工程科学学报, 第 45 卷, 第 7 期: 1232−1237, 2023 年 7 月
Chinese Journal of Engineering, Vol. 45, No. 7: 1232−1237, July 2023
https://doi.org/10.13374/j.issn2095-9389.2022.05.03.004; http://cje.ustb.edu.cn

面向材料数据的主动回归学习方法

张　函[1)], 钱　权[1,2)], 武　星[1,2)]✉

1) 上海大学计算机工程与科学学院, 上海 200444　2) 之江实验室, 杭州 311100
✉通信作者, E-mail: xingwu@shu.edu.cn

摘　要　材料的生产环境和测量条件不同, 导致用于机器学习的材料数据的噪声较大. 对材料数据进行标注需要一定的专业知识和专业技能, 因此标注成本也相对较高. 这两方面的因素给机器学习应用于材料领域带来了巨大挑战. 为应对这个挑战, 提出了一个主动回归学习方法, 由离群点检测模块、贪婪采样模块和最小变化采样模块组成. 同其他主动学习方法相比, 该方法整合了离群点检测机制, 选取高质量样本的同时有效地排除了噪声数据的影响, 避免了沉没成本. 在公开数据集和非公开数据集上与最新的主动回归学习方法进行了对比实验, 实验结果表明本文方法在相同的数据量下训练的任务模型性能指标相比于其他模型平均提高 15%, 且只需 30%~40% 的数据量作为训练集就可以达到甚至超过使用全部数据训练任务模型的精度.

关键词　主动学习; 材料; 离群点检测; 回归; 高质量样本
分类号　TG142.71

Active regression learning method for material data

ZHANG Han[1)], *QIAN Quan*[1,2)], *WU Xing*[1,2)]✉

1) School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
2) Zhejiang Laboratory, Hangzhou 311100, China
✉ Corresponding author, E-mail: xingwu@shu.edu.cn

ABSTRACT　To date, artificial intelligence has been successfully applied in various fields of material science, but these applications require a large amount of high-quality data. In practical applications, many unlabeled data points but few labeled data points can be obtained directly. The reason is that data annotations require fine and expensive experiments, and the cost of time and money cannot be ignored. Active learning can select a few high-quality samples from many unlabeled data points for labeling and use as little labeling cost as possible to optimize task model performance. However, active learning methods suitable for material attribute regression are poorly understood, and the general active learning method cannot easily avoid the negative effects of noise data, resulting in decreased costs. Therefore, we propose a new active regression learning method that includes the following features: (1) outlier detection module: using the labeled data prediction from a task model trained to fit and the labeled dataset to train the auxiliary classification model for classifying outliers and then excluding the samples that are most likely to be outliers in the unlabeled dataset; (2) greedy sampling: an iterative method is adopted to select the data farthest from the data in the labeled dataset and the selected data in the geometric space to fully consider sample diversity; and (3) minimum change sampling: selecting the unlabeled data with minimum change before and after the task model, which is trained on the labeled dataset. This part of the data is relatively lacking in the feature space of the labeled dataset. We performed experiments on the concrete slump test dataset and the negative coefficient of thermal expansion dataset and compared our method with the latest active regression learning methods. The results show that other methods do not necessarily improve

收稿日期: 2022−05−03
基金项目: 国家重点研发计划资助项目(2022YFB3707800); 云南省重大科技专项(202102AB080019-3, 202002AB080001-2); 之江实验室科研攻关项目(2021PE0AC02); 上海张江国家自主创新示范区专项发展资金重大项目(ZJ2021-ZD-006)

task model performance after labeling data in each active learning circle on noisy datasets, and the final performance cannot reach the level of the task model trained by all data. Under the same amount of data, the performance index of the task model trained by our method is improved by 15% on average compared with other models. Because of the addition of an outlier detection mechanism, our method can effectively avoid sampling outliers when selecting high-quality samples. The task model trained using only 30%–40% of the data can achieve or even exceed the accuracy of the task model trained by all data.

