基于池的无监督线性回归主动学习

刘子昂 [1]　蒋 雪 [1]　伍冬睿 [1]

摘 要 在许多现实的机器学习应用场景中, 获取大量未标注的数据是很容易的, 但标注过程需要花费大量的时间和经济成本. 因此, 在这种情况下, 需要选择一些最有价值的样本进行标注, 从而只利用较少的标注数据就能训练出较好的机器学习模型. 目前, 主动学习 (Active learning) 已广泛应用于解决这种场景下的问题. 但是, 大多数现有的主动学习方法都是基于有监督场景: 能够从少量带标签的样本中训练初始模型, 基于模型查询新的样本, 然后迭代更新模型. 无监督情况下的主动学习却很少有人考虑, 即在不知道任何标签信息的情况下最佳地选择要标注的初始训练样本. 这种场景下, 主动学习问题变得更加困难, 因为无法利用任何标签信息. 针对这一场景, 本文研究了基于池的无监督线性回归问题, 提出了一种新的主动学习方法, 该方法同时考虑了信息性、代表性和多样性这三个标准. 本文在 3 个不同的线性回归模型 (岭回归、LASSO (Least absolute shrinkage and selection operator) 和线性支持向量回归) 和来自不同应用领域的 12 个数据集上进行了广泛的实验, 验证了其有效性.

关键词 主动学习, 无监督学习, 线性回归, 支持向量回归, LASSO, 岭回归

引用格式 刘子昂, 蒋雪, 伍冬睿. 基于池的无监督线性回归主动学习. 自动化学报, 2021, **47**(12): 2771−2783

DOI 10.16383/j.aas.c200071

Unsupervised Pool-Based Active Learning for Linear Regression

LIU Zi-Ang[1]　JIANG Xue[1]　WU Dong-Rui[1]

Abstract In many real-world machine learning applications, unlabeled data can be easily obtained, but it is very time-consuming and/or expensive to label them. So, it is desirable to be able to select the optimal samples to label, so that a good machine learning model can be trained from a minimum number of labeled data. Active learning (AL) has been widely used for this purpose. However, most existing AL approaches are supervised: they train an initial model from a small number of labeled samples, query new samples based on the model, and then update the model iteratively. Few of them have considered the completely unsupervised AL problem, i.e., starting from zero, how to optimally select the very first few samples to label, without knowing any label information at all. This problem is very challenging, as no label information can be utilized. This paper studies unsupervised pool-based AL for linear regression problems. We propose a novel AL approach that considers simultaneously the informativeness, representativeness, and diversity, three essential criteria in AL. Extensive experiments on 12 datasets from various application domains, using three different linear regression models (ridge regression, LASSO (least absolute shrinkage and selection operator), and linear support vector regression), demonstrated the effectiveness of our proposed approach.

Key words Active learning (AL), unsupervised learning, linear regression, support vector regression, least absolute shrinkage and selection operator (LASSO), ridge regression

Citation Liu Zi-Ang, Jiang Xue, Wu Dong-Rui. Unsupervised pool-based active learning for linear regression. *Acta Automatica Sinica*, 2021, **47**(12): 2771−2783

在机器学习任务中, 往往需要大量的有标签训练数据以获得更好的性能. 但是, 在许多实际应用场景中, 获取未标注的数据相对容易, 标注过程却很困难, 通常需要投入大量的时间和经济成本. 例如, 在语音信号的情感估计问题中, 可以很容易地记录大量语音, 但是要对语音进行 3 个维度 (愉悦度、唤醒度和优势度) 的评估[1], 评估者必须反复倾

收稿日期 2020-02-17　录用日期 2020-08-14

Manuscript received February 17, 2020; accepted August 14, 2020

湖北省技术创新专项基金 (2019AEA171), 国家自然科学基金 (61873321), NSFC-深圳机器人基础研究中心重点项目基金 (U1913207), 科技部政府间国际科技创新合作重点专项基金 (2017YFE0128300) 资助

Supported by Technology Innovation Project of Hubei Province of China (2019AEA171), National Natural Science Foundation of China (61873321), NSFC-Shenzhen Robotics Basic Research Center (U1913207), and International Science and Technology Cooperation Program of China (2017YFE0128300)

本文责任编委 王立威

Recommended by Associate Editor WANG Li-Wei

1. 华中科技大学人工智能与自动化学院图像信息处理与智能控制教育部重点实验室 武汉 430074

1. Ministry of Education Key Laboratory on Image Information Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074

听, 仔细检查. 此外, 由于情感估计具有主观性, 而且部分数据可能只存在细微差异, 通常需要多个评估者, 例如, 素材来自于德国脱口秀节目 Vera am Mittag (VAM) 的语料库[2] 用到 6~17 个评估者, 国际情感数字化声音 (第 2 版) (International affective digitized sounds V2, IADS-2)[3] 用到至少 110 个评估者. 在石油和天然气行业中, 研究油井压裂后 180 天的累计产油量预测问题[4]有利于提高采收率, 输入信息 (油井的压裂参数, 例如油井位置、射孔长度、区域/孔的数量、注入的泥浆/水/砂的体积等) 可以在压裂操作期间记录, 但要获得地面产量 (压裂后 180 天累计产油量), 至少需要等待 180 天.

在很多这样的问题场景中, 如何确定最优的未标注样本进行标注是非常重要的. 主动学习 (Active learning, AL)[5] 可以用于解决此类问题, 它通过选择较少的有价值样本进行标注, 从而获得性能较好的机器学习模型, 减少了数据标注工作.

主动学习可用于分类问题和回归问题. 已有许多用于分类的 AL 方法[5] 相继提出, 但回归任务中的 AL 方法相对较少[6−20]. 这些主动学习回归 (Active learning for regression, ALR) 方法有基于流或基于池[13] 的应用场景, 本文考虑的是后者, 即给定一个未标注样本池, ALR 需要最优地选择一些样本进行标注, 从中训练出一个较好的线性回归模型.

现有的大多数基于池的 ALR 方法[6−12, 14−15, 19−20] 都考虑的是较简单的有监督场景, 即能够获得少量带标签的样本, 建立初始的回归模型, 然后根据模型选择后续的样本交给专家进行标注. 经过调研, 我们只发现在 4 项研究[13, 16−18] 中明确考虑了完全无监督的基于池的 ALR 场景 (将在下一节中详细介绍), 即在没有任何标签信息的情况下, 选择最有价值的初始样本进行标注, 这也是本文的重点.

具体地说, 本文考虑以下问题: 在给定大小为 N 的未标注样本池中, 如何最佳地选择初始的 M 个样本进行标注, 从而构建较好的线性回归模型? 这里的 M 是通过用户指定 (通常, 随着 M 变大, ALR 的优势会逐渐减弱). 在本文中, 我们仅关注线性回归模型.

针对上述问题, 本文提出了一种基于信息性−代表性−多样性 (Informativeness-representativeness-diversity, IRD) 的 ALR 方法. 通过同时考虑主动学习中的 3 个重要标准[17]: 信息性、代表性和多样性, 从而确定要查询的 M 个初始样本. 在 3 种不同的线性回归模型和来自不同应用领域的 12 个数据集上的实验表明, 与 3 种已提出的 ALR 方法相比, 本文提出的 IRD 方法所选择的 M 个样本可以实现更好的性能.

本文的主要贡献是:

1) 提出了一种无监督的 ALR 方法, 同时考虑要选择的 M 个样本的信息性、代表性和多样性 (这里 $M \le d+1$, 其中 d 是特征维数). 根据调研, 目前文献中 ALR 的信息性计算都必需输出信息, 还没有无需输出信息的信息性计算方法. 因此, 本文提出的方法是首个可考虑所选样本信息性的完全无监督 ALR 方法, 具有重要的理论创新性.

2) 提出了一种迭代式的 ALR 方法, 同时考虑代表性和多样性, 在 $M > d+1$ 时选择另外的 $M-d-1$ 个样本.

3) 在 3 种常见的线性回归模型和 12 个真实数据集上的大量实验, 证明了所提出的 IRD 方法的优越性能.

本文的组织架构如下: 第 1 节介绍 3 种现有的无监督 ALR 方法, 并指出了它们的局限性; 第 2 节详细介绍本文提出的 IRD 算法; 第 3 节对在 12 个数据集上的实验进行了讨论和分析; 最后, 第 4 节给出本文的结论.

1 现有的基于池的无监督 ALR 方法

Wu[17] 提出了以下 3 个基于池的有监督 ALR 方法应该考虑的标准. 这些标准也适用于无监督的 ALR 问题:

1) 信息性. 可以通过不确定性 (熵、到决策边界的距离、预测的置信度等)、模型改变期望 (Expected model change)、误差缩减期望 (Expected error reduction) 等来度量.

