多保真度数据训练特点。低保真度数据并非一无是处, 数据量巨大的低保真度数据即使最简单的混合后训练, 也可以大幅提高模型性能。后续章节将继续讨论的一系列多保真度数据利用算法将更好地提取数据中的信息。

在噪声达到一定程度后, 更大噪声的影响完全覆盖数据量的优势。如在图 1a 与图 1b-高噪声”及-中噪声”添加方案中, -大数据, 大噪声”的数据集 A 单独训练所得结果[图 1a: (1.32 eV, 0.95 eV), 图 1b: (1.04 eV, 0.97 eV)]已明显逊于-小数据, 低噪声”数据集 B[图 1a: (1.01 eV, 0.81 eV), 图 1b: (1.00 eV,

0.92 eV)]和数据集 C 单独训练结果[图 1a: (0.82 eV, 0.83 eV), 图 1b: (0.81 eV, 0.77 eV)]。而 ABCD 全集的训练结果(方法"All together"结果), 随着噪声强度的增加, 也逐渐逊色于数据量最小的 D 数据集训练结果。需说明的是, 图 1a、图 1b 最右一列均为相同数据集无噪声添加的训练结果, 理论上结果应保持一致。在固定随机数种子后, 目前理解实际的细微差异是由 GPU 并行计算过程引入。但这仍不妨碍由此可给出"Only true"方法基线: 0.76~0.78 eV。

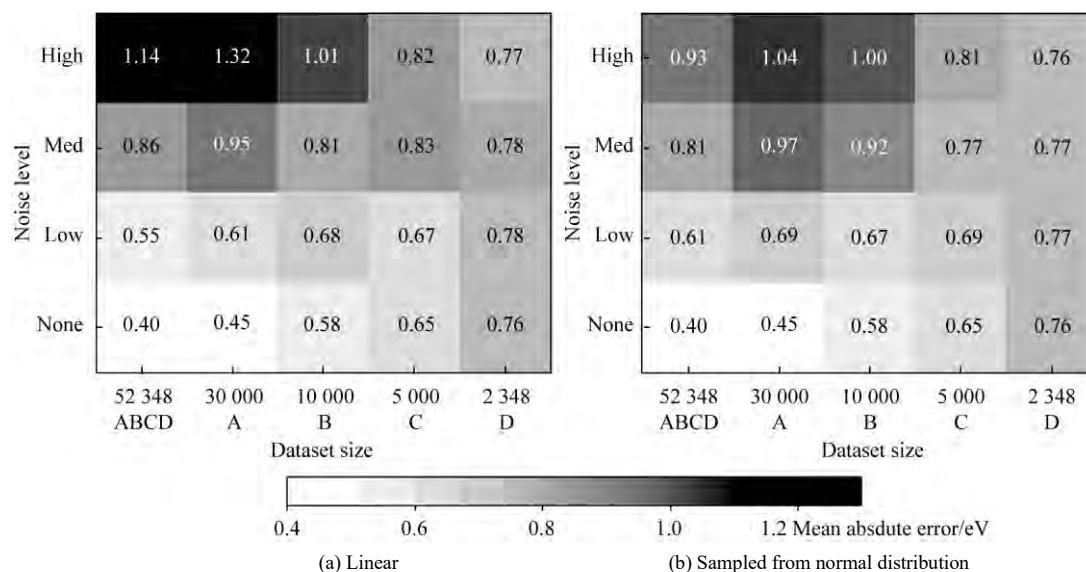


图 1 MEGNet 模型在不同保真度数据集上单独训练的结果和所添加噪声类型
Fig. 1 Training results of MEGNet on different single fidelity datasets and type of noise added