KEY WORDS　active learning; material; outlier detection; regression; high-quality samples

随着计算机技术的发展，机器学习在工程材料中的应用越来越普遍，预测材料物性参数是一项重要的应用. 一般而言，机器学习算法比较依赖训练数据，训练数据的好坏在很大程度上决定了机器学习模型的性能. 由于测量过程不尽相同，导致工程材料的工艺参数容易产生各式各样的误差，降低了数据的可靠性. 此外，虽然材料的工艺参数获取容易，但是获取实验数据的时间成本和金钱成本较高，容易产生沉没成本. 这些因素给机器学习应用于材料数据带来了一些挑战. 主动学习 (Active learning, AL) 认为数据样本对任务模型并不同等重要，在分类问题中只有少数样本定义了分离面，其他大多数样本都是多余的[1]. 在回归问题中，同样可以认为回归曲线附近的少数样本对回归模型的贡献较大，其他样本的贡献较小. 因此，只要掌握了这些贡献较大的小样本，就能解决整个问题. AL 选择信息最丰富的未标记数据样本，并请外部专家 (Oracle) 对它们进行标注[2]. 通过人机协作的方式，AL 可以实现用尽可能少的有标签样本，训练得到精度尽可能高的任务模型. 因此，将 AL 用于材料的物性参数预测可以降低任务模型对数据的要求，提高任务模型的鲁棒性.

1　研究背景

主动学习是被广泛应用于减少机器学习模型数据需求的方法，主要有三种范式：(1)基于池；(2)基于流；(3)查询合成[3]. 基于池的主动学习的基本问题场景是拥有一个由有标签数据构成的初始训练集和由大量无标签数据构成的数据池，通过迭代抽样的方式从数据池中基于一定的查询策略选择对任务模型性能提升最大的一批数据样本进行标注，然后更新训练集和数据池，直至达到一定的迭代次数或用尽预算. 在实践中很容易获得大量无标注数据，可以根据数据的特性来选择基于池或基于流的主动学习方法[4]. 主动学习的查询策略标准一般有：(1)信息量，选择最具信息量、对任务模型最有好处的无标签样本；(2)多样性，选择与有标签数据有最大差异的无标签样本；(3)代表性，选择无标签数据池中最具有代表性的样本. 近几年主动学习在图像领域研究较多的是图像分类任务. 例如，基于卷积神经网络的主动学习图像分类[5]、将 VAE(Variational autoencoder) 用于主动学习查询[6]、结合半监督学习的主动学习方法[4]、主动学习与自监督学习相结合[7]、考虑不平衡类的主动学习[8] 等. 除此之外，主动学习在图像检索[9]、图像字幕[10]、目标检测[11] 等领域也得到了广泛研究.

主动学习在分类问题中应用广泛，但在回归问题中得到的关注较少，且查询样本的方法也不适用于噪声较大的材料数据. 主动学习在材料科学领域应用较多的是设计具有理想性能的功能材料这一重要前沿. 机器学习相比于传统的试错法效率更高、代价更低，但受限于搜索空间的巨大，机器学习仍无法高效地完成这一任务. 而主动学习则可以起到导航的作用，在巨大的搜索空间中引导机器学习算法朝着性能更好的材料结构方向搜索，从而加速功能材料的预测. 例如，Kusne 等[12] 设计了一个闭环自主的主动学习算法，用于寻找相位之间具有最大光学带隙差的最佳相变材料；类似的还有 Min 和 Cho[13]、Lookman 等[14]、Bassman 等[15] 利用主动学习加速新型材料的设计. 机器学习在材料的物性参数预测方面的应用很广泛，如对热导率[16−17]、拉伸强度[18]、比热容[19−20] 等预测. 但主动学习在这个方面应用较少，没有针对材料数据噪声较多这一特点而设计的主动学习算法. 因此，针对材料数据噪声较多的特点，本文研发了适用于材料数据的、普适性较强的主动学习算法. 本文的贡献在于：

(1) 在主动学习查询过程中结合了样本的多样性和模型信息并排除了离群点的负面影响；

(2) 超参数可以灵活设置，适用于不同噪声程度的数据集；

(3) 提出的主动学习方法与任务模型无关，适用于机器学习的所有回归模型.

2　面向材料数据的主动回归学习

针对材料数据噪声较多的特点，本文提出了

一种新的主动回归学习方法, 有效解决了主动查询噪声数据产生沉没成本的问题, 具体包括三个模块: 离群点检测模块、贪婪采样模块、最小变化采样模块.