2) 代表性. 可以通过与目标样本相似或接近的样本数量来度量. 跟目标样本相似或接近的样本越多, 那么该目标样本代表性越强. 此标准优先选择靠近簇中心的样本, 或者分布稠密处的样本, 可防止选择离群点. 例如在图 1 中, 需要构建一个回归模型从输入 x_1 和 x_2 中预测输出. 两个实心的点是已经选中的待标注样本, 现在需要从空心的点中选

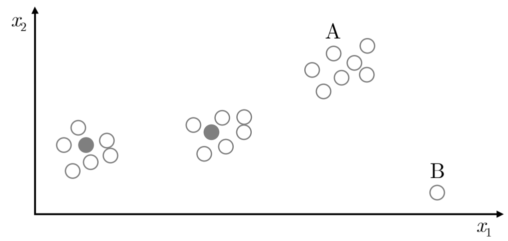


图 1 基于池的 ALR 中样本的代表性与多样性[17]

Fig.1 Illustration of representativeness and diversity in pool-based ALR[17]

出第 3 个待标注样本. 很显然, 从包含 "A"的簇中选出一个样本比选择样本 "B"更好, 因为 "A"处样本稠密, 代表性强, 而样本 "B"远离其他样本, 很可能是个离群点, 选出后对构建回归模型有害无利, 反而不如只用最初选出的两个样本的效果.

3) 多样性. 所选样本应尽可能分散在整个输入空间中, 而不是一个小的局部中, 以便学习一个良好的全局模型. 例如图 1 中, 绝大部分样本分布在 3 个簇中, 那么选择 3 个样本时, 应该从 3 个簇中分别选出一个, 让样本更加多样, 而不是只从其中一个或两个簇中选.

多样性和代表性经常会有一定的冲突, 所以应该折中平衡考虑. 一个常用的方法是先对所有待选样本聚类, 然后选取不同簇中靠近簇中心的样本, 如下文中的 RD 方法.

接下来, 我们介绍 3 种在文献中已有的基于池的无监督 ALR 方法, 并对照以上 3 个标准对其进行检查. 假设数据池由 N 个 d 维未标注样本 $\boldsymbol{x}_n = [x_{n,1}, x_{n,2}, \cdots, x_{n,d}]^{\rm T} \in \mathbf{R}^{d\times 1}$, $n=1,2,\cdots,N$ 组成, 用户将从中选择 M 个进行标注.

1.1 P-ALICE

Sugiyama 等[13] 提出了一种基于泛化误差条件期望的重要性加权最小二乘方法 (Pool-based active learning using the importance-weighted least-squares learning based on conditional expectation of the generalization error, P-ALICE), 这是一种无监督的 ALR 算法, 用于选择要标注的初始少量样本. 其主要思想是识别 M 个样本及其相关权重, 计算训练样本与测试样本之间的协变量偏移, 由这 M 个样本构建的加权线性回归模型可以最小化 N 个样本上的均方损失估计值.

设

$$\boldsymbol{U} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^{\rm T} \in \mathbf{R}^{d\times d} \tag{1}$$

其中, $\boldsymbol{U}^{-1} \in \mathbf{R}^{d\times d}$是 $\boldsymbol{U}$的逆, $\boldsymbol{U}_{i,j}^{-1}$表示 $\boldsymbol{U}^{-1}$的第 (i,j) 个元素. P-ALICE 首先定义关于 λ 的重采样偏差函数

$$b^{\lambda}(\boldsymbol{x}_n) = \left(\sum_{i,j=1}^{d} \boldsymbol{U}_{i,j}^{-1} x_{n,i} x_{n,j}\right)^{\lambda} \tag{2}$$

其中, $\lambda \in [0,1]$, 对于每个不同的 λ, 从样本池中选择 M 个未标注样本的概率与 $b^{\lambda}(\boldsymbol{x}_n)$ 成正比. 将所选样本表示为 $\{\boldsymbol{x}_m^{\lambda}\}_{m=1}^{M}$, 那么, 在 N 个样本上的均方损失可以如下进行估计:

$$Q(\lambda) = {\rm trace}[\boldsymbol{U}\boldsymbol{L}^{\lambda}(\boldsymbol{L}^{\lambda})^{\rm T}] \tag{3}$$

其中

$$\boldsymbol{L} = [\boldsymbol{X}^{\lambda}\boldsymbol{W}^{\lambda}(\boldsymbol{X}^{\lambda})^{\rm T}]^{-1}\boldsymbol{X}^{\lambda}\boldsymbol{W}^{\lambda} \in \mathbf{R}^{d\times M} \tag{4}$$

其中

$$\boldsymbol{X}^{\lambda} = [\boldsymbol{x}_1^{\lambda}, \boldsymbol{x}_2^{\lambda}, \cdots, \boldsymbol{x}_M^{\lambda}] \in \mathbf{R}^{d\times M} \tag{5}$$

$$\boldsymbol{W}^{\lambda} = {\rm diag}\left\{[b^{\lambda}(\boldsymbol{x}_1^{\lambda}), b^{\lambda}(\boldsymbol{x}_2^{\lambda}), \cdots, b^{\lambda}(\boldsymbol{x}_M^{\lambda})]^{-1}\right\} \in \mathbf{R}^{M\times M} \tag{6}$$

然后, P-ALICE 确定 $\lambda^* = \arg\min_{\lambda} Q(\lambda)$, 即选择使得在 N 个样本上均方损失最小的 λ^*, 并选择相应的 $\{\boldsymbol{x}_m^{\lambda^*}\}_{m=1}^{M}$ 进行标注. 由于每个选择的 $\boldsymbol{x}_i^{\lambda^*}$ 关联一个权重 $b^{\lambda^*}(\boldsymbol{x}_i^{\lambda^*})$, P-ALICE 最后会计算得到一个加权线性回归模型, 以适应训练样本和测试样本之间的协变量偏移.

综上所述, 对照 ALR 的 3 个标准, P-ALICE 只考虑了信息性 (均方损失估计值), 没有考虑代表性和多样性.

1.2 GSx

Yu 等[16] 提出了一种基于贪婪采样 (Greedy sampling, GS) 的 ALR 算法. 在给定一个初始未标注样本的情况下, GS 不需要任何标签信息就可以选择其他未标注的样本. 但是, GS 初始至少需要一个确定的未标注样本, 文中并没有对第 1 个样本的选取进行解释. 因此, Wu 等[18] 提出了 GSx 方法, 将第 1 个样本指定为最接近 N 个未标注样本中心的样本. 接下来对 GSx 算法进行介绍.

不失一般性, 假设 GSx 已选择了前 M_0 个样本. 对于剩余的 $N-M_0$个未标注样本 $\{\boldsymbol{x}_n\}_{n=M_0+1}^{N}$, GSx 首先计算每个样本到 M_0 个已选样本的距离, 即

$$d_{nm} = ||\boldsymbol{x}_n - \boldsymbol{x}_m||,\quad m=1,\cdots,M_0; n=M_0+1,\cdots,N \tag{7}$$

然后, 计算 $\boldsymbol{x}_n$ 到已选 M_0 个样本的最小距离 $\underline{d}_n$, 即

$$\underline{d}_n = \min_{m} d_{nm},\quad n=M_0+1,\cdots,N \tag{8}$$

再选择具有最大 $\underline{d}_n$ 的样本进行标注. 重复此过程, 直到选择的样本数量达到 M .

综上所述, 对照 ALR 的 3 个标准, GSx 仅考虑多样性, 没有考虑信息性和代表性.

1.3 RD

Wu[17] 提出了一种基于样本代表性 (Representativeness) 和多样性 (Diversity) 的方法, 简称 RD.

RD 主要由两部分组成: 一部分是初始化 (无监督过程), 另一部分是后续迭代 (有监督过程). RD 的无监督过程首先对 N 个未标注样本进行 k-means

聚类 ($k=d+1$), 然后选择最接近每个聚类中心的样本进行标注. 在文献 [15] 中也使用过类似的方法.

顾名思义, RD 在初始化时仅考虑代表性和多样性, 没有考虑信息性.

1.4 小结

表 1 中总结了 P-ALICE、GSx 和 RD 考虑的标准. 可见, 这 3 种方法都只考虑了 ALR 的 3 个基本标准中的 1 个或 2 个. 因此, 仍有改进的空间.

表 1 基于池的无监督 ALR 方法中考虑的标准

Table 1 Criteria considered in the three existing and the proposed unsupervised pool-based ALR approaches

	方法	信息性	代表性	多样性
现有方法	P-ALICE	✓	–	–
	GSx	–	–	✓
	RD	–	✓	✓
本文方法	IRD	✓	✓	✓

2 IRD 算法

本节对本文提出的基于池的无监督 ALR 算法——IRD 进行介绍. 顾名思义, IRD 同时考虑信息性、代表性和多样性.

设 M 为要选择的样本数量, d 为特征维数. 接下来分别讨论 IRD 算法在 3 种情形 ($M=d+1$, $M<d+1$, 以及 $M>d+1$) 下的实现.

2.1 情形 1: $M=d+1$

对于 d 维特征数据, 通常需要选择至少 $d+1$ 个样本来构造一个线性回归模型 $f(\boldsymbol{x})=\boldsymbol{x}^{\rm T}\boldsymbol{w}+b$, 其中 $\boldsymbol{w}\in\mathbf{R}^{d\times1}$ 为回归系数, b 为偏置. 接下来从 $d=2$ 维的特殊样本开始, 对 IRD 的基本思想解释说明 (图 2).

