DOI: 10.19650/j.cnki.cjsi.J2312180

# 基于自适应密度聚类的多准则主动学习方法*

贺忠海[1,2],朱温涵[1],陈旭旺[1],张晓芳[3]
(1.东北大学秦皇岛分校控制工程学院 秦皇岛 066004; 2.河北省微纳传感重点实验室 秦皇岛 066004;
3.北京理工大学光电学院 北京 100081)

**摘 要**:主动学习能够以更少的标注成本训练出更好的机器学习模型。现有的 RD 算法与 QBC 算法的结合有效地解决了只考虑单一标准的问题。然而,RD 所基于的 K-means 聚类会将离群点也包括在内进而造成模型性能降低,而 QBC 则需要维护于多个模型而间接返回样本的信息性. 针对上述问题,本文提出了一种基于自适应密度聚类的高斯过程回归(ADC-GPR)算法,通过先聚类后直接利用不确定性进而高效选择样本。该算法中的 ADC 聚类不仅对离群点鲁棒,还能根据数据集分布特性自适应聚类,并为后续的 AL 提供了代表性样本点和其对应的簇,该方法在无监督选择时保证了代表性和多样性,在有监督选择时考虑了信息性、代表性和多样性。实验结果表明,在相同的抽样次数下将 ADC-GPR 算法与 RS、KS 以及 RD-GPR 算法相比,其平均性能分别提升了 37.3%、8%和 2.8%,ADC-GPR 算法的选择效率更高。

**关键词**:主动学习;自适应密度聚类;高斯过程回归;离群点鲁棒;多标准融合
**中图分类号**:TH741 **文献标识码**:A **国家标准学科分类代码**:460.40

# A multi-criteria active learning method based on adaptive density clustering

He Zhonghai[1,2], Zhu Wenhan [1], Chen Xuwang[1], Zhang Xiaofang[3]
(*1. School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China*;
*2. Hebei Key Laboratory of Micro-Nano Sensing, Qinhuangdao 066004, China*; *3. School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China*)

**Abstract**: Active learning proves instrumental in training superior machine learning models while minimizing labeling costs. The combination of RD and QBC algorithms effectively addresses issues associated with considering only a single criterion. However, the K-means clustering upon which RD is based may include outliers, leading to a decrease in model performance, and QBC requires maintaining multiple models and indirectly provides sample information. To address these issues, we propose an adaptive density clustering-based Gaussian process regression (ADC-GPR) algorithm, which efficiently selects samples by first clustering and then utilizing uncertainty directly. The ADC clustering in this algorithm is not only robust against outliers but also adapts to the distribution characteristics of the dataset, providing representative sample points and their corresponding clusters for subsequent AL. This method ensures both representativeness and diversity in unsupervised selection and considers informativeness, representativeness, and diversity in supervised selection. The experimental results demonstrate that compared to the RS, KS, and RD-GPR algorithms, the ADC-GPR algorithm exhibits an average performance improvement of 37.3%, 8%, and 2.8% respectively, with the same number of sampling iterations. Furthermore, the ADC-GPR algorithm demonstrates higher selection efficiency.

**Keywords**: active learning; adaptive density clustering; Gaussian process regression; outlier robustness; multi-criteria fusion

收稿日期:2023-11-22 Received Date: 2023-11-22
*基金项目:河北省自然科学基金(F2020501040)项目资助

## 0 引 言

近红外光谱是一种非破坏性的物质成分分析方法,在实际应用中,光谱数据相对容易测量,而与其相关的理化值等标记通常却难以获取[1-2]。为有效节约资源,以更少的成本得到更精确的模型,主动学习(active learning,AL)近年来被应用到该领域。基于池的主动学习[3]是一种有效的机器学习方法,其核心思想是从未标记的样本池中选择出不确定性最大的样本进行标记,并将这些样本用来建立模型,从而在保持较低标记成本的同时,获得更高的模型性能。

AL 通常分为两步[4]:1)无监督地选出少量样本并标记,建立一个回归模型。2)有监督地选择一些对模型最有帮助的未标记样本并标记,然后将新标记的样本添加到训练集并用于更新模型,此过程迭代直到标记样本达到选定样本数为止[5]。

AL 的无监督选择,是要求算法在没有任何标签信息的情况下最佳地选择要标注的训练样本并建立初始的回归模型,这影响到主动学习第二步的有监督选择阶段。在早些年的相关研究中,通常是简单地通过随机抽样(random sample,RS)的方式标注训练样本,而将更多的精力专注到迭代过程,这使得 AL 效率低下。为了在初始阶段更有效地选择未标注样本,肯纳德-斯通(Kennard-Stone,KS)算法[6]能够基于样本在特征空间中的几何特征来选择一个新的样本,使它的位置远离之前已选择过的样本,然而却只考虑了样本在输入空间的均匀分布(多样性),没有考虑在样本点周围有相似或接近的样本数量(代表性)[7]。为同时考虑代表性和多样性,Wu 等[8]提出了一种基于样本代表性和多样性的算法(representativeness and diversity,RD),利用 K-means 聚类将数据点分为 K 个簇,选择并标注每个簇的质心,然而该聚类方法试图将包括离群点在内的所有点都进行聚类,对于不同大小和密度的簇,这会使得部分聚类中心偏离真实的簇中心[9]。而基于密度聚类的算法(density-based spatial clustering of applications with noise,DBSCAN)[10]能够捕获任意形状的高密度簇,由于离群点所在区域密度无法满足邻域最小数目的要求,会被算法单独识别出来,故解决了 K-means 不能识别和排除离群点的问题,但聚类结果过于依赖邻域半径和邻域最小数目两个参数,取值不当会导致聚类效果变差甚至不正确,传统的手动调参过于依赖经验[11]。为此,根据数据集本身性质自动确定聚类参数是可行的方法,提出的自适应密度聚类(adaptive density cluster,ADC)能够根据数据分布自动聚类进而发现高密度数据集合,每个集合内都有代表性的样本点,而不同的集合代表不同的簇,将代表性的样本点和其对应的簇用于整个 AL 阶段,使得每一次选择都具有代表性,在无监督阶段对这些代表点使用 KS 算法以确保选择具有代表性和多样性。