2.1 离群点检测模块

由于受外界干扰, 数据集中往往存在一些不符合真实数据分布的离群点. 在特征空间中, 这些离群点距离大部分样本点较远. 为避免离群点的负面影响, 本文提出了离群点检测模块, 旨在利用任务模型与有标签数据的信息, 剔除无标签数据集中最有可能是离群点的数据样本.

假设任务模型已经在有标签数据集上拟合, 对于有标签数据而言, 真值距离任务模型预测值越远就越可能是离群点. 检测离群点的问题是一个二分类问题, 因此需要构造值域在 0~1 之间的、代表是否是离群点的标签, 即:

$$\hat{y}=\begin{cases}1, & \text{if diff} > \text{mid}(\boldsymbol{\text{Diff}})\\ 0, & \text{otherwise}\end{cases},\ \text{diff}=\left|y_{\text{pred}}-y^{*}\right| \tag{1}$$

其中: $\hat{y}$为训练数据的标签, 1 代表是离群点, 0 代表不是离群点; mid(·)是对有标签数据集中所有数据的预测与真值的差值取中值操作, 旨在使训练数据的正负样本平衡, 防止分类器倾向于输出全为同一类, 即$y=1$的样本更接近离群点而不是真实的离群点; **Diff**为全部有标签样本的diff组成的矢量; y_{pred}为任务模型对当前有标签数据的预测值, y^{*}为当前有标签数据的真实标签.

利用有标签数据和公式 (1) 构造的标签训练离群点检测分类模型, 然后对所有无标签样本进行离群点预测, 选择预测概率最小的K个样本. 通过调整K的大小以确定下采样的数量, 等价于调整判断离群点的概率阈值, 可以适应不同噪声程度的数据集.

2.2 贪婪采样模块

GS_x(Greedy sampling on the inputs) 是一种优秀的多样性采样算法[21], 通过迭代的方法从无标签数据集中选择最具有多样性的样本. GS_x 使用最小距离作为标准选择最具多样性的样本, 这种方式在噪声大的数据集易选择到噪声点. 本文将距离标准改进为平均距离, 充分考虑而不过分追求多样性:

不失一般性, 假设已选择样本集合为$\boldsymbol{S}$, 剩余数据池$\boldsymbol{R}$, 计算剩余数据池中样本距所有已选择样本的距离

$$d_{\text{nm}}^{x}=\|x_n-x_m\|,\ n\in\boldsymbol{R};\ m\in\boldsymbol{S} \tag{2}$$

计算剩余数据池中样本距已选择样本的平均距离:

$$d_n^x=\frac{1}{k}\sum_{m=1}^{k}d_{nm}^x \tag{3}$$

选择据已选择样本平均距离最大的样本, 即选择的样本x为

$$x=\arg\max_x d_n^x$$

具体过程见算法 1.

Algorithm 1

Input: labeled data set $\boldsymbol{D}_{\text{l}}$; unlabeled data set $\boldsymbol{D}_{\text{u}}$; number of sampling K

Output: result of sampling $\boldsymbol{D}_K$, $|\boldsymbol{D}_K|=K$

Set $\boldsymbol{D}_K=\varnothing$

For k=1, …, K do

  Calculate d_{nm}^x using (2) where $\boldsymbol{R}=\boldsymbol{D}_{\text{u}}-\boldsymbol{D}_K$ and $\boldsymbol{S}=\boldsymbol{D}_{\text{l}}\cup\boldsymbol{D}_K$

  Calculate d_n^k using (3)