2.2 不同多保真度数据集利用方式的比较

在获得各数据集单独训练模型的结果后, 进一步将实验推进到对多保真度数据的利用上。

图 2 中, 从下至上的 4 条曲线分别代表了不同噪声添加量(无、低、中、高)的实验结果, 为了更清晰比较"lbyl"和"Onion"性能, 对应不同方法采用从两边向中间的方式进行绘图。图 2a 为线性噪声, 图 2b 为采样噪声。图中上 3 条曲线为含噪多保真度数据集训练结果, 2 种比较典型的数据利用方式"Onion"(ABCD→BCD→CD→D)与"lbyl"(A→B→C→D)在不同等级噪声存在下均优于直接混合训练结果(ABCD, 等效于"Onion"中的第 0 步)。不难理解含噪量越大, 最终训练所得模型的平均绝对值误差(MAE)越高。除线性噪声的高噪声情况,"Onion"训练方式在同样多保真度数据集上均优于"lbyl"的训练方式。图 2 中最下方线条组为无噪声添加时各方法训练效果。易看出单一保真度数据集训练第一步结果(与数据集大小强相关)便决定了最终性能表现, 后续步骤对性能提高影响不大。无论何种方法,

比较图 1a 和图 1b 2 种噪声单独训练的热力图易发现, 其中差别最大的是大数据集, 高噪声时的情况。一种合理的解释是线性噪声添加后, 所有数据会往一侧偏移; 采样噪声在均值两侧均会有被采样到的情况, 因此可能出现数据间噪声的“中和”。这与后续两节继续介绍的多保真度数据集综合利用情况略有不同。在"Onion"和"lbyl"方法中, 线性噪声虽存在整体偏移, 但更容易被少量真值数据修正, 取得更好的训练结果。

在任意数据噪声添加方案上, 最终第 3 步的结果均优于图 1 中"Only true"给出的 0.76~0.78 eV 基线, 这再次体现了多保真度数据中存在的信息价值。

2.3 迭代降噪的效果

在之前的工作中, 一个重要发现是多保真度数据集上迭代降噪的优异表现, 即使用前一次得到的模型对数据进行降噪, 之后再重复训练。遗憾当时受限于真实数据集的固定数据分布, 无法对此方法进行更深层次的研究。为此本节特意在不同多保真度数据集上进行评价, 力争更好地给出迭代降噪的效果指引。图 3a、图 3c 对应数据中添加线性噪声的训练结果, 而图 3b、图 3d 所对应的是添加采样噪声的训练结果, 图中虚线为不同方式在无噪声数据集上的训练结果。线性噪声对模型的影响小于采样噪声, 特别是图 3c 中低噪声数据, 采用"Onion"训练方式的训练结果与无噪声数据相差仅约 0.06 eV。相比于 30%、20%、10%的噪声添加比例, 这一结果无疑是十分出色的。目前主要认为这归因于神经网络强大的线性拟合能力, 使得最后一步使用少量