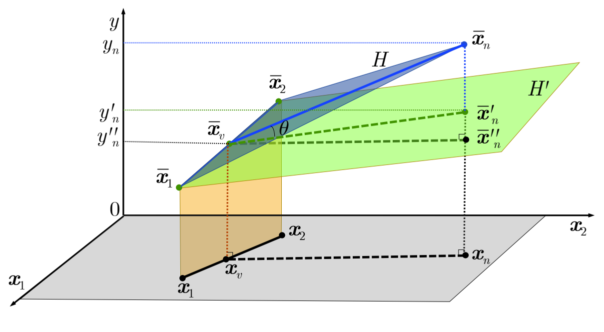


图 2 当 $d=2$ 时 IRD 算法图示

Fig. 2 Illustration of IRD when $d=2$

假设前两个未标注样本 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 已确定, 现在需要选择第三个样本 $\{\boldsymbol{x}_n\}_{n=3}^N$. 为了便于说明, 记 $\bar{\boldsymbol{x}}_n=[\boldsymbol{x}_n;y_n]\in\mathbf{R}^{(d+1)\times1}$, $n=1,\cdots,N$.

假设 H' 为通过 $\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$ 的 d 维最佳流形, 并且能够最佳地拟合其余的 $N-2$ 个样本. 在无监督问题中, H' 是未知的, 但如果给定所有 $\bar{\boldsymbol{x}}_n$, 并要求 H' 必须通过 $\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$, 那么一定会存在这样的 H'.

任何的 $\bar{\boldsymbol{x}}_n$ ($n=3,\cdots,N$) 和 $\{\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2\}$ 可以确定一个二维流形 H, 如图 2 所示. 那么 H 和 H' 会在 $\overrightarrow{\bar{\boldsymbol{x}}_1\bar{\boldsymbol{x}}_2}$ 上相交. 在 $\overrightarrow{\bar{\boldsymbol{x}}_1\bar{\boldsymbol{x}}_2}$ 上找到一点 $\bar{\boldsymbol{x}}_v$ 使得 $\overrightarrow{\bar{\boldsymbol{x}}_1\bar{\boldsymbol{x}}_2}\perp\overrightarrow{\bar{\boldsymbol{x}}_v\bar{\boldsymbol{x}}_n}$, 其中, $\bar{\boldsymbol{x}}_v\bar{\boldsymbol{x}}_n\in H$, 如图 2 所示.

那么, 为了获得最佳性能, 则希望 H 能够尽可能地接近 H'. 接下来将详细解释如何识别最佳的 $\boldsymbol{x}_n^*$.

设 $\overrightarrow{\bar{\boldsymbol{x}}_n\bar{\boldsymbol{x}}'_n}$ 通过 $\bar{\boldsymbol{x}}_n$, 并且平行于 y 轴, 和 H' 相交于 $\bar{\boldsymbol{x}}'_n$. 那么, 可以用 $\overrightarrow{\bar{\boldsymbol{x}}_v\bar{\boldsymbol{x}}_n}$ 和 $\overrightarrow{\bar{\boldsymbol{x}}_v\bar{\boldsymbol{x}}'_n}$ 之间的角度[1] θ 来表示 H 和 H' 间的接近程度.

从图 2 中, 可以得到:

$$|\theta|\propto\frac{|\bar{\boldsymbol{x}}_n-\bar{\boldsymbol{x}}'_n|}{|\bar{\boldsymbol{x}}_v-\bar{\boldsymbol{x}}'_n|}=\frac{|y_n-y'_n|}{\sqrt{(y'_n-y''_n)^2+|\bar{\boldsymbol{x}}_v-\bar{\boldsymbol{x}}''_n|^2}}\tag{9}$$

其中, 分子部分 $|y_n-y'_n|$ 取决于 y_n 和 y'_n, 这在无监督问题中是未知的, 只能忽略. 分母部分由两项组成, 第 1 项 $y'_n-y''_n$ 也是未知的, 同样忽略它. 第 2 项 $|\bar{\boldsymbol{x}}_v-\bar{\boldsymbol{x}}''_n|$ 可以在无监督的 ALR 中计算得到. 因为 $\bar{\boldsymbol{x}}_v$ 和 $\bar{\boldsymbol{x}}''_n$ 具有相同的 y, 所以 $|\bar{\boldsymbol{x}}_v-\bar{\boldsymbol{x}}''_n|$ 与 y 无关, 等于 $|\boldsymbol{x}_v-\boldsymbol{x}_n|$, 即 $\boldsymbol{x}_n$ 到 $\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_2}$ 的距离, 如图 2 所示.

因此, 基于以上推导和可以在基于池的无监督 ALR 中使用的所有信息, 可以近似得到:

$$|\theta|\propto\frac{1}{|\boldsymbol{x}_v-\boldsymbol{x}_n|}\tag{10}$$

式 (10) 从希望 H 和 H' 尽可能接近推导而来, 因此这考虑了 $\boldsymbol{x}_n$ 的信息性. 此外, $|\boldsymbol{x}_v-\boldsymbol{x}_n|$ 也可以看作从 $\boldsymbol{x}_n$ 到已确定样本 (在这里也就是 $\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$) 之间的距离. 要使 θ 变小, 则需要 $|\boldsymbol{x}_v-\boldsymbol{x}_n|$ 尽可能大, 即式 (10) 也保证了所选样本之间的多样性. 综上所述, 使用式 (10) 选择第 3 个样本时同时考虑了信息性和多样性.

但是, 如果仅使用式 (10) 作为选择第 3 个样本的准则, 它将始终选择距离 $\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_2}$ 最远的样本, 那很有可能是一个离群点. 为了同时考虑到代表性, 可以计算从 $\boldsymbol{x}_n$ 到 N 个样本的平均距离, 结合到式 (10) 中, 从而选择最佳的样本进行标注[2]:

[1] θ 并不是 H 和 H' 之间的真实角度, 当满足 $\overrightarrow{\bar{\boldsymbol{x}}_v\bar{\boldsymbol{x}}'_n}\perp\overrightarrow{\bar{\boldsymbol{x}}_1\bar{\boldsymbol{x}}_2}$ 时, θ 才是 H 和 H' 间的夹角, 通常 $\bar{\boldsymbol{x}}'_n$ 并不满足这一条件. 然而, 如同 H 和 H' 之间的真实角度, θ 会随着 H' 和 H 越来越接近而单调递减, 因此可以用它来度量 H' 和 H 之间的接近程度. 我们使用这样的 θ, 而不是 H 和 H' 之间的真实角度来进行计算, 将会使得推导更加容易.

[2] 我们还考虑了其他兼顾代表性、信息性和多样性的方法, 例如 $\boldsymbol{x}_n^*=\arg\min_{\boldsymbol{x}_n}(\frac{1}{N}\sum_{i=1}^N|\boldsymbol{x}_i-\boldsymbol{x}_n|^2)^{1/2}+\lambda\cdot|\boldsymbol{x}_v-\boldsymbol{x}_n|$, 但这种方法会引入超参数 λ, 并且我们的实验表明, 从最佳的 λ 获得的性能要比式 (11) 差. 因此, 由于其简单性和准确性, 我们最终使用式 (11) 作为选择准则.

$$\boldsymbol{x}_n^* = \arg\min_{\boldsymbol{x}_n} \frac{\left(\dfrac{1}{N}\sum\limits_{i=1}^{N}|\boldsymbol{x}_i-\boldsymbol{x}_n|^2\right)^{\frac{1}{2}}}{|\boldsymbol{x}_v-\boldsymbol{x}_n|} \tag{11}$$

当 $d>2$ 时, 同理, 可以用 $(d-1)$ 维流形 C 来代替 $\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_2}$, 所有已确定的 d 个样本 $\{\bar{\boldsymbol{x}}_t\}_{t=1}^{d}$ 都位于这个流形上. 那么, 可以将式 (11) 改写为

$$\boldsymbol{x}_n^* = \arg\min_{\boldsymbol{x}_n} \frac{\left(\dfrac{1}{N}\sum\limits_{i=1}^{N}|\boldsymbol{x}_i-\boldsymbol{x}_n|^2\right)^{\frac{1}{2}}}{\mathrm{dist}(\boldsymbol{x}_n, C)} \tag{12}$$

其中, $\mathrm{dist}(\boldsymbol{x}_n, C)$ 表示从 $\boldsymbol{x}_n$ 到流形 C 的距离.

为了计算 $\mathrm{dist}(\boldsymbol{x}_n, C)$, 首先需要找到一个垂直于 C 的向量 $\boldsymbol{w}\in\mathbf{R}^{d\times1}$, 即满足

$$\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_d \\ 1 & 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \boldsymbol{0} \tag{13}$$

那么,

$$\mathrm{dist}(\boldsymbol{x}_n, C) = \frac{\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{w}+b}{|\boldsymbol{w}|} \tag{14}$$

类似于式 (12) 的方法尚未出现在 ALR 中. 在用于分类的 AL 方法中, 有一些用于选择最接近当前分类边界的样本 (即具有最大不确定性的样本) 进行标注的方法[21−23], 但是它们与式 (12) 有 3 个显著区别:

1) 式 (12) 适用于回归问题, 而目前文献中的选择最接近当前分类边界样本的方法[21−23] 都是针对分类问题的.

2) 式 (12) 是完全无监督的, 即只需要知道样本的特征信息而无需知道其输出. 文献中针对分类问题的方法[21−23] 都是有监督的, 要求必须提供一些有标注的样本来初始化分类器, 从而计算待选样本到分类界面的距离.

3) 式 (12) 同时考虑了待选样本的信息性和多样性, 而传统分类问题中的方法[21−23] 只考虑信息性.

2.2 迭代改进

上述方法是在前 d 个样本确定的情况下, 选择第 $(d+1)$ 个样本. 第 $(d+1)$ 个样本的最优性还取决于前 d 个样本的最优性.