AL 的有监督选择,是在建立好初始回归模型的基础上,选择那些不确定度(信息性)最大的样本并标注,将新标记的样本添加到训练集用于更新模型[12]。为了选择不确定度最大的样本,Cai 等[13]提出了适用于回归的预模型变化最大化(expected model change maximization,EMCM)算法,旨在选择导致当前模型变化最大的样本,RayChaudhuri 等[14]提出了委员会查询(expected model change maximization,QBC)算法,将未标记样本的信息量计算为各委员会成员之间的分歧,然而这种两方法都是间接计算得到的信息性。为了解决这一问题,可以直接利用高斯过程回归(gaussian process regression,GPR)[15]提供相关样本不确定度的信息,并且由于 GPR 能够有效利用先验信息,所以其在小样本的数据集上有更好的表现。为了能同时考虑到样本的信息性和其空间几何特征(代表性和多样性),将 GPR 直接返回的置信度(信息性)与上文中 ADC 聚类得到代表性样本点和其对应的簇结合起来使得每一次有监督选择能够选择不同的簇,这不仅排除了离群点的影响,还综合了代表性、多样性和信息性。

本文提出的 ADC-GPR 算法首先在主动学习开始前通过 ADC 聚类自适应找到高密度聚类,并将其中有代表性的点和其对应的簇用于整个 AL 阶段以确保选择过程具有代表性。在无监督选择时,使用 KS 算法以保证代表性的和多样性;在有监督选择时,使用 GPR 直接返回预测点的不确定度,并判断不确定度最大的样本所属簇是否已经挑选过其他同簇样本,若是,则顺延考虑下一个不确定最大的样本(这样就防止了样本点在一个空间内过于集中进而体现了选择点的多样性),否则将该样本标注并加入训练集,若簇的所有编号都已选择过一遍,则在下次迭代时这些簇能够被重新考虑,迭代到标注样本达到规定数量为止。

本文的主要贡献有:

1)在主动学习开始前,基于数据集自身空间特性自适应完成了密度聚类,该聚类不仅对离群点鲁棒,还为 AL 过程提供了代表性的点和其对应的簇的编号,这使得整个 AL 学习过程都能够确保有代表性。

2)在初始的无监督选择中,对代表点使用 KS 算法选择标注,在没有标记的情况下充分考虑了样本的空间特性,使得选择的样本既有代表性又有多样性。

3)在后续的有监督选择中,使用 GPR 直接返回的不确定度衡量代表点的信息量,并利用先前聚类提供的簇信息使得每一次选择的代表点都处于不同的簇中,这样就综合了信息性、代表性和多样性。

# 1　自适应密度聚类方法

## 1.1　基本密度聚类方法

DBSCAN 是基于数据点的密度来确定簇的数量和形状,能够识别高密度区域的数据点,并将数据点划分为核心点、边界点和离群点。由于核心点的定义与代表性的定义相似,将核心点认为是具有代表性的点:对于任一样本 $x_n$,如果其邻域半径 *Eps* 内至少包含邻域最小数目 *MinPts* 个样本,则将 $x_n$ 记作核心点。密度可达的代表点和边界点将划分进同一个聚类,离群点则排除在外,DBSCAN 的基本思想如图 1 所示。

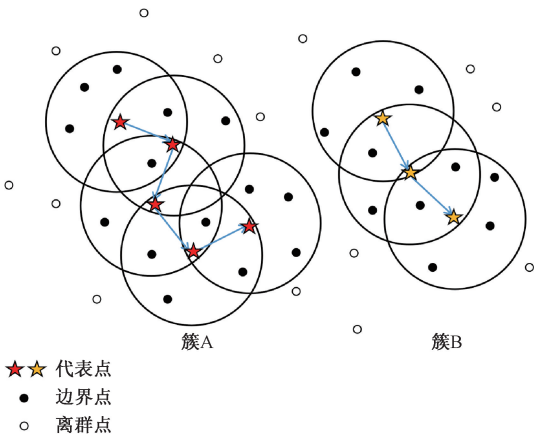



图 1　DBSCAN 聚类(MinPts=5)

Fig. 1　DBSCAN clustering (MinPts=5)

图 1 中,左侧五角星是属于簇 A 的代表点,右侧五角星是属于簇 B 的代表点,同一个簇内的代表点都有相同且唯一的簇编号,不同簇内的代表点编号则不同。通过一次 DBSCAN 聚类就能够得到代表点和对应的簇,在主动学习中将只从这些代表点中进行样本选择以提高效率。

DBSCAN 聚类的两个参数 *Eps* 和 *MinPts* 通常都是手动调整的,这需要主观的判断和丰富的经验。较小的 *Eps* 值会将更多的点划分为噪声或边界点,而较大的 *Eps* 值会将不同的聚类合并在一起使得核心点更具有代表性而聚类缺少了多样性;较小的 *MinPts* 值会形成更多的细小聚类,这使得核心点缺乏代表性而丰富了聚类的多样性,而较大的 *MinPts* 值可能会将较小的聚类合并成一个大聚类。