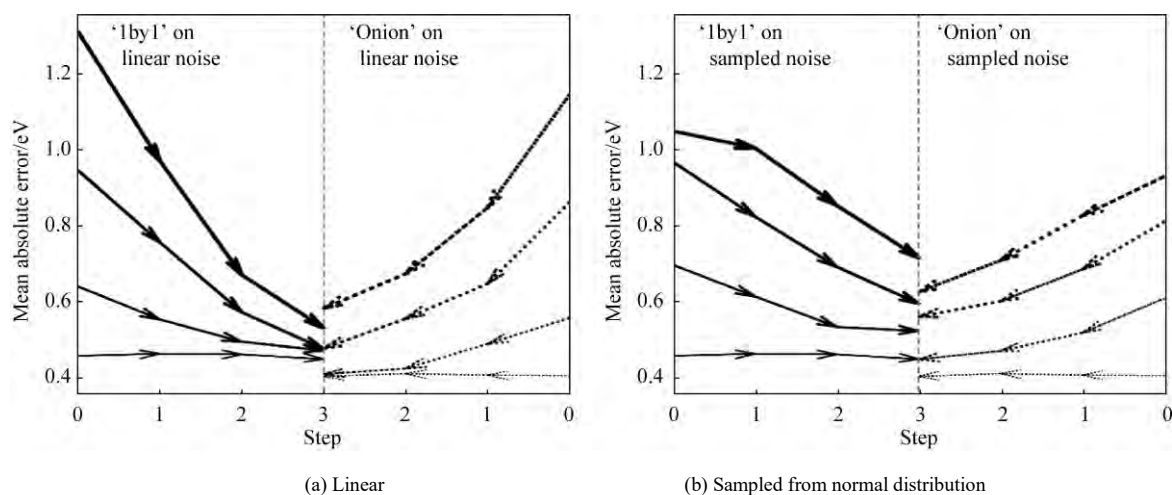


图 2 2 种多保真度数据利用方式在不同噪声等级(上下)、类型(左右)上的比较, 所添加噪声类型

Fig. 2 Comparison between '_lbyl_' and '_Onion_' to exploit the multi-fidelity data on different noise level and types, type of noise added