因此, 本小节提出一种交替优化方法, 以迭代方式优化第 $d+1$ 样本: 首先通过 GSx 或 RD 算法确定前 d 个样本, 再通过式 (12) 选择第 $(d+1)$ 个样本. 然后反复固定 d 个样本, 使用式 (12) 优化每个样本 $\boldsymbol{x}_t$ $(t=1,\cdots,d+1)$. 重复此过程, 直到选择的样本收敛或者达到最大迭代次数.

在 $M=d+1$ 情形下, IRD 的伪代码如算法 1 所示.

算法 1. $\boldsymbol{M=d+1}$ **时的 IRD 算法**

输入. N 个未标注样本, $\{\boldsymbol{x}_n\}_{n=1}^{N}$, 其中 $\boldsymbol{x}_n\in\mathbf{R}^{d\times1}$; $c_{\max}$, 最大迭代次数.

输出. $\{\boldsymbol{x}_t\}_{t=1}^{M}$, $M=d+1$ 个最佳的样本集.

用 GSx 或者 RD 算法获得初始的 M 个样本;
将 M 个样本的索引保存到矩阵 P 的第 1 行;
$c=0$;
while $c<c_{\max}$ **do**
  记 M 个已选择本为 $\{\boldsymbol{x}_t\}_{t=1}^{M}$, 其余样本记为 $\{\boldsymbol{x}_n\}_{n=M+1}^{N}$;
  for $t=1,\cdots,M$ **do**
    设 $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1},\cdots,\boldsymbol{x}_M\}$ 为流形 C 上的 d 个点;
    根据式 (14) 计算每个 $\boldsymbol{x}_n$ 到流形 C 的距离 $\mathrm{dist}\ (\boldsymbol{x}_n, C)$, $n=M+1,\cdots,N$;
    将从式 (12) 计算的 $\boldsymbol{x}_n^*$ 记为 $\boldsymbol{x}_t$;
  end for
  if 当前 M 个样本的索引和矩阵 P 的任一行相同 **then**
    Break;
  else
    将当前 M 个样本的索引保存到矩阵 P 的下一行;
  end if
  $c=c+1$;
end while.

2.3 情形 2: $\boldsymbol{M<d+1}$

情形 1 中考虑的是 $M=d+1$, 即所选样本数量刚好等于特征数加 1, 这是一种非常特殊的情况. 实际上 M 可能小于 $d+1$, 在这种情况下, 式 (12) 中的 $d-1$ 维流形 C 不能唯一确定, 因此 $\boldsymbol{x}_n^*$ 不能直接由式 (12) 得到.

对于这种情形, 本小节提出一种新的处理方法: 首先, 对 N 个样本 $\boldsymbol{x}_n$ 进行主成分分析 (Principal component analysis, PCA), 并确定前 $M-1$ 个主成分, 然后将每个 $\boldsymbol{x}_n$ 替换为其在 $M-1$ 个主成分方向的投影. 则式 (12) 可以在转换后的 $\boldsymbol{x}_n$ 上进行计算.

在 $M<d+1$ 情形下, IRD 的伪代码如算法 2 所示.

算法 2. $\boldsymbol{M<d+1}$ **时的 IRD 算法**

输入. N 个未标注样本, $\{\boldsymbol{x}_n\}_{n=1}^{N}$, 其中 $\boldsymbol{x}_n\in\mathbf{R}^{d\times1}$; M, 要选择的样本数量 $(M<d+1)$; $c_{\max}$, 最大迭代次数.

输出. $\{\boldsymbol{x}_t\}_{t=1}^{M}$, 最佳的样本集.

对 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]^{\mathrm{T}} \in \mathbf{R}^{N\times d}$ 进行 PCA 处理;
将每一个样本 $\boldsymbol{x}_n \in \mathbf{R}^{d\times 1}$ 替换为前 $M-1$ 个主成分方向的投影;
使用算法 1 得到 M 个样本.

2.4 情形 3: $M > d+1$

本小节考虑 $M > d+1$ 的情况.

首先, 使用算法 1 初始化 $d+1$ 个样本, 然后继续确定另外的 $M-d-1$ 个样本: 使用 k -means 聚类 ($k = M-d-1$) 在剩余的 $N-d-1$ 个样本中得到 $M-d-1$ 个簇, 然后从每个簇中选择一个样本. 这类似于 RD 方法, 但本文提出一种改进方法: 不是直接选择最接近每个聚类中心的样本, 而是使用迭代的方法来选择剩余的 $M-d-1$ 个样本.

不失一般性, 假设前 $d+1$ 个样本已通过算法 1 确定, 接下来的 $M-d-2$ 个样本也暂时确定 (例如最接近其簇中心的样本), 要优化将从第 $(M-d-1)$ 个簇中选择的第 M 个样本. 对于这个簇中的每个 $\boldsymbol{x}_n$, 将其与这个簇中其他样本的平均距离的倒数作为其代表性. 记 S 为第 $(M-d-1)$ 个簇中样本的索引. 则 $\boldsymbol{x}_n$ 的代表性可以表示为

$$R(\boldsymbol{x}_n) = \frac{|S|}{\sum\limits_{i\in S} |\boldsymbol{x}_n - \boldsymbol{x}_i|^2} \tag{15}$$

其中, $|S|$ 是 S 中元素的个数.

将 $\boldsymbol{x}_n$ 到 $M-1$ 个已选样本的最小距离作为其多样性的度量, 即

$$D(\boldsymbol{x}_n) = \min_{t=1,\cdots,M-1} |\boldsymbol{x}_n - \boldsymbol{x}_t| \tag{16}$$

再结合代表性和多样性:

$$RD(\boldsymbol{x}_n) = R(\boldsymbol{x}_n) \times D(\boldsymbol{x}_n) \tag{17}$$

选择样本

$$\boldsymbol{x}_n^* = \arg\max_{\boldsymbol{x}_n} RD(\boldsymbol{x}_n) \tag{18}$$

来代替第 M 个样本. 对每一个 $\{\boldsymbol{x}_t\}_{t=d+2}^{M}$ 重复此过程, 直到不再更新样本或达到最大迭代次数.

在 $M > d+1$ 情形下, IRD 的伪代码如算法 3 所示.

算法 3. $M > d+1$ 时的 IRD 算法

输入. N 个未标注样本, $\{\boldsymbol{x}_n\}_{n=1}^{N}$, 其中 $\boldsymbol{x}_n \in \mathbf{R}^{d\times 1}$; M, 要选择的样本数量 ($M > d+1$); $c_{\max}$, 最大迭代次数.

输出. $\{\boldsymbol{x}_t\}_{t=1}^{M}$, 待标注的最佳的样本集合.

使用算法 1 确定前 $d+1$ 个样本, 并把它们记为 $\{\boldsymbol{x}_t\}_{t=1}^{d+1}$;
在其余的 $N-d-1$ 个样本上进行 k -means聚类, 其中, $k = M-d-1$;
初始化 $\boldsymbol{x}_t$ 为第 $(t-d-1)$ 个簇中最接近中心点的样本, $t = d+2, \cdots, M$;
将 M 个样本的索引保存到矩阵 P 的第 1 行;
$c = 0$;
while $c < c_{\max}$ **do**
  for $t = d+2, \cdots, M$ **do**
    将式 (18) 计算的 $\boldsymbol{x}_n^*$ 记为 $\boldsymbol{x}_t$;
  end for
  if 当前 M 个样本的索引和矩阵 P 的任一行相同 **then**
    Break;
  else
    将当前 M 个样本的索引保存到矩阵 P 的下一行;
  end if
  $c = c+1$;
end while.

3 实验结果与分析

为了验证文中提出的基于池的无监督 ALR 算法 IRD 的有效性, 在 12 个数据集和 3 种线性回归模型上进行了实验. 本节将对实验结果进行分析讨论.

3.1 数据集

本文使用了 12 个来自不同应用领域的数据集进行实验, 其基本情况如表 2 所示.

其中 9 个数据集来自 UCI 机器学习数据库[3], 2 个来自 CMU StatLib Datasets Archive[4]. 这些数据集在其他的 ALR 实验[7−8, 16−18] 中也用过. 其中两个数据集 (autoMPG 和 CPS) 同时包含数字型和类别型特征, 因此首先使用 one-hot 编码进行处理, 将类别型特征转换为数字型特征, 再进行 ALR实验.

本文还使用了一个公开的情感计算数据集: VAM (Vera am Mittag) 数据库[2], 这个数据库也得到了广泛应用[20, 24−27]. 它包含来自 47 位讲话者的 947 条情感语音样本, 从中提取了 46 个声学特征[26−27], 其中包括 9 个音高特征、5 个持续时间特征、6 个能量特征和 26 个 MFCC 特征, 对情感的 3 个维度 (愉悦度、唤醒度和优势度) 进行预测. 在本文实验中, 只将唤醒度作为回归输出.

对于每个数据集, 采用 z-score 对输入的每一维进行标准化.