显然,代表性和多样性会有一定的冲突,为了能够实现自动聚类并平衡好代表性和多样性,本文提出的 ADC 算法利用了数据集自身的空间分布特性生成密度阈值列表,并将这些参数列表依次进行聚类分析并通过寻找最大轮廓系数以确保尽可能多地检索不同类型的代表性数据。

## 1.2　通过密度阈值列表得到数据集聚类基本参数

首先,要明确参数列表的取值范围和参数间距以确定合适的密度阈值参数列表,本文利用数据集自身的空间分布特性,计算每个数据点与其第 *K* 个最近邻数据点之间的 *K*-最近邻距离,并对所有数据点的 *K*-最近邻距离求平均值,得到数据集的 *K*-平均最近邻距离列表即为 *Eps* 参数列表,并根据给定的 *Eps* 参数列表求出每个 *Eps* 参数对应的 *Eps* 邻域对象数量,并计算所有对象的 *Eps* 邻域对象数量的数学期望值作为 *MinPts* 参数列表,具体步骤如下:

1)计算数据集 D 的距离矩阵 $\boldsymbol{D}_{n\times n}$,其中为数据集 D 中第 $i$ 个对象到第 $j$ 个对象的距离,即:

$$\boldsymbol{D}_{n\times n}=\{Dist(i,j)\mid 1\leqslant i,j\leqslant N\} \tag{1}$$

对 $\boldsymbol{D}_{n\times n}$ 的每一行进行升序排序得到 $\boldsymbol{D}_k$,则第 1 列的所有元素表示对象到自身的距离均为 0,第 *K* 列的元素构成所有数据点的 K-最近邻距离向量 $\boldsymbol{D}_k$;对其中的元素求平均,即可得到向量 $\boldsymbol{D}_k$ 的 *K* - 平均最近距离 $\overline{\boldsymbol{D}_k}$,并将其作为候选参数 $\boldsymbol{D}_{Eps}$,表示为:

$$\boldsymbol{D}_{Eps}=\{\overline{\boldsymbol{D}_k}\mid 1\leqslant K\leqslant N\} \tag{2}$$

2)对于给定的 $\boldsymbol{D}_{Eps}$ 参数列表 ,依次求出每个 *Eps* 邻域内的对象数量 $P_i$,并计算所有数据的 *Eps* 邻域样本数量的数学期望值,即:

$$MinPts=\frac{1}{N}\sum_{i=1}^{N}P_i \tag{3}$$

其中,*N* 为样本个数,得到与 $\boldsymbol{D}_{Eps}$ 对应的 *MinPts* 参数列表:

$$\boldsymbol{D}_{MinPts}=\{MinPts\mid 1\leqslant K\leqslant N\} \tag{4}$$

3)根据密度阈值列表 $\boldsymbol{D}_{Eps}$ 和 $\boldsymbol{D}_{MinPts}$,将 *N* 个参数对输入到 DBSCAN 进行聚类分析。为了从这 *N* 个参数对中自适应地寻找最优聚类,引入轮廓系数这一聚类评价指标。

## 1.3　结合轮廓系数得到最优聚类结果

在特征空间中,样本的代表性与密度有关,样本点周围相近样本数量越多,凝聚度越大,则代表性越好;样本的多样性与样本的差异性有关,簇的类别越多,分离度越大,则多样性越好。

轮廓系数[16]综合了样本与其所属簇内的相似度以及样本与最近的其他簇间的不相似度,用以评估聚类结果的紧密度和分离度。具体来说,如果一个数据点与自己所属的簇内的其他数据点的距离很小,但是与其他簇中的数据点的距离很大,就表示这个数据点所在的簇内紧密度高,簇间分离度大,那么该数据点的轮廓系数就会越大。由于平衡代表性和多样性就是要求簇内距离最小和簇间距离最大。故本文通过寻找 *N* 个 *Eps* 和 *MinPts* 参数对聚类后最大的轮廓系数来自动确定最优参数对。

针对数据集里的某一簇内样本 $i$,假设样本 $i$ 被聚类到簇 A,该样本个体的轮廓系数 $s_i$ 定义如下:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{5}$$

其中,$a_i$ 为样本 $i$ 与同一类别簇 A 中其余所有点之间的平均距离,体现了代表性;$b_i$ 为样本 $i$ 与距离最近的簇中所有点之间的平均距离,体现了多样性。

求出这一次聚类所有样本个体的轮廓系数的平均值,即可得到一次聚类算法的整体轮廓系数。对于使用 $N$ 个参数对的某次聚类来说,其聚类轮廓系数 $s$ 定义为:

$$s = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{6}$$

其中,$n$ 为聚类内样本的个数。聚类轮廓系数 $s$ 是衡量聚类质量的重要依据,其取值范围为[-1,1],当其得分越接近 1 时,聚类的效果就越好。通过计算 $N$ 个参数对聚类后的得分,就能够确定最优的聚类。

### 1.4 自适应密度聚类流程

自适应密度聚类的算法流程如图 2 所示。生成密度阈值列表后,将 $N$ 个参数对依次进行 DBSCAN 聚类,并计算每一次聚类的轮廓系数,通过寻找聚类轮廓系数最大值的方法,以确定最优的聚类。

