分 类 号 _____ 学号 M201772606

学校代码 10487 密级 不保密

华中科技大学

硕士学位论文

回归问题中的主动学习研究

学 位 申 请 人： 刘子昂

学 科 专 业： 控制科学与工程

指 导 教 师： 伍冬睿教授

答 辩 日 期： 2020 年 5 月 6 日

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree for the Master of Engineering

Research on Active Learning in Regression Problems

Candidate : Liu Ziang

Major : Control Science and Engineering

Supervisor : Wu Dongrui

Huazhong University of Science and Technology

Wuhan 430074, P.R. China

May, 2020

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：刘子昂

日期：2020年 5 月 3 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于 保密□，　　在_____年解密后适用本授权书。

不保密☑。

（请在以上方框内打“√”）

学位论文作者签名：刘子昂　　指导教师签名：伍冬睿

日期：2020年 5 月 3 日　　日期： 2020 年 5 月 18 日

摘 要

回归是一类机器学习问题，带标签的训练样本对于回归模型的训练非常重要。而在某些实际应用中，原始样本很容易获得，但是给它们打上真实标签却非常困难，例如需要花费大量的人力、物力或时间。对于这类回归问题，应用主动学习可以有效地降低打标成本。目前的主动学习研究大多针对分类问题，针对回归问题的较少。本文考虑离线的基于池的主动学习回归问题，即给定一个样本池，如何从中选择尽可能少的最有价值的样本来打标，从而训练出性能尽可能好的回归模型。

本文首先对比了有监督主动学习回归算法和无监督主动学习回归算法，并指出后者的一些优势；然后为无监督主动学习回归算法建立数学模型，并提出一种无需真实标签也能预测回归模型的精度的新指标；本文随后将有监督主动学习回归算法中的三个核心指标“分散度”、“代表性”和“信息量”迁移到了无监督主动学习回归算法中，并基于提出的数学模型和新指标为它们提供了理论解释；本文随后提出一种用于优化无监督主动学习回归算法中待打标样本集合的框架，该框架利用交替优化算法将多目标优化问题拆分为多个单目标优化问题；本文随后基于该框架，提出两种新的无监督主动学习回归算法 iRDM 和 IRD，其中 iRDM 算法度量并融合了“分散度”和“代表性”指标，IRD 算法不仅度量并融合了“分散度”和“代表性”指标，还针对线性回归模型度量了“信息量”指标，并融合到单次优化的目标函数中；本文最后在涵盖多个实际应用领域的 12 个公开的回归数据集上进行了大量的实验，实现了现有的经典主动学习回归算法和本文提出的两种新的无监督主动学习回归算法，使用“岭回归（Ridge）”和“基于径向基核函数的支持向量机回归（RBF SVR）”分别测试各算法在线性回归和基于核的非线性回归中的效果，从多个角度进行了数据分析，并进行了统计检验，验证了本文提出的两种新算法的性能和稳定性均优于现有的无监督主动学习回归算法，且在带标签训练样本很少时甚至优于有监督主动学习回归算法，还验证了使用本文提出的新算法为有监督主动学习回归算法选择初始的少量待打标样本能够有效提升有监督主动学习回归算法的性能。

本文为无监督主动学习回归算法建立的数学模型和提出的预测指标能为后续的无监督主动学习回归算法的研究提供理论支持及新思路。本文提出的两种无监督主动学习回归算法合理地度量并融合了分散度、代表性和信息量三个核心指标。相比现有算法，它们能更有效地减少打标工作量，它们还能用于任何有监督主动学习回归算法中以提升初始回归模型的性能。

关键词：主动学习，无监督学习，线性回归，核函数

Abstract

Regression is a type of machine learning problem. Labeled samples are very important for training regression models. However, in some practical applications, the original samples are easy to obtain, but it is very difficult to label them with true labels. For example, the labeling process usually takes a lot of manpower, material resources or time. For such regression problems, applying active learning can effectively reduce the cost of labeling. However, among the existing active learning researches, most of them only focus on classification problems, and few on regression problems. This thesis focuses on the offline pool-based active learning regression problem (ALR), that is, given a sample pool, how to select a few valuable samples from it to label, so that the regression model trained from them can achieve the best possible performance.