  Choose $x=\text{argmax}_x d_n^x$

  Reset $\boldsymbol{D}_K=\boldsymbol{D}_K\cup\{x\}$

End

2.3 最小变化采样模块

为了进一步提高选择样本的质量, 本文提出了最小变化采样模块, 将模型信息整合到主动学习过程. 在主动学习每次循环过程中, 任务模型训练前后对无标签数据的预测变化幅度并不一样. 变化幅度较小的无标签样本更具有可选择性, 因为经剔除离群点和贪婪采样后, 已选择的数据样本中与有标签数据集中样本有一定的分布差异, 而样本在训练前后变化小证明这部分样本在训练集中是缺乏的, 所以更具有标注的价值. 无标签样本在任务模型训练前后的预测变化Δy的具体计算方法为:

$$\Delta y=|y_{\text{before}}-y_{\text{after}}| \tag{4}$$

其中, y_{before}为任务模型训练前对无标签数据样本的预测; y_{after}为任务模型训练后对无标签数据样本的预测. 选择的样本x为

$$x=\arg\min_x \Delta y \tag{5}$$

2.4 方法总结

本文提出的主动学习方法包括离群点检测、贪婪采样和最小变化采样三个部分. 首先进行离群点检测, 减少后续工作采样到离群点的概率; 然后进行贪婪采样, 尽可能地选择具有多样性的样本; 最后进行最小变化采样, 选择在任务模型训练前后预测变化最小的无标签数据进行标注, 加入训练集进行下一轮循环. 一次主动回归学习循环

的具体步骤如下：

(1) 使用公式 (1) 计算$\hat{y}$，并结合有标签数据训练离群点检测模型；

(2) 利用离群点检测模型预测无标签数据集$\boldsymbol{D}_{\mathrm{u}}$的离群点概率$y_{\mathrm{ODM}}$；

(3) 选择y_{ODM}最小的K个样本放入集合$\boldsymbol{D}_K$；

(4) 使用算法 1 得到$\boldsymbol{D}_K$的子集$\boldsymbol{D}_M$，$|\boldsymbol{D}_M| = M$；

(5) 使用公式 (4) 计算$\boldsymbol{D}_M$中数据的Δy；

(6) 选择Δy最小的B个样本放入集合$\boldsymbol{D}_B$；

(7) 将集合$\boldsymbol{D}_B$中的数据移出$\boldsymbol{D}_{\mathrm{u}}$，标注后放入有标签数据集$\boldsymbol{D}_{\mathrm{l}}$；

(8) 进行下一轮循环，直至满足循环终止条件.

数据流示意图如图 1 所示.

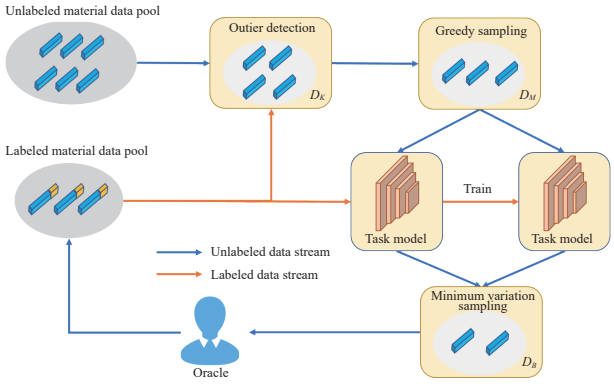

图 1　面向材料数据的主动回归学习数据流示意图

Fig.1　Data flow diagram of active regression learning for material data

3　实验

本小节分别在公开数据集和非公开数据集上将本文的方法与最新的几种主动回归学习方法进行对比，最后进行了消融实验来进一步地验证本文提出的主动回归学习方法的合理性.

3.1　数据集和实验设置

3.1.1　数据集

(1) 混凝土坍落度测试数据集 (Concrete slump test)：坍落度是检测混凝土和易性、黏聚性、保水性能的一种重要指标，常用于检测铁路建设中混凝土的性能. 混凝土坍落度实验数据集包含 103 条数据样本，每条数据样本有 7 个特征，预测标签数为 3 个，分别为坍落度 (Slump, cm)、流量 (Flow, cm)、28 d 后抗压强度 (28-day compressive strength, MPa).

(2) 负热膨胀材料数据集 (Negative thermal expansion material)：负热膨胀为调控膨胀系数，解决现代科学技术中的一些难题提供了新的机遇[22]. 然而人工合成的负热膨胀材料的热膨胀系数难以调控，需要进行大量的实验. 为了降低成本，我们进行实验并且搜集了一些相关数据，希望利用机器学习对材料的负热膨胀系数进行预测. 数据集共包含 132 条数据样本，其中 9 条数据样本是我们进行精确实验得到的，其余 123 条数据样本是在各类文献中搜集得到的；每条数据样本有 18 个特征，预测标签数为 3 个，分别为出现负膨胀现象的起始温度 (T_1, ℃)、终止温度 (T_2, ℃)、坐标图上 T_1 与 T_2 连线的斜率 (α).