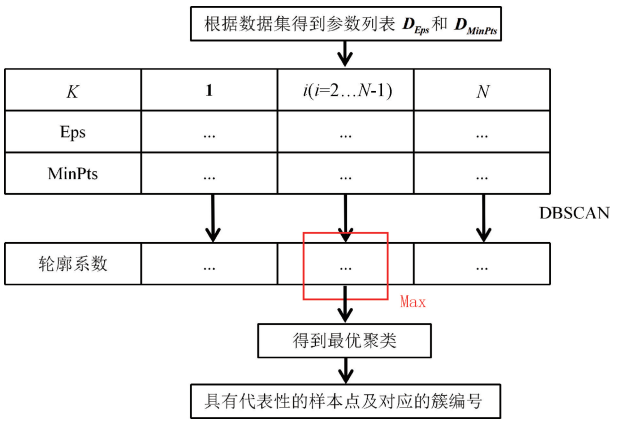



图 2 ADC 算法

Fig. 2 ADC algorithm

自动聚类完成后,就得到了具有代表性的样本点和其对应的簇编号。本文将在后面的 AL 过程中从代表点中抽样而不再考虑边界点和离群点,这样做确保了选择过程始终能够考虑代表性,而簇的编号将用于 AL 的有监督选择阶段。

## 2 结合聚类结果的主动学习方法

### 2.1 结合聚类和 KS 的无监督选择

完成聚类后,使用 KS 算法对代表性样本池中进行多样性抽样。KS 选择的第一个样本点为最靠近质心的样本(即该样本点与其余 $N-1$ 个样本的距离最短),对于剩余的 $N-1$ 个核心点,计算每个代表性样本池中未选择的数据集 U 和已选择数据集 $S$ 之间的最小距离,并从最小距离中选择距离最大的 $U$,如下式所示:

$$\arg\max_{u(\in U)}\left(\min_{\forall s\in S} dist(s,u)\right) \tag{7}$$

这样抽样的目的是从 U 中选择多样性最大的数据点,当数据点周围没有点时,回归函数在该数据点的周围区域很可能是不确定的,故通过 KS 抽样,可以有效地在整个空间内搜索并选择最具多样性的点。

### 2.2 基于多准则高斯过程回归的有监督选择

高斯过程以概率分布的方式对函数进行建模,它为每个输入点提供了完整的不确定性信息,直接给出了预测的方差,从而允许对不确定度进行推断。在这一阶段,使用高斯过程回归模型[17]在剩余代表点中寻找不确定性最大的样本时,应同时考虑该样本与已标记的样本来自不同的簇,这样选择的样本将综合 3 个准则。高斯过程回归模型 $f(X)$ 由均值函数 $\mu(X)$ 和核函数 $K(X,X)$ 决定,记作:

$$f(X) \sim GP(\mu(X), K(X,X)) \tag{8}$$

均值函数和核函数的选择对 GPR 来说至关重要。在均值函数选择方面,为了使得模型具有更强的灵活性和适应性,本文将均值函数 $\mu(X)$ 设置成 0;在核函数选择方面,为了可以更好地捕捉样本之间的连续变化,本文选择常用的高斯核函数 $K(X,X)$:

$$\begin{cases}\mu(X) = 0\\ K(x_i, x_j) = \Sigma^2 \exp\left(-\dfrac{1}{2l^2}(x_i - x_j)(x_i - x_j)'\right)\end{cases} \tag{9}$$

将其中的 $\theta = \{\Sigma, l\}$ 定义为超参数,在 GPR 训练过程中,使用最大对数似然函数[18]训练得到超参数。每输入一个预测点 $X^*$,即可得到该点的预测方差 $\hat{\sigma}$:

$$\hat{\sigma} = K(X^*, X^*) - K(X, X^*)^{\mathrm{T}}(K(X,X) + \Sigma^2\mathbf{I})^{-1}K(X, X^*) \tag{10}$$

将根据样本点的预测方差和对应簇的编号,进行基于多准则高斯过程回归的有监督选择,其伪代码如算法 1 所示。

在伪代码中,首先将方差降序排序,选择方差最大的代表点,并判断该点的簇与已标记样本的簇是否相同,若不同,则将该点进行标注并加入训练集,否则按排序顺延选择下一个点,直到判断到该点的簇从未被选择过,则将该点标注加入到训练集中;在这个过程中需要判断所有的簇是否已经被选择过一遍,如果是,则将已选择的簇集合清空,这样才可以重新选择之前的簇(这样做是由于每个簇内的代表点有多个,即代表点的数量大于簇的数量);最后,使用新的训练集训练 GPR 并得到其性能指标 *RMSE*。以上主动学习过程在每一次迭代中只选择一个样本标注,直至标注样本达到预定数量为止。

**算法 1**　有监督选择时的 GPR 算法

**输入**:$N$ 个未标注的代表性样本 $\{R_n\}_{n=1}^{N}$ 以及对应簇编号 $\{c_n\}_{n=1}^{N}$

已选择的样本集 $\{x,y\}_{train}^{g}$,规定标注样本数量 $G$

测试集 $\{x,y\}_{test}$,初始训练的 GPR

**输出**:GPR 回归模型 $f(x)$

设置已选择的簇 $C_{selected}=\varnothing$,根据 $\{s_n\}_{n=1}^{N}$ 得到簇的共有 $M$ 个;
**while** $G>g$ **do**
　使用 GPR 对 $\{x,y\}_{test}$ 预测,返回不确定度 $\sigma$;
　对 $\sigma$ 降序排序:$\{\sigma_{Ri}\}^{1}>\{\sigma_{Ri}\}^{2}>\cdots>\{\sigma_{Ri}\}^{N}$,$R_i$ 为 $\{R_n\}_{n=1}^{N}$ 中任一点;
　**for** $j=1,2,\cdots,N$ **do**
　　**if** $idx_{\sigma i}\notin C_{selected}$ **then**
　　将 $R_s$ 从 $\{R_n\}_{n=1}^{N}$ 移除后,标注并加入 $\{x,y\}_{train}$,$N=N-1$;
　　将 $idx_{\sigma\max}$ 加入 $C_{selected}$;
　　**break**
　**end**
　　**if** $number(C_{selected})=M$
　　$C_{selected}=\varnothing$;
　**end**
　使用 $\{x,y\}_{train}$ 训练 GPR,并计算新的 $RMSEP$;
**end while**