This thesis first compares the unsupervised ALR algorithms with the supervised ALR algorithms, and points out some advantages of the unsupervised ALR. Secondly, this thesis establishes a mathematical model for unsupervised ALR algorithms, and proposes a new indicator that can predict the accuracy of the regression model without any true label information. Thirdly, this thesis migrates the three essential criteria: "diversity", "representativeness" and "informativeness" in the supervised ALR to the unsupervised ALR, and theoretically explains them using the proposed mathematical model and the new indicator. Next, this thesis proposes a framework for optimizing the set of candidate samples in unsupervised ALR. The framework uses an alternating optimization approach to split the multi-objective optimization problem into multiple single-objective optimization problem. Based on this framework, two new unsupervised active learning regression algorithms, iRDM and IRD, are proposed. The iRDM algorithm measures and integrates "diversity" and "representativeness". The IRD algorithm not only considers "diversity" and "representativeness", but also measures "informativeness" for the linear regression model, and integrates it into the objective function. At the end of this thesis, a large number of

experiments were performed on 12 public regression datasets. These datasets cover multiple practical application areas. For each dataset, the proposed two unsupervised ALR algorithms (iRDM and IRD), and seven state-of-the-art ALR algorithms were implemented by MATLAB2019 and were used in linear regression (Ridge) and kernel function regression (RBF SVR). The results showed that the two new proposed unsupervised ALR algorithms (iRDM and IRD) performed better and more stable than the existing unsupervised ALR algorithms. Furthermore, they even performed better than supervised ALR algorithms when the number of labeled samples is very small. It was also verified that using these two algorithms to select a few samples to label at the beginning for a supervised ALR algorithm can improve its performance.

The mathematical model and the evaluation indicator proposed in this thesis for unsupervised ALR can provide theoretical support and new ideas for subsequent researches on unsupervised ALR. The two unsupervised active learning regression algorithms proposed in this thesis reasonably measure and integrate the three essential criteria: "diversity", "representativeness", and "informativeness". Compared with the existing ALR algorithms, they can reduce the labeling effort more effectively. They can also be used in any supervised ALR to further improve their performance by generating a better initial regression model.

Key words：Active learning, Unsupervised learning, Linear regression, Kernel function

目 录

1 绪论

1.1 研究背景、目的及意义

回归问题是机器学习中的一类常见问题，在该类问题中，样本的标签是连续值，我们需要根据已知的样本和真实标签来训练回归模型，用于对新样本的标签进行预测。

样本的真实标签对于回归模型的训练十分重要，但在实际应用中存在一类问题，那就是获取大量的原始数据很容易，而给它们打上真实标签却很费力，例如需要消耗大量的人力、物力或时间。

例如在“根据语音或视频进行情感估计的任务”[1]中，情感标签是二维空间（效价、唤醒度）[2]或三维空间（效价、唤醒度和优势度）[3]中的连续值。虽然记录大量的语音或视频较为容易，但是为每个语音或视频样本打上对应的情感标签却较为困难。这是由于情感是主观、微妙且不确定的，往往需要多位专业评估者评估多次，才能获得准确的标签。例如，在 VAM 语料库中每个语音样本的情感标签需要 6-17 位评估者来评估[4]，在数据集“使用生理信号数据集进行情感分析的数据库”中的每个视频片段的情感标签需要 14–16 个评估者来评估[5]，在 IADS-2 数据集中每个语音样本的情感标签则至少需要 110 位评估者来评估[6]。因此获取情感标签的工作量非常大，需要消耗大量人力。

再例如在“通过脑电信号进行疲劳驾驶检测的任务”[7][8][9]中，采集脑电信号相对容易，但是获取它们真实的疲劳程度则需要进行持续注意力驾驶实验[10][11]，其物力成本是相对较高的。

再例如在“油井压裂后 180 天的累计产油量预测任务”[12]中，其输入（油井的压裂参数，例如它的位置、射孔的长度、区域/孔的数量和注入的泥浆/水/砂的体积等）可以在压裂操作期间很容易地记录下来，但是要获得真实的标签，即压裂后 180 天的累计产油量，则必须至少等待 180 天。因此获取标签的时间成本极高。