3.1.2　实验设置

支持向量回归 (Support vector regression, SVR) 是支持向量机 (Support vector machine, SVM) 在回归领域的应用. SVR 和 SVM 基于结构风险最小化原则，以提高学习机的泛化能力并降低经验风险和置信区间[23]，在小型数据集上表现良好. 因此，在实验中使用 SVR 作为任务模型，SVM 作为离群点检测辅助模型.

由于 2.3 节所述的最小变化采样需要模型训练前后的预测结果，SVR 需要先进行训练才可以进行预测，因此在使用本文的方法前需要先进行一次采样. 为了使采样方法更多样，首次采样考虑样本的代表性，即对无标签数据集进行聚类，聚类中心个数为B，选择距离每个聚类中心最近的样本进行标注，并加入有标签数据集.

具体地，根据实验的数据集大小，设置$K = 40$、$M = 20$、$B = 10$，主动学习循环次数固定为 4 次. 负热膨胀材料数据集中 9 条数据样本是我们精确实验得到的，且其中 3 条数据样本实验较为理想，因此取这 3 条数据样本作为测试集，其余 6 条数据样本作为初始训练集. 混凝土坍落度实验数据集上进行的试验也进行同样的变量设置和测试集划分. 需要注意的是，第一次聚类采样不计入主动学习循环次数.

3.2　对比方法和评价指标

3.2.1　对比方法

(1) 随机采样 (Random sampling, RS)：随机从无标签数据集中选取样本进行标注，是一种最简单的采样方法，也是主动学习中常用的基准 (baseline) 方法.

(2) 改进的贪婪采样 (Improved greedy sampling, iGS)[21]：结合 GS_x 和 GS_y (Greedy sampling on the output)，充分考虑了输入空间和输出空间样本的多样性，实现了主动学习算法的有效性和鲁棒性.

(3) 基于池的序列主动回归学习结合 iGS (Representativeness and diversity with iGS, RD−iGS)：Wu[24] 将所有数据以有标签数据集中样本个数加 1 为类

簇个数进行聚类, 在不包含有标签数据的类簇中结合其他选择方法选择无标签样本进行标注, 直至达到标记样本的最大数量. 本文将其与 iGS 进行结合, 同时考虑了代表性和多样性, 保证了有效性.

3.2.2　评价指标

决定系数 (R^2) 是最广泛使用的“拟合优度”的指标[25]. 因此本文的实验采取决定系数 (R^2) 作为评价指标, 计算公式为

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^* - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i^* - \bar{y}_i)^2}$$

其中, y_i^*为样本的真实标签; $\hat{y}_i$为回归模型的预测值; $\bar{y}_i$为样本真实标签的平均值. R^2 的取值范围是负无穷到 1, 越接近 1, 模型的预测越准确.

3.3　对比实验

在两个数据集上的实验结果如图 2、图 3 所示. 图像的 x 轴是训练集数据个数, y 轴为三个回归参数的决定系数平均值$\bar{R}^2$.

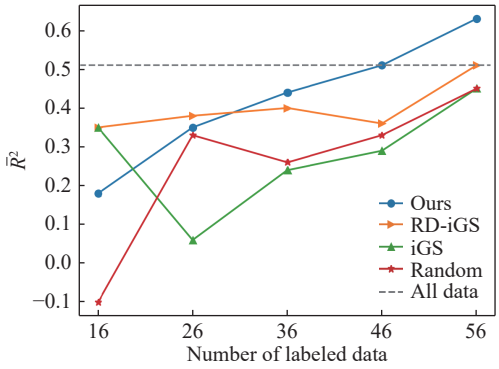

图 2　四种算法在混凝土坍落度测试数据集上的表现

Fig.2　Performance of four algorithms on the concrete slump test dataset

由图 2、图 3 可知, 在这两个材料数据集上, iGS 和 RD−iGS 每轮主动学习循环后模型性能提升不明显, 有时模型性能甚至会有些下降. 原因在于数据集含有大量噪声, iGS 和 RD−iGS 都考虑了样本的多样性, 所以容易选择离群点进行标注.