至此,算法依次通过聚类、无监督选择和有监督选择实现了整个样本的选择过程,基于自适应密度聚类的多标准高斯过程回归主动学习流程图如图 3 所示。

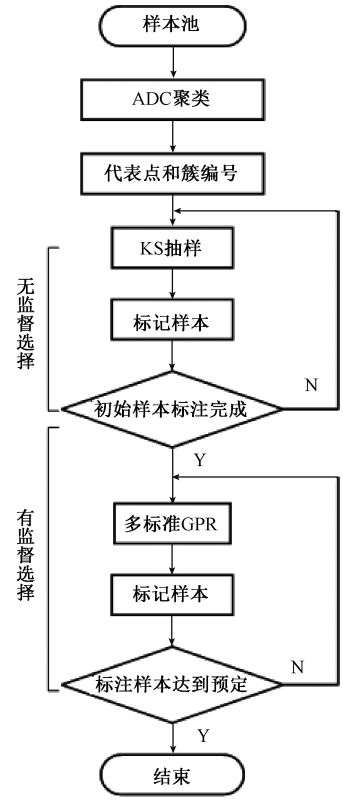



图 3　基于自适应密度聚类的多准则高斯过程回归算法流程图

Fig. 3　Flowchart of Multi-Criterion Gaussian Process Regression Algorithm Based on Adaptive Density Clustering

## 3　实验和结果

### 3.1　数据集测定和光谱测量

一组豆粕含有 928 个样本,一组玉米蛋白粉含有 306 个样本,这些样本是分别从两家工厂采集并分析的,本文将在每一次实验时从样本中随机抽取 50 个样本作为测试集,250 个样本作为训练集。在每个算法的主动学习过程中,先无监督选择 15 个样本作为最开始的回归模型并得出其性能,然后从剩余的样本集中迭代标注 150 个样本并计算其性能指标。

本文将分别测定豆粕数据集和玉米蛋白粉数据集的水分和蛋白质:水分将根据 GB/T 12087－2008,采用 105℃干质量法间接测定,而蛋白质含量采用 GB 5009.5－2010 中凯氏定氮法使用 NYK6160 分析仪(上海亿鸿分析仪器公司)测定,测定的理化值统计参数见表 1。

**表 1　数据集统计参数**

**Table 1　Statistical parameters of the dataset**

| 数据集 | 理化值 | 范围 | 均值 | 标准差 |
|---|---|---|---|---|
| 豆粕 | 水分 | 8.25~14.9 | 12.799 5 | 0.661 0 |
| | 蛋白质 | 37.7~48.6 | 44.218 7 | 1.546 6 |
| 玉米蛋白粉 | 水分 | 5.1~10.9 | 7.852 3 | 1.013 5 |
| | 蛋白质 | 44.4~62.8 | 57.770 7 | 3.459 5 |

豆粕和玉米蛋白粉的光谱数据都是使用二极管阵列分析仪(DA 7200, Perten Instruments, Sweden)在 950~1 650 nm 的近红外区域以 5 nm 为单位通过透光率收集的,每个样品的副本扫描两次。豆粕和玉米蛋白粉的典型光谱如图 4 所示。

### 3.2　比较方法和评价指标

为了检验本文所提出的主动学习算法的有效性,将本文的 ADC-GPR 算法与以下 3 种主动学习算法进行了比较:

1)RS:在整个 AL 阶段从训练集中选择样本随机标注。

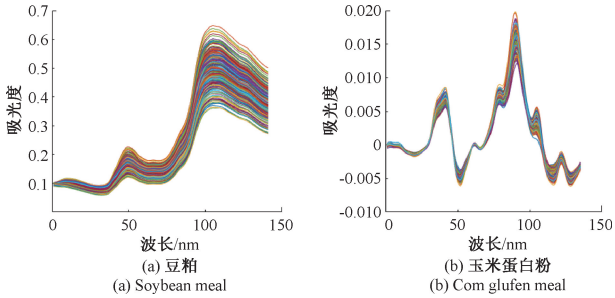



图 4 豆粕和玉米蛋白粉的典型光谱

Fig. 4 Typical spectra of soybean meal and corn protein meal

2) KS:在整个 AL 阶段使用 KS 算法无监督标注样本,这样只考虑了多样性。

3) RD-GPR:在无监督阶段使用 RD 标注聚类内最靠近质心的点,在有监督阶段通过迭代 RD 聚类找到未标注的聚类,然后使用 GPR 返回最不确定的样本并进行标注。

为了进行评估,使用了一种广泛采用的回归指标均方根误差(*RMSEP*)来衡量回归模型在测试集上的性能:

$$RMSEP = \sqrt{\frac{1}{N}\sum_{i=1}^{N_i}(y_i - \hat{y}_i)^2} \tag{11}$$

其中,$N$ 为测试集样本数,$y_i$ 为 $x_i$ 的真实标签,$\hat{y}_i$ 为预测值。*RMSEP* 是最直接反映模型性能的参数,*RMSEP* 的值越小,模型预测效果越好。

### 3.3 实验结果

1) 样本的空间分布

本文能够通过可视化不同算法的样本选择结果来验证基于本文提出的方法的优越性。然而,近红外光谱的特征空间维度很高,因此很难直接将其可视化。所以,对每个数据集进行 PCA 降维,并将其中所有的样本表示为它们对前两个主成分的投影。图 5 为玉米蛋白粉数据集在一次实验时降维得到的散点图,其中不同颜色代表不同的聚类,RD-GPR 是将整个空间内的样本都进行聚类,ADC-GPR 是排除了部分离群值后进行聚类,标注的样本表示最初无监督选取的 15 个样本。