3.2 对比算法

本文将 IRD ($c_{\max} = 5$) 与以下 4 种算法进行

[3]http://archive.ics.uci.edu/ml/index.php
[4]http://lib.stat.cmu.edu/datasets/

表 2 12 个数据集的总结

Table 2 Summary of the 12 regression datasets

数据集	来源	样本个数	原始特征个数	数字型特征个数	类别型特征个数	总的特征个数
Concrete-CS[a]	UCI	103	7	7	0	7
Yacht[b]	UCI	308	6	6	0	6
autoMPG[c]	UCI	392	7	6	1	9
NO2[d]	StatLib	500	7	7	0	7
Housing[e]	UCI	506	13	13	0	13
CPS[f]	StatLib	534	10	7	3	19
EE-Cooling[g]	UCI	768	7	7	0	7
VAM-Arousal[h]	ICME	947	46	46	0	46
Concrete[i]	UCI	1030	8	8	0	8
Airfoil[j]	UCI	1503	5	5	0	5
Wine-Red[k]	UCI	1599	11	11	0	11
Wine-White[l]	UCI	4898	11	11	0	11

[a] https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test
[b] https://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics
[c] https://archive.ics.uci.edu/ml/datasets/auto+mpg
[d] http://lib.stat.cmu.edu/datasets/
[e] https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
[f] http://lib.stat.cmu.edu/datasets/CPS_85_Wages
[g] http://archive.ics.uci.edu/ml/datasets/energy+efficiency
[h] https://dblp.uni-trier.de/db/conf/icmcs/icme2008.html
[i] https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
[j] https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise
[k] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

比较:

1) 随机采样 (Random sampling, RS): 随机选择 M 个样本进行标注.

2) P-ALICE: 在第 1.1 节中已经介绍. 参数 λ 从 {0, 0.1, 0.2, 0.3, 0.4, 0.41, 0.42, ···, 0.59, 0.6, 0.7, 0.8, 0.9, 1} 中选择最佳的一个.

3) GSx: 在第 1.2 节中已经介绍.

4) RD: 在第 1.3 节中已经介绍.

3.3 评价过程

对于每个数据集, 每一次重复实验随机选择 50% 的样本作为样本池, 其余 50% 作为测试集, 每种算法从完全未标注的样本池中选择 $M \in [5, 15]$ 个样本进行标注, 然后建立线性回归模型. 所有实验均重复 100 次.

在测试集上进行预测, 使用均方根误差 (Root mean squared error, RMSE) 和相关系数 (Correlation coefficient, CC) 作为性能评价指标.

对于每种方法, 训练 3 个不同的线性回归模型[5]:

1) 岭回归 (Ridge regression, RR), L2 正则化系数 $\lambda = 0.5$. 由于选择的样本数量很少, 本文使用较大的 λ 以减小回归模型的方差.

2) LASSO, L1 正则化系数 $\lambda = 0.5$.

3) 线性支持向量回归 (Support vector regression, SVR), $\epsilon = 0.1 \times \mathrm{std}(y)$ ($\mathrm{std}(y)$ 是 M 个选择样本真实标签的标准差), box constraint $C = 1$. SVR 包含 L2 正则项, 其等效正则化系数为 $\frac{1}{2C}$, 与 RR 和 LASSO 中的大小相同.

在后面的小节中主要给出了 RR 模型上的结果, 因为它的 RMSE 和 CC 通常比 LASSO 和线性 SVR 更稳定, 尤其对于 RS 方法而言. 但是, 如第 3.5 节所示, 当使用 LASSO 或线性 SVR 时, IRD 相对于其他算法 (尤其是 RS) 的提升效果可能更大.

3.4 RR 上的结果

图 3 中展示了使用 RR 作为回归模型, 在 12 个数据集上 5 种采样方法的平均 RMSE 和 CC[6].

通常, 随着 M 的增加, 5 种采样方法的 RMSE 和 CC 也会随之得到改善, 因为有更多的训练样本加入回归训练, 逐渐提升了回归性能. 但仍然可能会存在一些波动, 尤其是在样本数量较少的情况下. 因为仅从少量标注样本中训练得到的线性回归模型可能存在很多随机性和不确定性.

在大多数数据集和大多数 M 取值上, RS 和 GSx 具有更大的 RMSE 和更小的 CC, 即它们的性能相对于另外 3 种算法较差. IRD 在大多数数据集和大多数 M 取值上都取得了最小的 RMSE 和最大的 CC, 表明 IRD 是表现最佳的样本选择方法.

[5] 对普通最小二乘法 (Ordinary least squares, OLS) 回归也进行了尝试, IRD 依然取得了最佳表现. 但当训练样本较少时, OLS 非常不稳定, 因此在实际中不是一个合理的选择, 本文不讨论其结果.

[6] 由于页面限制, 只展示了 RR 的详细结果, 因为它通常表现更稳定. 其他两个回归模型上的结果类似.

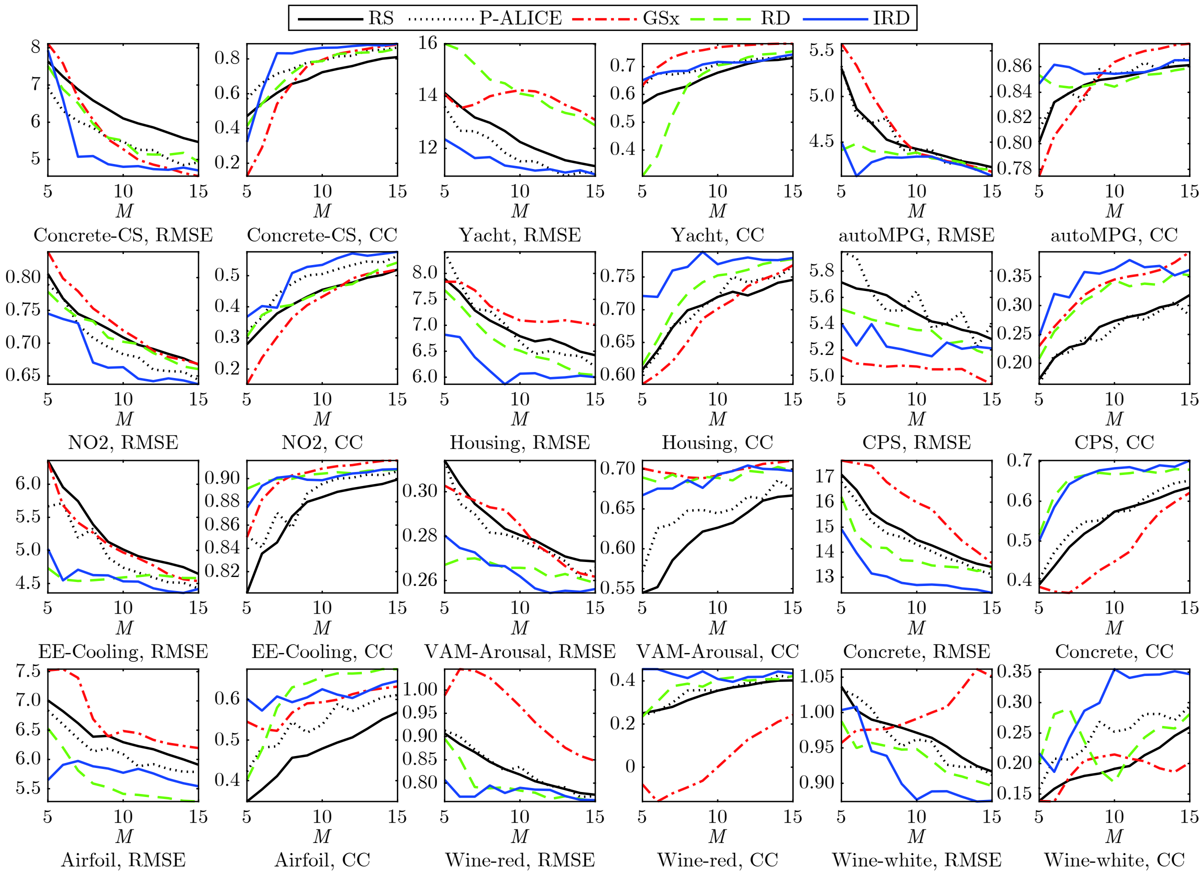


图 3 12 个数据集上的平均 RMSE 和 CC (mRMSE 和 mCC; 重复运行 100 次) (回归模型为 RR ($\lambda=0.5$))

Fig.3 Mean of the RMSEs and the CCs on the 12 datasets, averaged over 100 runs (RR ($\lambda=0.5$) was used as the regression model)

为了更全面地进行比较, 我们还计算了 100 次重复实验 RMSE 和 CC 平均值的曲线下面积 (Area under curve, AUC), 分别记为 AUC-mRMSE 和 AUC-mCC, 结果如图 4(a) 所示. 由于不同数据集上 AUC 的大小差异很大, 不便在一张图中展示, 因此根据 RS 的结果进行了归一化处理, 使图 4(a) 中 RS 的结果始终为 1. 图 4(a) 表明:

1) IRD 在 12 个数据集中的 10 个上均获得了最小的 RMSE, 在其余两个数据集中排名第 2. 平均而言, IRD 取得了最小的 RMSE. 它在 10 个数据集上也取得了最大的 CC, 在其余 2 个数据集上排名第 2 和第 3. 平均而言, IRD 也取得了最大的 CC.

2) 平均而言, RD 的性能略优于 P-ALICE, 两者均优于 RS.