在训练数据较少时, 噪声数据占比较大, 影响了模型拟合真实数据分布, 因此图 2 和图 3 中的性能指标在训练前期波动较大. RD−iGS 同 iGS 相比增加了聚类的操作, 在不含有标签样本的类簇中运行 iGS 算法, 因此先考虑代表性再考虑多样性, 减少了标注离群点的概率, 前几轮循环性能较好. 但在图 3 中 RD−iGS 在最后两轮循环性能低于 iGS, 是因为负热膨胀材料数据集中的数据大多来自文献, 相对于混凝土测试数据集噪声较大; 并且 RD−iGS 由于先考虑代表性的原因, 选择的数据与有标签

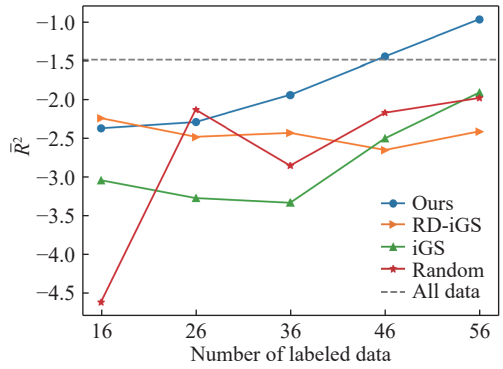

图 3　四种算法在负热膨胀材料数据集上的表现

Fig.3　Performance of four algorithms on the negative thermal expansion material dataset

数据差异更大, 在后几轮循环中选择样本的质量并没有显著提升; 而 iGS 在无标签数据池规模变小后, 离群点占比也变小, 更易选择高质量的样本, 所以导致 iGS 在图 3 中、后期表现优于 RD−iGS.

为增加可比性, 以相同的初始条件运行五次随机采样算法, 最终将结果取平均. 尽管如此, 随机采样还是有很大的波动, 并且噪声越大波动也就越大. 本文的方法加入了离群点检测机制, 在这两个噪声较大的数据集中表现都很优秀. 尽管前几轮循环性能不如其他算法, 但没有出现波动, 随着标注样本的增加, 任务模型的性能也在不停地上升. 在训练集有 46 个样本时, 模型已经达到使用全部数据训练的模型性能, 并且性能还可以进一步的提升, 这是其他算法所达不到的.

3.4　消融实验

我们分别去除离群点检测模块、贪婪采样模块和最小变化采样模块, 并在混凝土坍落度测试数据集和负热膨胀材料数据集上执行 4 轮主动学习循环. 表 1 展示了实验结果, $\bar{R}^2$表示 4 轮主动学习循环后任务模型预测的平均决定系数.

离群点检测模块的功能是剔除无标签数据池中的离群点, 在这两个数据集中去除离群点检测模块后性能指标都有最明显的下降, 表明在噪声比较大的材料数据中离群点检测模块能起到较大的作用. 同时, 在混凝土坍落度测试数据集去除贪婪采样模块后性能指标下降的幅度比去除最小变化采样模块后性能指标下降的幅度大, 而在负膨胀材料数据集上却相反, 由于以多样性为标准的采样方法容易选择到噪声数据, 这也从侧面表明了负膨胀材料数据集的噪声相对更大一些. 从整体来看, 缺少这三个模块中的任意一个都会导致性能指标降低, 说明这三个模块对整体采样都有不可或缺的贡献.

表 1　在混凝土坍落度测试数据集和负膨胀材料数据集上的消融实验结果

Table 1　Ablation experimental results on the concrete slump test dataset and the negative thermal expansion material dataset

Method	$\bar{R}^2$ for concrete slump test dataset	$\bar{R}^2$ for negative coefficient of thermal expansion dataset
Without outlier detection module	0.39	−2
Without greedy sampling module	0.43	−1.42
Without minimum change sampling module	0.46	−1.83
Complete method	0.52	−0.97

4　总结

(1) 针对材料科学领域数据噪声较多的特点，本文提出了一个包含离群点检测模块的主动学习框架.