通过观察,可以得出以下结论:

(1) 使用 RS 初始抽样没有对大部分的特征空间进行采样,随机性太强。

(2) 使用 KS 虽然会在特征空间边界附近选取一个或多个样本,这只过度关注了多样性,并有极大可能标注到离群值。

(3) 使用 RD 进行了均匀的初始化,然而离群值的存在也影响到了部分聚类质心的选择,显然部分标注点的周围没有样本,这也造成了部分代表性的缺失。

(4) 使用 ADC 在初始聚类时就排除了无代表性的点

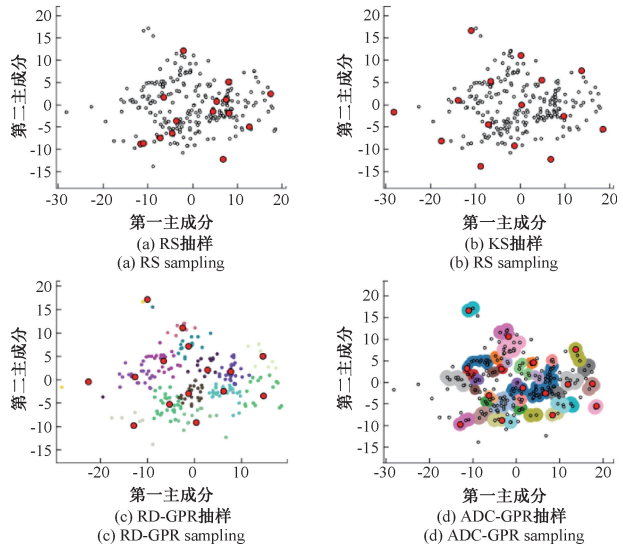



图 5 无监督选择时不同的算法在玉米蛋白粉数据集上的可视化分析

Fig. 5 Visualization analysis of different algorithms during the unsupervised selection on the corn protein dataset

以及离群点,这样选择的样本不仅是聚类内有代表性的点,同时基本覆盖了整个特征空间,相较于 RD 有误差的迭代聚类,仅通过一次聚类的 ADC-DBSCAN 鲁棒性更好。

2) 主动学习结果

为了减少随机误差对实验结果的影响,每种算法对不同划分的样本集分别进行 10 次测试,在两个数据集上,使用不同的算法从训练集挑选出 135 个数据进行标注,得到 10 次试验模型的平均性能结果如表 2 所示。

**表 2 光谱预测模型的 RMSEP**

**Table 2 RMSEP of spectral prediction model**

| 数据集 | 理化值 | RS | KS | RD-GPR | ADC-GPR |
|---|---|---|---|---|---|
| 豆粕 | 水分 | 1.000 8 | 1.003 0 | 0.953 4 | 0.915 9 |
| | 蛋白质 | 0.445 0 | 0.394 7 | 0.357 7 | 0.344 1 |
| 玉米蛋白粉 | 水分 | 0.367 4 | 0.355 1 | 0.340 1 | 0.335 4 |
| | 蛋白质 | 0.412 7 | 0.419 1 | 0.411 2 | 0.409 0 |

显然,在挑选的样本数量一致的情况下,本文提出的方法在两个数据集中的表现都优于其他算法,说明综合了代表性和多样性并排除异常值干扰的算法是有效的;而只基于多样性抽样的 KS 算法在两个数据集上的表现有时略低于综合了代表性和多样性的 RD 算法,说明基于多个标准进行样本选择这一策略是有效的,但 RD 算法并没有排除离群点的干扰。

图 6 展示了不同方法选择标注的训练集对水分和蛋白质的 RMESP 变化趋势。横轴为迭代次数,迭代次数为 1

时的 RMSEP 是各个方法在无监督标注后得到的,之后每一次迭代的 RMSEP 都是添加了 5 个样本后的 RMSEP。

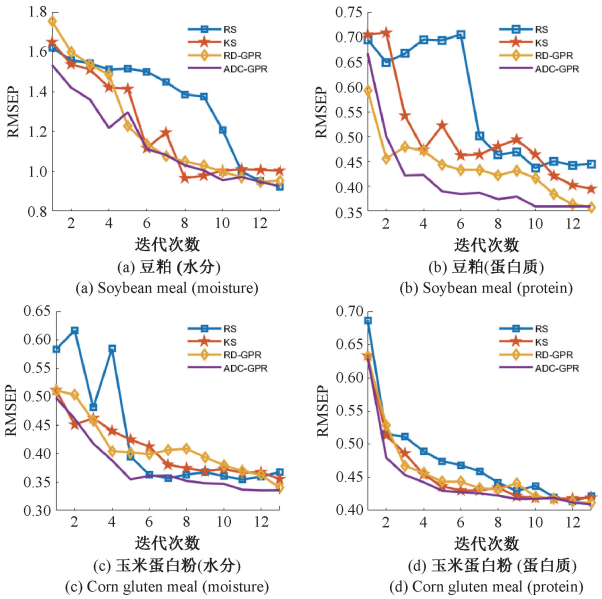



图 6　不同算法在豆粕和玉米蛋白粉数据集上逐步添加样品的模型性能趋势

Fig. 6　The model performance trends of different algorithms when incrementally adding samples in the datasets of soybean meal and corn protein powder

能够得到以下结论:

(1)随着迭代次数的增加,4 种算法都取得了更好的性能(RMSEP 不断变小),这是很直观的,因为标记的训练样本也多,高斯过程回归模型就越可靠。

(2)KS 算法的表现优于 RS 而低于其他 AL 算法,这表明在无监督情况下使用 KS 考虑代表性是有效的,而综合考虑多个标准使得 AL 效率更高。

(3)本文提出的方法优于其他所有算法,使用本文提出的方法进行无监督选择,在大多数数据集上的性能较优,为后续有监督选择提高了标注效率,说明了将自适应密度聚类和高斯过程回归结合起来进行样本选择更有效率。

### 3.4　显著性检验

为了确定实验结果是否具有统计学意义,采用 Wilcoxon 符号秩检验[19],设显著性水平 $\alpha = 0.05$。首先提出 ADC-GPR 不显著的假设,然后比较成对样本的绝对差数,并根据这些差数分配秩次和符号,计算正负符号秩和($T+$和 $T-$)确定 $Z$ 统计量,如下式所示:

$$Z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{12}$$

其中,$T = \min(T+, T-)$。成对样本总数为 151(包括无监督后和 150 次迭代后的 RMSEP),临界值 $|Z_\alpha| = 1.96$,表 3 展示了不同算法的显著性检验结果,表中符号"vs."表示两种方法的比较。从表 3 中可以看出,ADC-GPR 与其他算法的测试值都明显小于临界值,这说明在所有的情况下,原假设都被拒绝,所以 ADC-GPR 得到的结果不是随机的,具有统计学意义。

**表 3　Wilcoxon 显著性检验结果**

**Table 3　Results of Wilcoxon significance test**

| 数据集 | 理化值 | 算法 | $Z$ | 是否显著 |
|---|---|---|---|---|
| 豆粕 | 水分 | ADC-GPR vs. RS | −9.546 | 显著 |
| | | ADC-GPR vs. KS | −7.681 | |
| | | ADC-GPR vs. RDGPR | −7.903 | |
| | 蛋白质 | ADC-GPR vs. RS | −9.546 | 显著 |
| | | ADC-GPR vs. KS | −9.546 | |
| | | ADC-GPR vs. RDGPR | −8.271 | |
| 玉米蛋白粉 | 水分 | ADC-GPR vs. RS | −8.811 | 显著 |
| | | ADC-GPR vs. KS | −8.481 | |
| | | ADC-GPR vs. RDGPR | −9.546 | |
| | 蛋白质 | ADC-GPR vs. RS | −9.533 | 显著 |
| | | ADC-GPR vs. KS | −8.351 | |
| | | ADC-GPR vs. RDGPR | −9.142 | |

为了进一步评价该方法相对于其他算法的优越性,基于显著性水平,从中位数的角度对不同方法的 RMSEP 进行分析,如图 7 所示。其中(S)为豆粕数据集(C)为玉米蛋白粉数据集。ADC-GPR 算法对应的 *RMSEP* 中位数

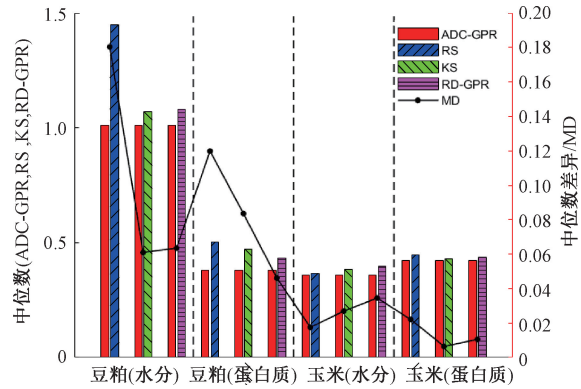



图 7　ADC-GPR 与其他算法的 RMSEP 中位数比较以及两者 RMSEP 差异中位数比较。左 y 轴表示不同方法的 RMSEP 中位数,右 y 轴表示 ADC-GPR 与其他算法 RMSEP 中位数的差异

Fig. 7　Comparison of the median RMSEP values between ADC-GPR and other algorithms, as well as the median difference (MD) in RMSEP between them. The left y-axis represents the median RMSEP values for different methods, while the right y-axis indicates the differences in median RMSEP values between ADC-GPR and other algorithms

低于其他算法,配对样本差异的中位数也小于 0。这表明,与其他算法相比,ADC-GPR 算法具有更低的 RMSEP 值,获得了更好的模型性能。

## 4 结 论

本文提出的 ADC-GPR 算法在主动学习的无监督选择时,有效结合了样本在输入空间上的代表性和多样性,并通过一次自适应聚类得到代表点和簇的编号,这不仅使得初始建立的回归模型具有更强的泛化能力和鲁棒性,还让后续的有监督选择更高效;在主动学习的有监督选择时,使用 GPR 直接返回每个未标记样本的信息性,并结合代表点和簇,使得每一次选择都能兼顾信息性、代表性和多样性,这样做避免了模型过度关注特定类型的样本,有助于减少标注偏差,使得模型更加客观地学习和理解数据。因此,先通过自动聚类得到信息,在 AL 选择新样本时再综合考虑 3 个标准并加以改进,就能够以更少的标注成本获得更高的模型性能。在豆粕和玉米蛋白粉数据集上进行了大量实验,充分验证了 ADC-GPR 算法的有效性。

本文将自适应聚类和主动学习算法结合在一起,先通过聚类进行信息筛选,再对处理后的样本进行主动学习。聚类的效果直接影响着样本选择的效率,由于本文的距离度量均是使用欧氏距离,这对于光谱数据来说可能存在维度灾难的问题,选择适当的距离度量和聚类评价指标是优化聚类的关键。未来本文将在如何对高维光谱数据中更好地自动聚类并提取到有效信息进行主动学习作进一步研究。

## 参考文献


[1] MOGHADDAM H N, TAMIJI Z, LAKEH M A, et al. Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics [J]. Journal of Food Composition and Analysis, 2022, 107: 104343.