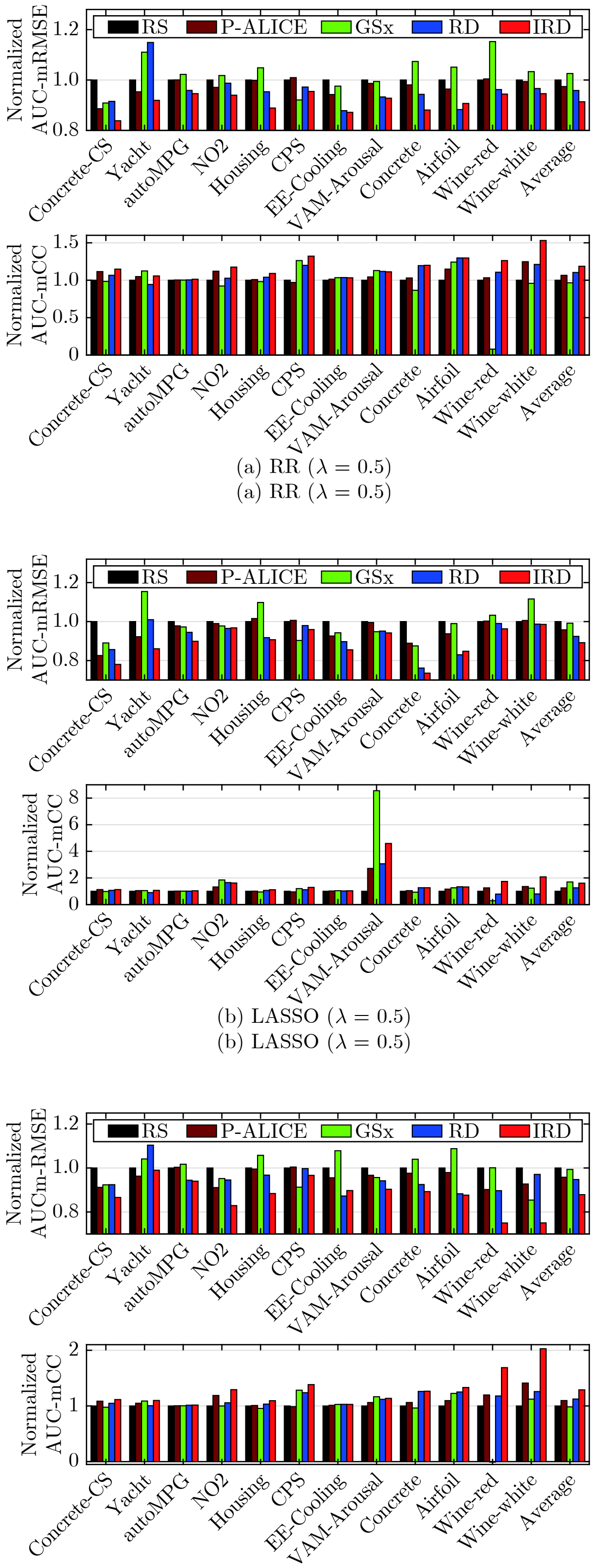
3) GSx 在 7 个数据集上的 RMSE 表现最差, 在另外 3 个数据集上排名倒数第二, 平均而言, GSx 的 RMSE 最差. 它在 6 个数据集中的 CC 也是最低, 因此其 CC 平均值也最低.

因此, 5 种算法的性能整体排名是: IRD > RD > P-ALICE > RS > GSx.

表 3 中展示了 3 个回归模型、5 种无监督采样方法在 12 个数据集上的平均 AUC 情况. 当 M 较小时, GSx 表现较差的原因可能是其选择的样本大多是离群点, 而离群点的负面影响超过了 GSx 多样性的正面影响. IRD 同时考虑了信息性、代表性和多样性, 因此表现最好.

除了准确性, 算法的稳定性也很重要. 实际情况中, 如果多种算法具有相似的性能, 通常首选变化较小, 也就是更稳定的算法. 表 3 展示了运行 100 次的 AUC-mRMSE 和 AUC-mCC 在 12 个数据集上的平均标准差 (Standard deviation, std) 提升结果. 可以看到, IRD 在标准差上相对于 RS 的提升最大, 即它是最稳定的 ALR 方法.

对于不同的 M , 我们统计了 P-ALICE、GSx、RD 和 IRD 对应的 RMSE (CC) 相对于 RS 的比率, 重复 100 次实验在 12 个数据集上取平均, 结果如图 5 所示. 可见, 当 M 较小时, IRD 相对于其他 4 种方法的提升很大, 因为 IRD 同时考虑了信息性、代表性和多样性. 随着 M 的增加, IRD 的优越性逐渐下降, 因为随着标注样本数量的增加, 每个样本



(a) RR ($\lambda = 0.5$)
(a) RR ($\lambda = 0.5$)

(b) LASSO ($\lambda = 0.5$)
(b) LASSO ($\lambda = 0.5$)

(c) 线性 SVR ($C = 1$)
(c) Linear SVR ($C = 1$)

图 4 12 个数据集上归一化 AUC-mRMSE 和 AUC-mCC
Fig. 4 Normalized AUCs of the mean RMSEs and the mean CCs on the 12 datasets

表 3 AUC-mRMSE/sRMSE 和 AUC-mCC/sCC 的提升百分比
Table 3 Percentage improvements of the AUCs of the mean/std RMSEs and the mean/std CCs

回归模型	性能指标		相对于 RS 的提升百分比			
			P-ALICE	GSx	RD	IRD
RR	RMSE	Mean	2.58	−2.57	4.15	**8.63**
		std	2.75	3.98	**36.60**	34.84
	CC	Mean	6.54	−3.43	10.39	**18.70**
		std	12.74	29.47	35.03	**42.97**
LASSO	RMSE	Mean	4.22	0.84	7.58	**10.81**
		std	6.77	0.85	**43.45**	39.84
	CC	Mean	25.06	**69.41**	25.67	60.63
		std	6.39	**31.05**	22.46	29.82
SVR	RMSE	Mean	4.21	0.66	5.23	**12.12**
		std	6.62	−0.19	33.99	**38.69**
	CC	Mean	9.71	−1.65	12.46	**28.99**
		std	11.10	25.78	34.97	**43.25**

最优性的影响就会减小.

3.5 LASSO 和线性 SVR 上的结果

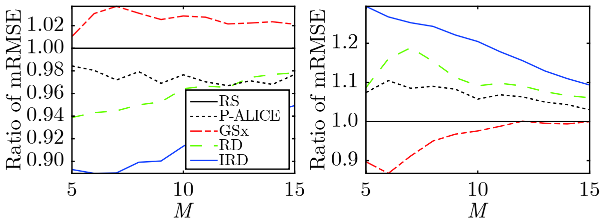
当使用 LASSO 和线性 SVR 作为线性回归模型时, 我们也重复了上述实验. 结果如图 4(b) 和图 4(c) 所示. 可以得到和图 4(a) 类似的结论, 例如 IRD 始终取得最佳的平均性能, 而 RD 则优于 P-ALICE、RS 和 GSx. 此外, 整体看来, 相对于 RR, 4 种 ALR 算法 (特别是 IRD) 在这两个模型上相对于 RS 的性能提升更为明显.

为了量化 4 种无监督 ALR 算法相对于 RS 的改善效果, 我们也计算了其 AUC-mRMSE 和 AUC-mCC 的提升百分比, 如表 3 所示. 无论使用哪种线性回归模型或性能指标, IRD 的平均表现都优于其他 4 种方法.

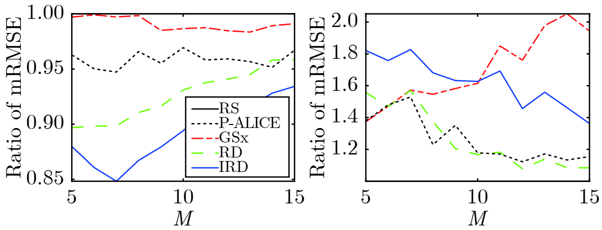
3.6 统计分析

为了确定 IRD 与其他 4 种算法之间的性能差异是否具有统计意义, 我们使用 Dunn 检验[28] 对几种方法的 AUC-mRMSE 和 AUC-mCC 在 12 个数据集上的平均值进行了非参数多重比较检验, 使用错误发现率 (False discovery rate) 方法[29] 进行 p 值校正. 结果如表 4 所示, 其中具有统计意义的结果以粗体标出.

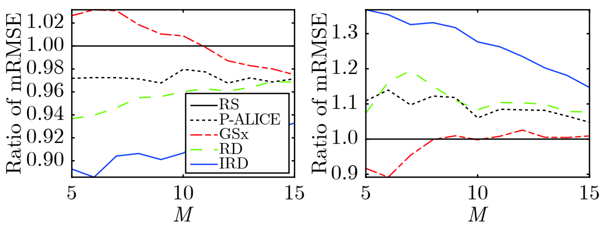
结果表明, 无论使用哪种线性回归模型, IRD 的 RMSE 和 CC 相对于 RS、P-ALICE 和 GSx 的提升始终具有统计学意义; 相对于 RD, CC 的提升具有统计学意义; 使用线性 SVR 时, RMSE 的提升也具有统计学意义.



(a) RR ($\lambda = 0.5$)
(a) RR ($\lambda = 0.5$)



(b) LASSO ($\lambda = 0.5$)
(b) LASSO ($\lambda = 0.5$)



(c) 线性 SVR ($C = 1$)
(c) Linear SVR ($C = 1$)

图 5 对于不同的 M, 4 种 ALR 方法的 mRMSE 和 mCC 相对于 RS 在 12 个数据集上的平均比率

Fig. 5 Ratios of the mean RMSEs and the mean CCs for different M, averaged across 12 datasets

表 4 非参数多重检验的 p 值 ($\alpha = 0.05$; 如果 $p < \alpha/2$ 拒绝 H_0).