经典的回归算法是被动学习算法，即对样本池的所有样本都打标，或随机选择部分样本来打标，然后构建带标签的训练集之后再训练回归模型。而对于上述无标签样本获取容易但打标困难的问题，在实际应用中对所有样本都打标的成本是非常高的，往往是没有必要甚至是无法实现的，而随机选择部分样本来打标往往不能训练出性能达标的回归模型。

主动学习[13]是机器学习中减少打标工作量的常用方法，将主动学习应用于回归问题，即“主动学习回归算法”。该类算法能主动地选择最有价值的待打标样本，并只给这些样本打上真实标签，从而提升打标的性价比，从而在训练出性能相当的回归模型的条件下，有效减少需要打标的样本的数量。

现在已经有很多关于主动学习算法的研究，它们都取得了一定的效果[8][13][14][15][16][17][18][19]，但其中大多数主动学习算法都是为分类问题而设计的，为回归问题设计的主动学习算法较少[20]，在这些为数不多的主动学习回归算法中，适用于机器学习中“基于池”的应用场景的算法则更少了[20]。

本课题研究“基于池的主动学习回归算法”（后文中简称为“主动学习回归算法”）。在实际应用中，很多应用场景（例如前文介绍的三个应用场景）亟需基于池的主动学习回归算法，但现有的算法数量少，理论体系尚不完善。综合以上分析，本课题的研究在理论和实践中均具有较高的研究价值。

1.2 主动学习回归算法的国内外研究现状

依据不同的样本查询情景，学者们将“主动学习回归算法”划分为如下两类：“基于总体的主动学习回归算法”和“基于池的主动学习回归算法”[21]。基于总体的主动学习回归算法的应用场景是：样本空间中任意数据点都有效且能够被查询标签；测试数据的分布是已知的；其目标是：找到最优的训练数据分布，然后使用该分布在样本空间中采集少量训练数据，对它们打上真实标签后生成带标签的训练数据集。基于池的主动学习回归算法的应用场景是：给定一个无标签样本池，样本池中的样本可以被打标，样本空间中其他样本点无意义不能被打标；其目标是：从给定的无标签样本池中采集少量最有价值的样本，随后仅对这些样本打标以生成带标签的训练数据集。

在机器学习领域，大多数的数据集都是基于池的，因此本文只考虑“基于池的主动学习回归算法”。“基于池的主动学习回归算法”包含：“顺序主动学习回归算法”和“批模式主动学习回归算法”[22]。顺序主动学习回归算法是指在每次迭代过程中只选择一个样本来打标。该类算法能够快速地更新回归模型，因此性能更好，但是计算代价较高，且需要多次与人工专家交互。批模式主动学习回归算法是指在每次迭代过程中选择一批样本来打标，这样做的目的是减少计算代价，同时减少与人工专家的交互次数，但其性能往往略低于顺序主动学习回归算法。本文只考虑性能较好的“顺序主动学习回归算法”。

本文将现有的“基于池的主动学习回归算法”分为如下几类：“基于回归模型信息的主动学习回归算法”、“基于贪婪采样的主动学习回归算法”和“同时考虑分散度和代表性的主动学习回归算法”。

第一类是“基于回归模型信息的主动学习回归算法”[21][23][24][25]，其基本思想是依据当前回归模型的预测信息来选择最有价值的待打标样本，例如打标之后能对当前回归模型产生最大影响的样本，或当前模型预测的最不准确的样本。该类方法的优点是算法往往能取得较好的性能；该类方法的缺点是每种算法往往只适用于一种或一类回归模型，因为算法的推导过程需要用到某一种或一类回归模型的信息。该类方法的几种代表算法如下：

基于委员会分歧的主动学习算法（Query By Committee，简称 QBC）在分类问题[13][26][27][28][29]和回归问题[13][24][30][31][32][33]中均有广泛应用。本文考虑 RayChaudhuri 和 Hamey[24]提出一种 QBC 算法，该方法同时适用于线性回归和非线性回归，其基本思想是：建立模型委员会，选择委员会预测分歧最大的样本来打标。