[2] 张峰, 汤晓君, 仝昂鑫, 等. 一种基于频率与回归系数相结合的自举柔性收缩变量选择方法[J]. 仪器仪表学报, 2020, 41(1): 64-70.
ZHANG F, TANG X J, GONG A X, et al. A bootstrap flexible contraction variable selection method based on the combination of frequency and regression coefficient[J]. Chinese Journal of Scientific Instrument, 2020, 41(1): 64-70.

[3] SUGIYAMA M, NAKAJIMA S. Pool-based active learning in approximate linear regression[J]. Machine Learning, 2009, 75(3): 249-274.

[4] HE Z, SONG S, SHEN K, et al. Performance enhancement-based active learning sample selection method[J]. Journal of Chemometrics, 2022, 36(3): e3386.

[5] KRISHNAKUMAR A. Active learning literature survey[J]. Tech. rep., Technical reports, University of California, Santa Cruz., 2007, 42.

[6] RAMIREZ-LOPEZ L, SCHMIDT K, BEHRENS T, et al. Sampling optimal calibration sets in soil infrared spectroscopy[J]. Geoderma, 2014, 226: 140-150.

[7] 刘子昂,蒋雪,伍冬睿. 基于池的无监督线性回归主动学习[J]. 自动化学报, 2021, 47(12): 2771-2783.
LIU Z ANG, JIANG X, WU D R. Pool-based unsupervised linear regression active learning[J]. Acta Automatica Sinica, 2021, 47(12): 2771-2783.

[8] WU D R. Pool-based sequential active learning for regression[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 30(5): 1348-1359.

[9] AHMED M, SERAJ R, ISLAM S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[10] CHEN F, ZHANG T, LIU R. An active learning method based on variational autoencoder and dbscan clustering[J]. Computational Intelligence and Neuroscience, 2021, DOI: org/10.1155/2021/9952596.

[11] LI Y, YANG Z, JIAO S, et al. Partition KMNN-DBSCAN algorithm and its application in extraction of rail damage data[J]. Mathematical Problems in Engineering, 2022, DOI:org/10.1155/2022/4699573.

[12] BARTON T, BRUNA T, KORDIK P. Chameleon 2: An improved graph-based clustering algorithm [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(1): 1-27.

[13] CAI W, ZHANG Y, ZHOU J. Maximizing expected model change for active learning in regression[C]. 2013 IEEE 13th International Conference on Data Mining. IEEE, 2013: 51-60.

[14] RAYCHAUDHURI T, HAMEY L G C. Minimisation of data collection by active learning[C]. Proceedings of ICNN'95-International Conference on Neural Networks. 1995, 3: 1338-1341.

[15] GE Z. Active probabilistic sample selection for intelligent soft sensing of industrial processes[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 151: 181-189.

[16] ROUSSEEUW P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis [J].

Journal of Computational and Applied Mathematics, 1987, 20: 53-65.

[17] RASMUSSEN C E, NICKISCH H. Documentation for GPML Matlab Code version 4.2 [EB/OL]. (2020) [2023-11-10]. http://gaussianprocess.org/gpml/code/matlab/doc.

[18] NGUYEN V H, LE T T, TRUONG H S, et al. Applying Bayesian optimization for machine learning models in predicting the surface roughness in single-point diamond turning polycarbonate [J]. Mathematical Problems in Engineering, 2021, 2021: 1-16.

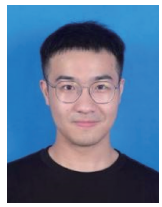
[19] WILCOXON F, KATTI S K, WILCOX R A. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test [J]. Selected Tables in Mathematical Statistics, 1970, 1: 171-259.

## 作者简介

**贺忠海**(通信作者),1994 年于合肥工业大学获得学士学位,1999 年于天津大学获得硕士学位,2002 年于天津大学获得博士学位,现为东北大学秦皇岛分校副教授,主要研究方向为光谱在线检测技术。

E-mail:professorhe@qq.com

**He Zhonghai** (Corresponding author) received his B. Sc. degree in 1994 from Hefei University of Technology, received his M. Sc. degree in 1999 from Tianjin University, and received his Ph. D. degree in 2002 from Tianjin University. Now he is an associate professor in Northeastern University at Qinhuangdao. His main research interest is spectrum on-line detection technology.



**朱温涵**,2022 年于南京工业大学浦江学院获得学士学位,现为东北大学硕士研究生,主要研究方向为机器学习。

E-mail:wenhanzhu2000@163.com

**Zhu Wenhan** received his B. Sc. degree in 2022 from Nanjing Tech University Pujiang Institute, and now he is a master student in Northeastern University. His main research interest is machine learning.

**陈旭旺**,2021 年于天津理工大学获得学士学位,现为东北大学硕士研究生,主要研究方向为机器学习。

E-mail:353504320@qq.com

**Chen Xuwang** received his B. Sc. degree in 2021 from Tianjin University of Technology, and now he is a master student in Northeastern University. His main research interest is machine learning.

**张晓芳**,1996 年于天津大学获得学士学位,2002 年于天津大学获得博士学位,现为北京理工大学副研究员,主要研究方向为主动光学、计算光学、自适应光学。

E-mail:zhangxf@bit.edu.cn

**Zhang Xiaofang** received her B. Sc. degree in1996 from Tianjin University, received her Ph. D. degree in 2002 from Tianjin University, and now she is an associate researcher in Beijing Institute of Technology. Her main research interests include active optics, adaptive optics and computational optics.