Table 4 p-values of non-parametric multiple comparisons ($\alpha = 0.05$; reject H_0 if $p < \alpha/2$)

回归模型	性能指标	IRD versus			
		RS	P-ALICE	GSx	RD
RR	RMSE	**0.0000**	**0.0003**	**0.0000**	0.0284
	CC	**0.0000**	**0.0000**	**0.0000**	**0.0005**
LASSO	RMSE	**0.0000**	**0.0004**	**0.0000**	0.0596
	CC	**0.0000**	**0.0000**	**0.0000**	**0.0000**
SVR	RMSE	**0.0000**	**0.0000**	**0.0000**	**0.0018**
	CC	**0.0000**	**0.0000**	**0.0000**	**0.0000**

3.7 选择样本的可视化

为了更直观地了解不同 ALR 算法选择样本之间的差异, 我们在一个典型数据集 (Housing) 上使用 t-SNE[30] 将样本映射到 2 维空间. 图 6 展示了 3 个不同的 M 值对应的 4 种 ALR 算法选择的样本. P-ALICE 的样本权重在绘图中没有显示.

图 6 中, GSx 倾向于选择位于边界的样本, 这样的样本很有可能是离群点, 且所选样本的分布情况与池中的样本不一致. 因此, 它的平均性能在 4 种算法中是最差的. 与 GSx 相比, P-ALICE 和 RD 选择的样本在池中分布更均匀. IRD 选择的样本倾向于靠近池的边界, 但不完全位于边界, 这样的样本不太可能是异常点, 并且选择的样本的分布情况与池中样本更一致. 这些都可能是 IRD 表现较好的原因.

3.8 $c_{\max}$ 对 IRD 的影响

算法 1~3 中有一个重要参数: $c_{\max}$, 即最大迭代次数. 当 $c_{\max} = 0$ 时, IRD 等效于 RD. 本小节通过设置 $c_{\max} > 0$ 来探究 IRD 的性能是否优于 RD.

图 7 展示了在 3 种线性回归模型上, $c_{\max} \in [0, 10]$ 的归一化 AUC (相对于 RS) 的变化趋势, 这是在 12 个数据集上重复 100 次 实验的平均结果.

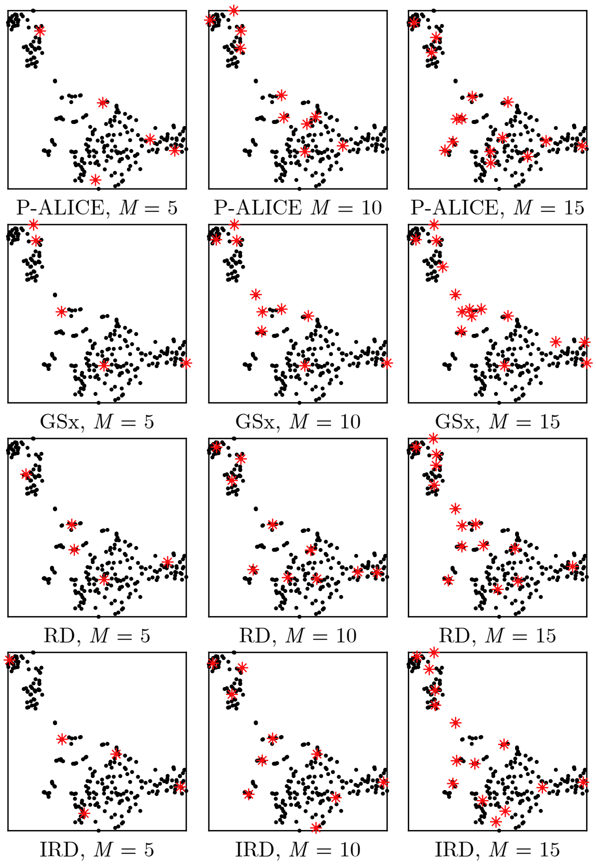
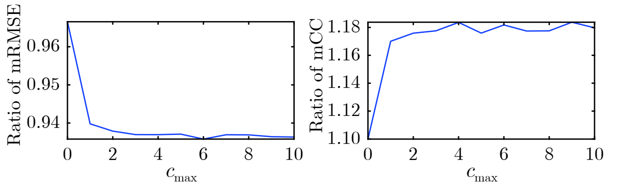


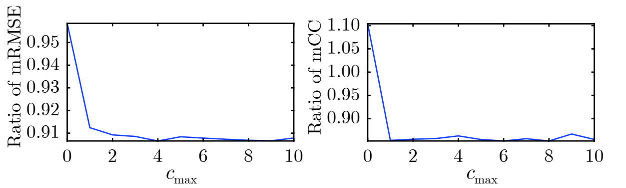
图 6 在 Housing 数据集上不同 ALR 算法所选样本 (星号) 的 t-SNE 可视化

Fig. 6 t-SNE visualization of the selected samples (asterisks) from different ALR approaches on Housing dataset

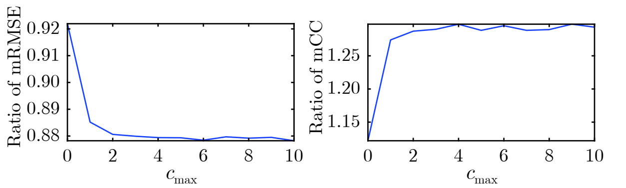
如图所示, IRD 的性能随着 $c_{\max}$ 的增加而迅速提升, 并且总是在 $c_{\max}=5$ 之前就达到了最优, 这意味着 IRD 是一种既有效又高效的算法.



(a) RR ($\lambda=0.5$)
(a) RR ($\lambda=0.5$)



(b) LASSO ($\lambda=0.5$)
(b) LASSO ($\lambda=0.5$)



(c) 线性 SVR ($C=1$)
(c) Linear SVR ($C=1$)

图 7　对于不同的 $c_{\max}$, 4 种 ALR 算法的 AUC-mRMSE 和 AUC-mCC 相对于 RS 在 12 个数据集上的平均比率

Fig. 7　Ratios of AUCs of the mean RMSEs and the mean CCs for different $c_{\max}$, averaged across 12 datasets

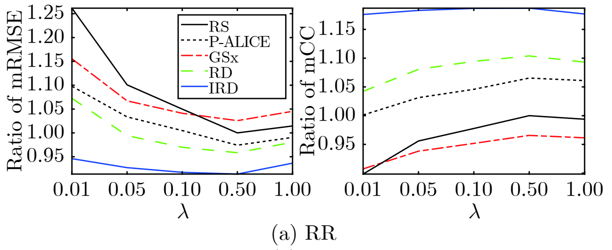
3.9　λ 对 IRD 的影响

为了研究 5 种无监督采样方法的性能对 3 个线性回归模型正则化系数的敏感性, 我们对 $\lambda\in\{0.01,0.05,0.1,0.5,1\}$ 进行了重复实验. 线性 SVR 有一个等价的 L2 正则化系数 $\lambda=\frac{1}{2C}$, 等效设置为 $C\in\{50,10,5,1,0.5\}$. 将每种采样方法在不同参数回归模型下的 AUC 结果相对于 RS ($\lambda=0.5$) 进行归一化, 如图 8 所示.

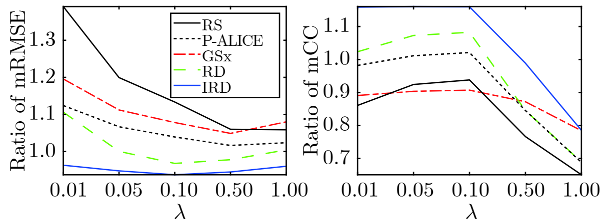
整体来看, 5 种无监督采样方法的性能首先随着 λ 的增大而提高, 然后下降. 然而, 无论 $\lambda(C)$ 取值为多少, IRD 的表现通常都是最好的, RD 次优. 当 λ 较小时, IRD 相对于其他 4 种方法的提升更大. 此外, 可以看出 IRD 对参数 λ 不是很敏感, 这将有利于实际应用.

3.10　三个标准的影响

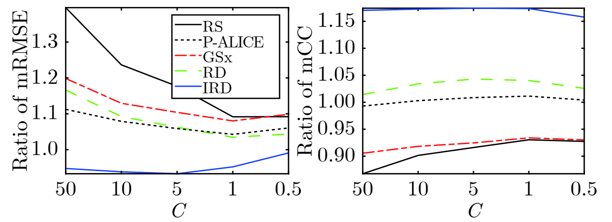
为了研究信息性、代表性和多样性分别对 IRD 的影响, 我们将 IRD 与三个变体进行比较:



(a) RR
(a) RR



(b) LASSO
(b) LASSO



(c) 线性 SVR
(c) Linear SVR

图 8　对于不同的 λ (RR 和 LASSO) 和 C (线性 SVR), 4 种 ALR 算法的 AUC-mRMSE 和 AUC-mCC 相对于 RS 在 12 个数据集上的平均比率

Fig. 8　Ratios of the AUCs of the mean RMSEs and the mean CCs, averaged across 12 datasets, for different λ (RR and LASSO) and C (linear SVR)

1) IRD ($c_{\max}=5$): 本文提出的方法, 在第 2 节中已介绍.