基于模型变化的期望最大化（Expected Model Change Maximization，简称 EMCM）的主动学习算法在分类问题[13][34][35][36]、回归问题[22][23]和排序问题[37]中皆有应用。本文考虑 Cai 等学者提出一种 EMCM 算法[23]，该方法适用于线性回归和梯度提升决策树回归，本文只考虑其在线性回归中的部分。

基于泛化误差条件期望的重要性加权的最小二乘算法（Pool-based Active Learning using the Importance-weighted least-squares learning based on Conditional

Expectation of the generalization error, 简称 P-ALICE），由 Sugiyama 和 Nakajima[21]提出，该算法仅适用于线性回归，其基本思想是，通过同时考虑训练集和测试集的协变量偏移，从样本池中采集M个样本，并确定它们的权重，然后训练一个加权线性回归模型。

基于残差的主动学习回归方法（Active Learning using Residual regression 简称 RSAL），由 Douak 等人[25]提出，该算法只适用于基于核的非线性回归。其基本思想是：预测当前回归模型对每一个无标签样本的“预测误差”，然后选择“预测误差”最大的样本来打标。

第二类是“基于贪婪采样的主动学习回归算法”[38][39]，其基本思想是利用贪婪采样的方法，在每一步中只选择最有利于当前已打标样本集合的样本。该类方法的优点是操作简单，运算代价小；该类方法的缺点是只能保证当前选择的样本对之前选择的样本最有利，但无法保证其对之后选择的样本也最有利，因此缺乏全局优化的观念，且贪婪采样容易采集到离群点，从而降低算法性能和鲁棒性。该类方法的几种代表算法如下：

基于贪婪采样的主动学习回归算法（Greedy Sampling，简称 GS），由 Yu 和 Kim[38]提出，它适用于线性回归和非线性回归。Wu[39]改进了 GS 算法中第一个样本的选择方法，本文考虑该算法（GSx），其基本思想是让选择的样本在样本空间中尽可能分散。GSx 算法首先计算样本池的中心，然后找到距离样本池中心最近的样本作为第一个待打标样本；然后在每一轮迭代中计算每个未被选中的样本$\mathbf{x}_n$到已确定的待打标样本的距离的最小值$d_n^{\mathbf{x}}$，最后选择$d_n^{\mathbf{x}}$最大的样本为新的待打标样本。该算法既可以在选择一个新的待打标样本后立即交给人工专家打标，也可以等选择完所有待打标样本后，一次性交给人工专家打标。

改进的基于贪婪采样的有监督主动学习回归算法（improved Greedy Sampling，简称 iGS），由 Wu[39]提出，它仅适用于线性回归。其基本思想是同时考虑特征空间中的贪婪距离和标签维度上的贪婪距离。在每一轮迭代中，对每一个无标签样本$\mathbf{x}_n$，计算其到已打标样本的距离的最小值$d_n^{\mathbf{x}}$，再依据当前的回归模型计算其预测标签，

然后计算该预测标签到已打标样本的真实标签的差距的最小值d_n^y，最后选择$d_n^x \times d_n^y$最大的样本来打标。

第三类是“同时考虑分散度和代表性的主动学习回归算法”[20]，其基本思想是在保证选择的待打标样本之间尽可能分散的前提下，又使得每一个样本周围的无标签样本尽可能多。该类方法的优点是算法选择的待打标样本集能更好地代表整个样本池，且每一个样本都能更好地代表其周围的样本，从而提升打标的性价比，提升算法的鲁棒性；该类方法的缺点是属于启发式算法，缺乏理论基础。该类方法的几种代表算法如下：

同时考虑待打标样本的分散度和代表性的主动学习回归算法（Representativeness and Diversity，简称 RD），由 Wu[20]提出，它适用于线性回归和非线性回归，其基本思想是让算法选中的待打标样本分散的同时具有代表性。RD 首先对样本池使用k-均值聚类算法（$k = M$）聚类成M个簇，然后选择各簇中距离簇中心最近的样本作为待打标样本，依此法获得的M个待打标样本，即为算法选择的待打标样本。