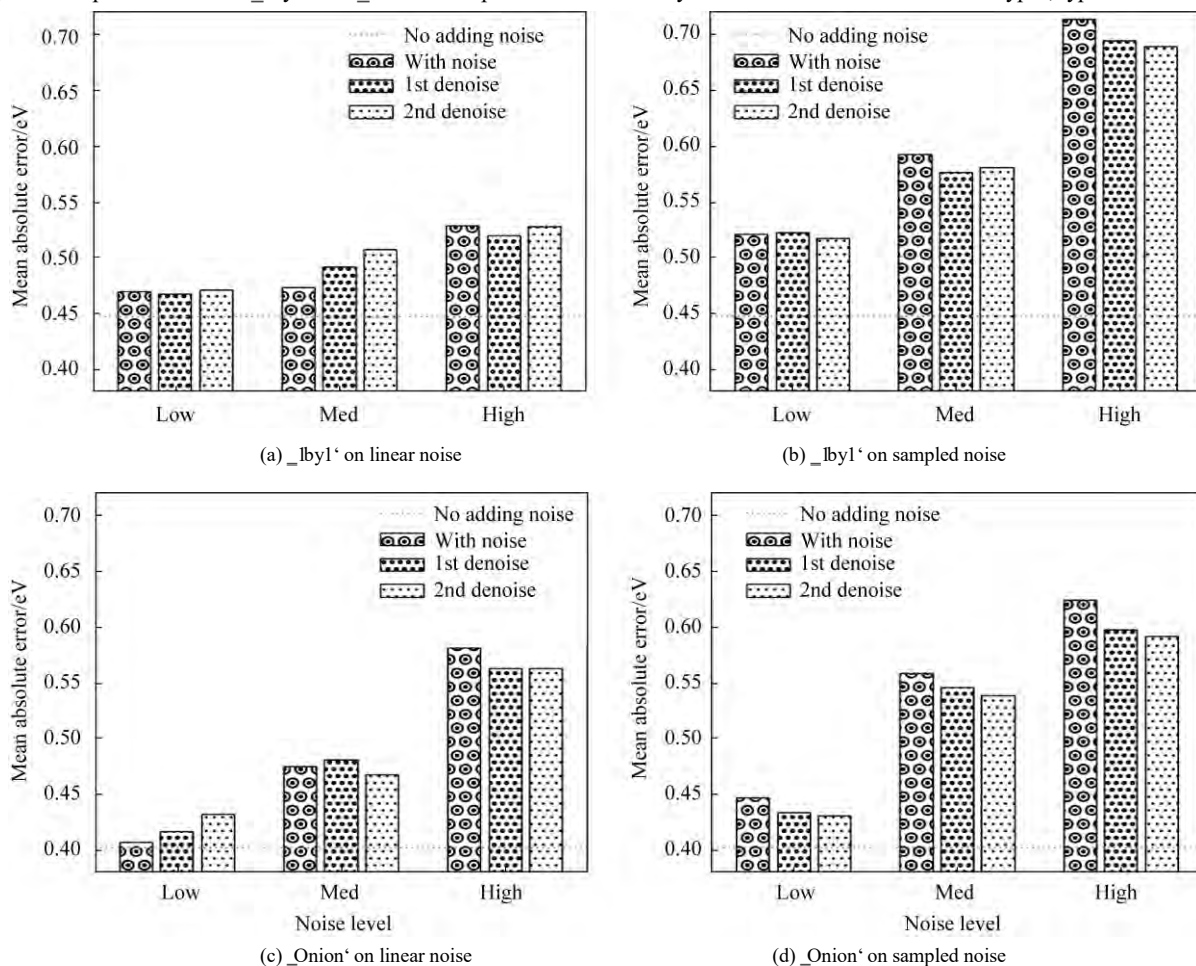


图 3 2 种多保真度数据利用方式在不同噪声等级及强度上的迭代降噪效果

Fig. 3 Denoise effects of '_lbyl_' and '_onion_' on different noise levels and types:

(仅 2 348 条数据)真值 D 数据集可以很容易地将模型调校准确。然而, 在迭代降噪方面, 除图 3c 中 "onion" 方法在高噪声数据上的效果外, 其余均未出现成功的下降, 而这一组所谓的下降, 也是由于初次太差导致(与 2.2 节相一致)。这进一步说明模型第

一轮已经将数据中的信息学习较成功, 线性噪声较易被修正。与之相对应, 对于图 3b、图 3d 中的采样噪声, 仅仅使用 "_lbyl_" 或 "_Onion_" 训练方式已很难应对, 未经迭代降噪的 2 种多保真度利用方法所得结果皆与无噪声数据训练结果有明显差距。这时迭

代降噪的威力便有所显现, 两轮迭代降噪可以分别平均降低"1by1"及"Onion"2种方法的预测误差2.1%和4.1%。

综上所述, 采样噪声被认为是一种与现实噪声更接近的定量噪声。少量的真值数据难以对其训练出的模型进行高效的纠偏, 需使用其他降噪算法进一步对整体数据集进行提升。

3 结论

验证了多保真度数据定量噪声评价方法, 实现了在单一保真度数据集上对模型多保真度性能的评估。在材料禁带宽度数据集上探究了不同噪声分布, 不同训练方式对最终模型预测性能及迭代降噪效果的影响。结果表明, 多保真度数据的利用方式至关重要, 需对各子数据集中数据量及含噪情况进行综合考量。在使用不同噪声类型与强度构造出的多种数据集上, 得益于数据间的协同效应, 逐步删除低保真度数据的-Onion"训练方式明显优于按数据集所含噪声减小方向逐个进行的训练方式。无论对于何种噪声强度及训练方式, 线性噪声对模型的影响更小, 训练所得模型结果与无噪声真值数据训练所得模型结果更为接近。对于采样噪声来说, 在各环节更好地模拟了真实多保真度数据, 建议被后续研究所采用。定量噪声评价策略有望为未来合理规划数据生产提供指导, 更好地解决数据准备环节成本与精度之间的矛盾。

补充信息:

相关模型训练及文中图片绘制脚本已开源至:
https://gitee.com/liuxiaotong15/denoise_jcc

参考文献:

- [1] MUFTI B, CHEN M, PERRON C, et al. A Multi-Fidelity Approximation of the Active Subspace Method for Surrogate Models with High-Dimensional Inputs[C]//AIAA AVIATION 2022 Forum. Chicago: the American Institute of Aeronautics and Astronautics, Inc, 2022: 3488.
- [2] DE HOOP M V, ZHENG YU HUANG D, QIAN E, et al. The cost-accuracy trade-off in operator learning with neural networks [J] *Mathemat: Numerical Anal*, 2022. Doi: 10.48550/arXiv.2203.13181.
- [3] ZIMMER L, LINDAUER M, HUTTER F. Auto-pytorch: Multi-fidelity metalearning for efficient and robust AutoDL[J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, 43(9): 3079–3090.
- [4] ARRIETA A. Multi-Fidelity digital twins: a means for better cyber-physical systems testing?[J]. *Comput Sci: Software Eng*. 2021. Doi: 10.48550/arXiv.2101.05697
- [5] 伊恩·古德费洛. 深度学习[M]. 北京: 人民邮电出版社, 2017: 70.
- [6] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation[J]. *APL Mater*, 2013, 1(1): 011002.
- [7] KIRKLIN S, SAAL J E, MEREDIG B, et al. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies[J]. *Comput Mater Sci*, 2015(1): 15.
- [8] MORALES-GARCÍA Á, VALERO R, ILLAS F. An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations[J]. *J Phys Chem C*, 2017, 121(34): 18862–18866.
- [9] 蒋鸿. 带隙问题: 第一性原理电子能带理论研究现状[J]. *化学进展*, 2012, 24(6): 910–927.
JIANG Hong. *Prog Chem (in Chinese)*, 2012, 24(6): 910–927.
- [10] CHEN C, ZUO Y, YE W, et al. Learning properties of ordered and disordered materials from multi-fidelity data[J]. *Nat Comput Sci*, 2021, 1(1): 46–53.
- [11] BATRA R, PILANIA G, UBERUAGA B P, et al. Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia[J]. *ACS Appl Mater Interf*, 2019, 11(28): 24906–24918.
- [12] TRAN A, TRANCHIDA J, WILDEY T, et al. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys[J]. *J Chem Phys*, 2020, 153(7): 074705.
- [13] GREENMAN K P, GREEN W H, GÓMEZ-BOMBARELLI R. Multi-fidelity prediction of molecular optical peaks with deep learning[J]. *Chem Sci*, 2022, 13(4): 1152–1162.
- [14] LIU X, DE BREUCK P-P, WANG L, et al. A simple denoising approach to exploit multi-fidelity data for machine learning materials properties[J]. *NPJ Comput Mater*, 2022, 8(1): 233.
- [15] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. *硅酸盐学报*, 2022, 50(3): 863–876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. *J Chin Ceram Soc*, 2022, 50(3): 863–876.
- [16] SCHUTT K, KESSEL P, GASTEGGER M, et al. SchNetPack: A deep learning toolbox for atomistic systems[J]. *J Chem Theory Comput*, 2018, 15(1): 448–455.
- [17] RAMAKRISHNAN R, DRAL P O, RUPP M, et al. Quantum chemistry structures and properties of 134 kilo molecules[J]. *Sci Data*, 2014, 1(1): 1–7.
- [18] RUDDIGKEIT L, VAN DEURSEN R, BLUM L C, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17[J]. *J Chem Inform Model*, 2012, 52(11): 2864–2875.
- [19] PERDEW J P, BURKE K, ERNZERHOF M. Generalized gradient approximation made simple[J]. *Phys Rev Lett*, 1996, 77(18): 3865.
- [20] HEYD J, SCUSERIA G E, ERNZERHOF M. Hybrid functionals based on a screened Coulomb potential[J]. *J Chem Phys*, 2003, 118(18): 8207–8215.
- [21] JIE J, WENG M, LI S, et al. A new MaterialGo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps[J]. *Sci Chin Technol Sci*, 2019, 62(8): 1423–1430.
- [22] SUN J, RUZSINSZKY A, PERDEW J P. Strongly constrained and appropriately normed semilocal density functional[J]. *Phys Rev Lett*, 2015, 115(3): 036402.
- [23] GRITSSENKO O, VAN LEEUWEN R, VAN LENTHE E, et al. Self-consistent approximation to the Kohn-Sham exchange potential[J]. *Phys Rev A*, 1995, 51(3): 1944.
- [24] KUISMA M, OJANEN J, ENKOVAARA J, et al. Kohn-Sham potential with discontinuity for band gap materials[J]. *Phys Rev B*, 2010, 82(11): 115106.
- [25] CHEN C, YE W, ZUO Y, et al. Graph networks as a universal machine learning framework for molecules and crystals[J]. *Chem Mater*, 2019, 31(9): 3564–3572.