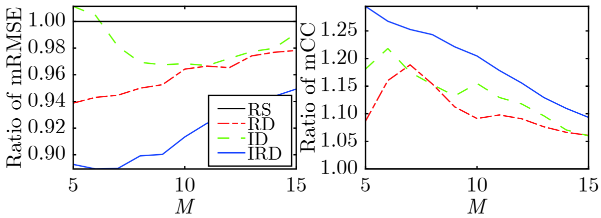
2) ID: 当 $M=d+1$ 时, 只考虑式 (12) 的分母部分; 当 $M>d+1$ 时, 只考虑式 (17) 中的 $D(\boldsymbol{x}_n)$. 即只考虑信息性和代表性.

3) RD: 等同于 $c_{\max}=0$ 时使用 RD 进行初始化的 IRD. 即只考虑代表性和多样性.

对于 $M\in[5,15]$, 每种方法在 12 个数据集上运行 100 次, 训练 3 种线性回归模型: RR ($\lambda=0.5$)、LASSO ($\lambda=0.5$) 和线性 SVR ($C=1$). 图 9 展示了对于不同的 M 取值, IRD 及变体的 RMSE 和 CC 相对于 RS 的平均比率. 3 个回归模型上的结论是类似的. 通常, 3 种 ALR 方法都优于 RS. IRD 仍然表现最好, 这表明同时考虑信息性、代表性和多样性至关重要.

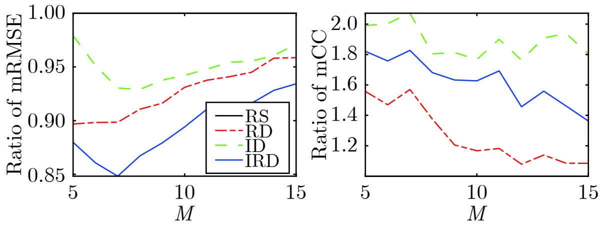
4　结论

主动学习通过选择最有价值的样本进行标注,



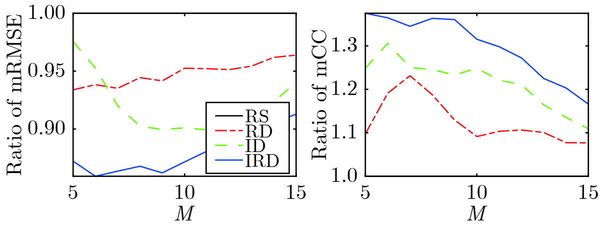
(a) RR ($\lambda = 0.5$)

(a) RR ($\lambda = 0.5$)



(b) LASSO ($\lambda = 0.5$)

(b) LASSO ($\lambda = 0.5$)



(c) 线性 SVR ($C = 1$)

(c) Linear SVR ($C = 1$)

图 9 对于不同的 M, IRD 及其变体的 mRMSE 和 mCC 相对于 RS 在 12 个数据集上的平均比率

Fig. 9 Ratios of the mean RMSEs and the mean CCs w.r.t. different M, averaged across 12 datasets

从而利用较少的训练数据就可以建立较好的机器学习模型. 这在许多实际应用中有着重要的作用, 因为数据的标注过程往往需要耗费大量的时间和经济成本. 大多数现有的主动学习方法是有监督的: 能够从少量的标注样本中建立一个初始的模型, 基于模型查询新的数据, 然后进行迭代更新. 本文考虑了线性回归中完全无监督的基于池的主动学习问题, 即在完全不知道任何标签信息的情况下, 最优地选择初始的少量样本进行标注. 文中提出一种新的主动学习算法 IRD, 该算法同时考虑了主动学习中的 3 个重要标准: 信息性、代表性和多样性. 在来自于不同应用领域的 12 个数据集和 3 种不同的线性回归模型 (RR、LASSO 和线性 SVR) 上进行了大量实验, 充分验证了本文提出方法的有效性.

References

1 Mehrabian A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Cambridge: Oelgeschlager, Gunn & Hain, 1980.

2 Grimm M, Kroschel K, Narayanan S. The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo (ICME). Hannover, Germany: IEEE, 2008. 865−868

3 Bradley M M, Lang P J. The International Affective Digitized Sounds: Affective Ratings of Sounds and Instruction Manual, Technical Report No. B-3, University of Florida, USA, 2007.

4 Joo J I, Wu D R, Mendel J M, Bugacov A. Forecasting the post fracturing response of oil wells in a tight reservoir. In: Proceedings of the 2009 SPE Western Regional Meeting. San Jose, USA: SPE, 2009.

5 Settles B. Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin, USA, 2009.

6 Burbidge R, Rowland J J, King R D. Active learning for regression based on query by committee. In: Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning. Birmingham, UK: Springer, 2007. 209−218

7 Cai W B, Zhang M H, Zhang Y. Batch mode active learning for regression with expected model change. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(7): 1668−1681

8 Cai W B, Zhang Y, Zhou J. Maximizing expected model change for active learning in regression. In: Proceedings of the 13th International Conference on Data Mining. Dallas, USA: IEEE, 2013. 51−60

9 Cohn D A, Ghahramani Z, Jordan M I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996, **4**(1): 129−145

10 Freund Y, Seung H S, Shamir E, Tishby N. Selective sampling using the query by committee algorithm. *Machine Learning*, 1997, **28**(2−3): 133−168

11 MacKay D J C. Information-based objective functions for active data selection. *Neural Computation*, 1992, **4**(4): 590−604

12 Sugiyama M. Active learning in approximately linear regression based on conditional expectation of generalization error. *The Journal of Machine Learning Research*, 2006, **7**: 141−166

13 Sugiyama M, Nakajima S. Pool-based active learning in approximate linear regression. *Machine Learning*, 2009, **75**(3): 249−274

14 Castro R, Willett R, Nowak R. Faster rates in regression via active learning. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2005. 179−186

15 Wu D R, Lawhern V J, Gordon S, Lance B J, Lin C T. Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression. In: Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Budapest, Hungary: IEEE, 2016. 730−736

16 Yu H, Kim S. Passive sampling for regression. In: Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia: IEEE, 2010. 1151−1156

17 Wu D R. Pool-based sequential active learning for regression. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(5): 1348−1359

18 Wu D R, Lin C T, Huang J. Active learning for regression using greedy sampling. *Information Sciences*, 2019, **474**: 90−105

19 Liu Z, Wu D R. Integrating informativeness, representativeness and diversity in pool-based sequential active learning for regression. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE, 2020. 1−7

20 Wu D R, Huang J. Affect estimation in 3D space using multi-task active learning for regression. *IEEE Transactions on Affective Computing*, 2019, DOI: 10.1109/TAFFC.2019.2916040

21 Tong S, Koller D. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2001, **2**: 45−66

22 Lewis D D, Catlett J. Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of the 8th International Conference on Machine Learning (ICML). New Brunswick, Canada: Elsevier, 1994. 148−156

23 Lewis D D, Gale W A. A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland: Springer, 1994. 3−12

24 Grimm M, Kroschel K. *Emotion Estimation in Speech Using a 3D Emotion Space Concept*. Austria: I-Tech, 2007. 281−300

25 Grimm M, Kroschel K, Mower E, Narayanan S. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 2007, **49**(10−11): 787−800

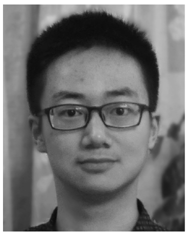
26 Wu D R, Parsons T D, Mower E, Narayanan S. Speech emotion estimation in 3D space. In: Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME). Singapore, Singapore: IEEE, 2010. 737−742

27 Wu D R, Parsons T D, Narayanan S S. Acoustic feature analysis in speech emotion primitives estimation. In: Proceedings of the 2010 InterSpeech. Makuhari, Japan: ISCA, 2010. 785−788

28 Dunn O J. Multiple comparisons among means. *Journal of the American Statistical Association*, 1961, **56**(293): 52−64

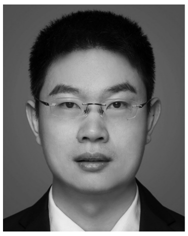
29 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, **57**(1): 289−300

30 Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, **9**: 2579−2605



刘子昂 2017 年获得武汉理工大学自动化专业学士学位, 2020 年获得华中科技大学控制科学与工程硕士学位. 主要研究方向为机器学习.
E-mail: ziangliu@hust.edu.cn
(**LIU Zi-Ang** Received his bachelor degree in automation from Wuhan University of Technology in 2017, and his master degree in control science and engineering from Huazhong University of Science and Technology in 2020. His main research interest is machine learning.)

蒋　雪 华中科技大学人工智能与自动化学院博士研究生. 2019 年获得西南大学电子信息工程学院学士学位. 主要研究方向为机器学习, 脑机接口和情感计算.
E-mail: xuejiang@hust.edu.cn
(**JIANG Xue** Ph.D. candidate at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. She received her bachelor degree in communications engineering from Southwest University in 2019. Her research interest covers machine learning, brain-computer interfaces, and affective computing.)



伍冬睿 华中科技大学人工智能与自动化学院教授. 主要研究方向为机器学习, 脑机接口, 计算智能, 情感计算. 本文通信作者.
E-mail: drwu@hust.edu.cn
(**WU Dong-Rui** Professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interest covers machine learning, brain-computer interfaces, computational intelligence, and affective computing. Corresponding author of this paper.)