Wu[20]还将 RD 算法与其他主动学习回归算法结合，例如将 RD 与 QBC 算法结合形成 RD-QBC 算法，将 RD 与 EMCM 结合形成 RD-EMCM 算法，其基本思想是再同时考虑分散度和代表性的基础上，再考虑回归模型的信息。该文实验部分验证了 RD-EMCM 算法的效果优于 RD-QBC 算法，因此本文使用 RD-EMCM 算法。RD-EMCM 算法首先使用 RD 算法从无标签样本池中选择M_0个样本来打标，然后开始迭代，首先使用k-均值聚类算法（$k = M_0 + 1$）将整个样本池聚类为$M_0 + 1$个簇，其中至少有一个簇不含有已打标样本，选择其中最大的一个不含已打标样本的簇$\bar{C}$，使用 EMCM 算法在簇$\bar{C}$中选择最有价值的样本来打标，打标后加入已打标样本集，将M_0加一，开始下一轮迭代，直到结束。

此外，“基于池的主动学习回归算法”还可以分为“有监督主动学习回归算法”和“无监督主动学习回归算法”两种[40]。有监督主动学习回归算法需要迭代式地依次选择待打标样本来打标[22]，在选择最有价值的无标签样本的过程中需要使用到样本池中部分已知的真实标签。无监督主动学习回归算法则可以一次性选择所有待打标样本，在选择无标签样本的过程中不需要用到任何真实标签信息，且可以只与人工

专家交互一次。上面介绍的八种现有算法中，QBC、EMCM、RSAL、RD-EMCM 和 iGS 五种算法是有监督主动学习回归算法，P-ALICE、GSx 和 RD 三种算法是无监督主动学习回归算法。

1.3 研究内容及主要贡献

本文将深入研究回归问题中的主动学习，意图让主动学习更好地为回归问题服务，更有效地减少训练回归模型所需要的打标代价。

本文的主要贡献如下：

为无监督主动学习回归算法建立了数学模型，并提出一种新的指标来预测回归模型的性能，从而判断当前选中的待打标样本集合的优劣。该预测指标可为新的无监督主动学习回归算法的设计提供思路。

将分散度、代表性和信息量三个核心指标从有监督主动学习回归算法迁移到无监督主动学习回归算法，并基于本文建立的数学模型和提出的预测指标，在线性回归和基于核函数的非线性回归中，对这三个核心指标的作用进行理论解释。

提出一种适用于无监督主动学习回归算法的交替优化框架，用于优化待打标样本集合，它将多目标优化问题拆分为多个单目标优化问题迭代进行。

提出一种新的无监督主动学习回归算法 iRDM，该算法的思想是同时最大化分散度和代表性，它同时适用于线性回归和基于核的非线性回归。本文还为该算法在两种回归模型中的有效性提供了理论解释。

提出一种新的无监督主动学习回归算法 IRD，该算法同时融合了分散度、代表性和信息量三个核心指标，其中信息量指标是针对线性回归模型设计的。本文还针对线性回归，在一维、二维以及高维特征空间中为 IRD 算法提供了理论解释。

在涵盖多个实际应用领域的 12 个公开的回归数据集上进行了大量的实验，在线性回归和基于核的非线性回归中，验证了本文提出的两种新算法（iRDM 和 IRD）相对于现有算法的优势，以及 IRD 算法在线性回归中相对 iRDM 算法的优势。依据多次重复实验的结果进行了统计检验，验证了各差异的显著性，让结论更可靠。实验还

验证了两种新算法对回归模型超参数变化和样本池分布变化的鲁棒性，以及分散度、代表性和信息量三个核心指标各自的有效性。

本文各章节的主要内容如下：

第一章介绍主动学习回归算法的背景、意义、现状和本文的主要创新点。

第二章介绍现有的基于池的主动学习回归算法，简述各算法的原理和思想并给出各算法的步骤。这些算法都将在本文的实验部分得以实现，作为本文提出的两种新算法的对照算法。

第三章从理论角度研究无监督主动学习回归算法。对无监督主动学习回归算法进行了数学建模；还提出了一种在无监督主动学习回归算法中用于预测回归模型精度的新指标；还将分散度、代表性和信息量三个核心指标从有监督主动学习回归算法推广到无监督主动学习回归算法，并基于建立的数学模型和提出的预测指标，在线性回归和基于核的非线性回归中为三个核心指标提供理论解释。