(2) 本文的主动学习框架与任务模型无关，适用于机器学习的所有回归模型，并且根据数据集噪声大小和预算可以适量调整超参数K、M、B的大小，具有较强的普适性.

(3) 在公开数据集和非公开数据集上经实验验证，本文的方法只需不到一半的数据量就可以达到全体数据训练模型的精度，极大降低了成本.

参 考 文 献

[1] Cao X Y, Yao J, Xu Z B, et al. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Trans Geosci Remote Sens*, 2020, 58(7): 4604

[2] Tsymbalov E, Panov M, Shapeev A. Dropout-based active learning for regression // *International Conference on Analysis of Images, Social Networks and Texts*. Moscow, 2018: 247

[3] Kumar P, Gupta A. Active learning query strategies for classification, regression, and clustering: A survey. *J Comput Sci Technol*, 2020, 35(4): 913

[4] Guo J N, Shi H C, Kang Y Y, et al. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 2896

[5] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach // *International Conference on Learning Representations*. Vancouver, 2018: 38

[6] Sinha S, Ebrahimi S, Darrell T. Variational adversarial active learning // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, 2019: 5971

[7] Bengar J Z, van de Weijer J, Twardowski B, et al. Reducing label effort: Self-supervised meets active learning // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 1631

[8] Choi J, Yi K M, Kim J, et al. Vab-al: Incorporating class imbalance and difficulty with variational bayes for active learning // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, 2021: 6745

[9] Barz B, Käding C, Denzler J. Information-theoretic active learning for content-based image retrieval // *German Conference on Pattern Recognition*. Stuttgart, 2018: 650

[10] Deng Y, Chen K W, Shen Y L, et al. Adversarial active learning for sequences labeling and generation // *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 4012

[11] Bengar J Z, Gonzalez-Garcia A, Villalonga G, et al. Temporal coherence for active learning in videos // *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. Seoul, 2019: 914

[12] Kusne A G, Yu H S, Wu C M, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat Commun*, 2020, 11: 5966

[13] Min K, Cho E. Accelerated discovery of novel inorganic materials with desired properties using active learning. *J Phys Chem C*, 2020, 124(27): 14759

[14] Lookman T, Balachandran P V, Xue D Z, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput Mater*, 2019, 5: 21

[15] Bassman L, Rajak P, Kalia R K, et al. Active learning for accelerated design of layered materials. *Npj Comput Mater*, 2018, 4: 74

[16] Juneja R, Yumnam G, Satsangi S, et al. Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity. *Chem Mater*, 2019, 31(14): 5145

[17] Loftis C, Yuan K P, Zhao Y, et al. Lattice thermal conductivity prediction using symbolic regression and machine learning. *J Phys Chem A*, 2021, 125(1): 435

[18] Le T T. Prediction of tensile strength of polymer carbon nanotube composites using practical machine learning method. *J Compos Mater*, 2020, 55(6): 787

[19] Çolak A B, Yıldız O, Bayrak M, et al. Experimental study for predicting the specific heat of water based Cu–Al_2O_3 hybrid nanofluid using artificial neural network and proposing new correlation. *Int J Energy Res*, 2020, 44(9): 7198

[20] Assad M E H, Mahariq I, Ghandour R, et al. Utilization of machine learning methods in modeling specific heat capacity of nanofluids. *Comput Mater Continua*, 2022(1): 361

[21] Wu D R, Lin C T, Huang J. Active learning for regression using greedy sampling. *Inf Sci*, 2019, 474: 90

[22] Liang E J, Sun Q, Yuan H L, et al. Negative thermal expansion: Mechanisms and materials. *Front Phys*, 2021, 16(5): 53302

[23] Mehr A D, Nourani V, Khosrowshahi V K, et al. A hybrid support vector regression–firefly model for monthly rainfall forecasting. *Int J Environ Sci Technol*, 2019, 16(1): 335

[24] Wu D R. Pool-based sequential active learning for regression. *IEEE Trans Neural Netw Learn Syst*, 2019, 30(5): 1348

[25] Onyutha C. A hydrological model skill score and revised R-squared. *Hydrol Res*, 2022, 53(1): 51