第四章首先提出一种交替优化框架，用于在无监督主动学习回归算法中优化待打标样本集；然后基于该框架提出两种新的无监督主动学习回归算法 iRDM 和 IRD，其中 iRDM 算法同时考虑了分散度和代表性，IRD 算法在同时考虑了分散度和代表性的基础上，还针对线性回归模型给出分散度-信息量指标的联合度量。本章还给出两种算法的理论解释及具体流程。

第五章展示了大量的实验结果，验证了本文提出的新算法 iRDM 和 IRD 的有效性和稳定性并给出分析；还验证了无监督主动学习回归算法中分散度、代表性和信息量三个核心指标各自的作用；还验证了在回归模型超参数变化或样本池分布变化时，新算法的表现依然优于现有算法；还验证了使用无监督主动学习回归算法代替随机采样为有监督主动学习回归算法选择初始的少量待打标样本，能有效提升该有监督主动学习回归算法的性能。

第六章对全文进行总结和归纳，并指出未来的研究方向。

2 无监督主动学习回归算法的理论研究

“主动学习回归算法”可以分为“无监督主动学习回归算法”和“有监督主动学习回归算法”两类[40]。本章首先详细分析无监督主动学习回归算法的特性及其相对于有监督主动学习回归算法的三个核心优势；本章随后为无监督主动学习回归算法建立数学模型，并提出一种用于预测回归模型性能的指标；本章随后将有监督主动学习回归算法中的分散度、代表性和信息量三个核心指标迁移到无监督主动学习回归算法中，并基于提出的数学模型和预测指标，在线性回归和基于核的非线性回归中为它们提供理论解释。

2.1 无监督主动学习回归算法的特性及优势

主动学习回归算法在从样本池中采集最有价值的样本的时候，依据是否需要真实标签信息（部分已打标的样本），可以将主动学习回归算法分为“有监督”与“无监督”两大类。有监督主动学习回归算法的采样过程需要真实标签信息，而无监督主动学习回归算法的采样过程不需要任何真实标签信息。

研究发现，大多数现有主动学习回归算法为“有监督”的，然而，无监督主动学习回归算法同样具有重要的研究意义，具体如下：

第一是替代随机采样，帮助有监督主动学习回归算法选择更好的初始待打标样本。在没有任何真实标签的情况下，任何有监督主动学习回归算法都必须需要一种无监督的采样算法来选择最初的几个样本以进行标签并构建初始回归模型，最简单的方法便是利用随机采样。因此，使用更好的无监督主动学习回归算法可以简单地提高有监督主动学习回归算法的总体性能，因为可以获得更好的初始模型。

第二是减少人类专家的在线时间或实验时间。有监督主动学习回归算法需要与人类专家多次互动，这是因为在每次交互中，有监督主动学习回归算法都会向专家提供一个或一批候选样本，然后专家会在经过专业分析后返回其真实标签（或经过实验后返回实验结果）。这要求专家（或实验设备）一直保持在线状态，这非常消耗时间和成本，有时甚至很难实现。但是，无监督主动学习回归算法可以一次性选择出所有

的待打标的候选样本，人类专家可以一次完成标记工作（或多组实验可以并行进行，同时得出多个实验结果），这样可以节省人类专家的时间（或节省实验时间）。

第三是在训练样本数量较少的条件下提升回归模型性能。在特殊情况下，我们的资源仅足以标记极少量样本（例如 5 个样本），若使用有监督主动学习回归算法，用于训练初始回归模型的标记样本会更少（例如 3 个标记样本），这样初始模型会非常粗糙，可能会误导另外 2 个待打标样本的选择，降低最终训练得到的回归模型的性能。而若使用无监督主动学习回归算法，则不会存在这个问题，因此在训练样本数量极少的情况下，在回归问题中应用无监督主动学习回归算法可能比应用有监督主动学习回归算法获得更好的效果。

“有监督主动学习回归算法”和“无监督主动学习回归算法”的流程图见图 2-1。

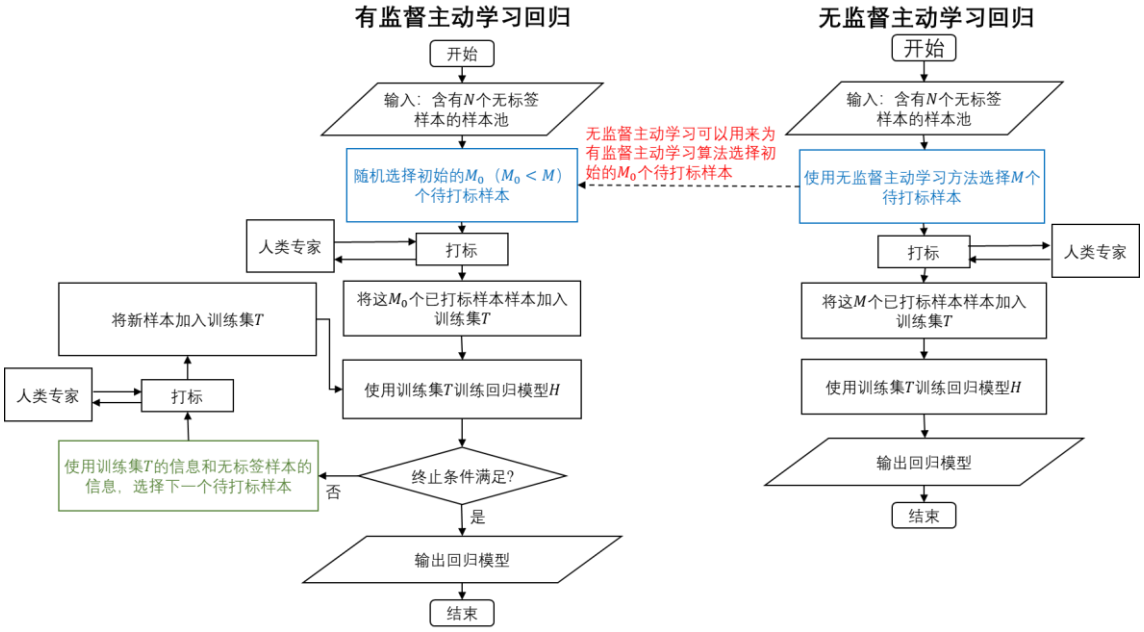

图 2-1 有监督主动学习回归算法和无监督主动学习回归算法流程图

2.2 无监督主动学习回归算法的数学建模

本节将首先定义需要用到的数学变量，并给它们分配指定的符号。

设给定的样本池中含有N个样本，样本的特征维度是d，那么特征空间就是一个d维的线性空间$\mathbb{R}^d$，且这N个样本分别对应$\mathbb{R}^d$中的N个特征点$\{\mathbf{x}_n\}_{n=1}^N$。设这N个特征

点的真实标签是$\{y_n\}_{n=1}^N$，它们是未知的，在主动学习采样过程中需要预测，设这N个样本点的预测标签是$\{\hat{y}_n\}_{n=1}^N$。给每个样本的特征向量增加代表标签的维度，则生成一个$d+1$维的线性空间$\mathbb{R}^{d+1}$，本文中叫它样本空间。样本池中的N个样本加上预测标签$\{\hat{y}_n\}_{n=1}^N$后，在样本空间$\mathbb{R}^{d+1}$中对应N个预测样本点$\{\hat{\mathbf{x}}_n\}_{n=1}^N$。任意样本$\mathbf{x}_i$的预测标签$\hat{y}_i$可以看作是一个随机变量，那么样本$\mathbf{x}_i$在样本空间$\mathbb{R}^{d+1}$中对应的预测样本点$\hat{\mathbf{x}}_i$也是一个随机变量。

线性回归模型在样本空间$\mathbb{R}^{d+1}$中对应一个超平面。设使用样本池中所有的样本训练出的最优回归超平面是H^*（存在但未知），设样本池中的N个样本在最优回归超平面H^*上对应的最优预测标签值是$\{y_n^*\}_{n=1}^N$，设它们在最优回归超平面H^*上对应的最优预测样本点是$\{\mathbf{x}_n^*\}_{n=1}^N$。样本池中的$N$个样本对应的$N$个预测样本点$\{\hat{\mathbf{x}}_n\}_{n=1}^N$也能通过线性回归获取一个预测超平面$\widehat{H}$。预测样本点$\{\hat{\mathbf{x}}_n\}_{n=1}^N$和预测回归超平面$\widehat{H}$都可以看作围绕最优回归超平面$H^*$震荡的随机变量。

图 2-2 直观展示了特征空间为一维时，无监督主动学习回归算法中预测回归模型$\widehat{H}$的不确定性：

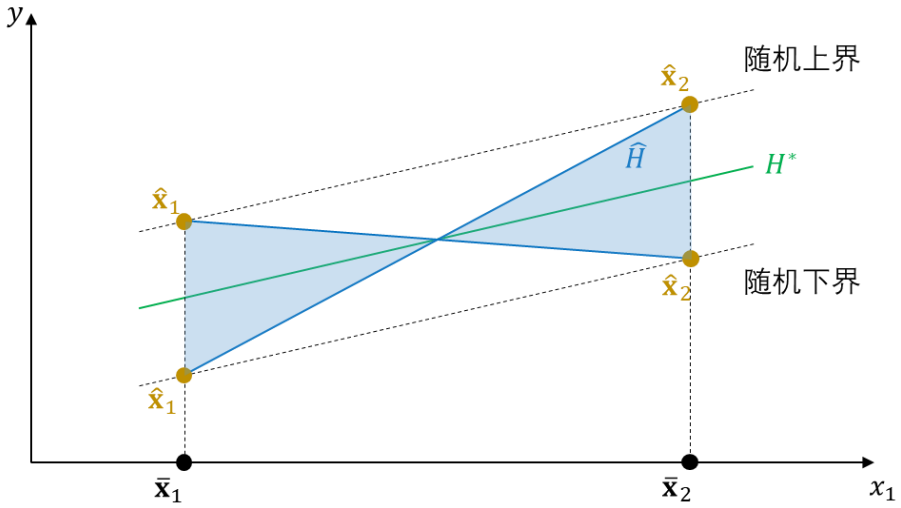

图 2-2 特征空间为一维时，预测回归模型的不确定性的示意图

图 2-2 中，设待打标样本点为$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2\}$，预测样本点$\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2\}$和预测回归直线都是围绕最优回归直线$H^*$上下震荡的随机变量，图中展示了其上界和下界。可以用图中两个蓝色三角形的对顶角大小来度量回归模型的不确定性。

2.3 评价回归模型性能的新指标

设目前的资源只够打M个标签，那么无监督主动学习回归算法的目标是一次性地从样本池的N个样本中采集M个最优的待打标样本，输出这个最优待打标样本集合$\{\mathbf{x}_m\}_{m=1}^{M}$给人工专家进行打标。而采样的方式非常多，有$C_M^N$种，想要找到最优的待打标样本集合，首先需要找到一种指标来比较不同待打标样本集合的优劣，即预测为不同待打标样本集合打标后，训练得到的回归模型的精度。遗憾的是，在无监督主动学习回归算法的场景下，没有任何已知的标签信息，因此经典的均方误差、相关系数等指标均无法使用，此时急需一种新的指标。

本文提出：使用预测回归超平面$\hat{H}$的正法向量$\hat{\boldsymbol{\gamma}}$与最优回归超平面$H^*$的正法向量$\boldsymbol{\gamma}^*$之间的夹角大小$|\theta|$来评价待打标样本集合的优劣。夹角$|\theta|$越小，代表当前待打标样本集合越优（备注：这里区分超平面的正面和背面，因此法向量分为正法向量和负法向量，在标签轴方向上的分量为正的法向量为正法向量）。下面给出解释：

首先，新指标可以在几何学上直观理解。每一个待打标样本集都对应一个预测回归超平面$\hat{H}$，若预测回归超平面$\hat{H}$越接近最优预测超平面H^*，便说明预测超平面$\hat{H}$的精度越高，对应的待打标样本集越好。根据高维几何知识，每个超平面的正法向量方向唯一，因此超平面之间的差距可以通过正法向量间的夹角大小来衡量，夹角$|\theta|$越小，代表预测回归超平面$\hat{H}$和最优回归超平面H^*越接近。值得注意的是，我们无需知道夹角$|\theta|$的具体大小，只需要比较不同预测回归超平面$\hat{H}$对应的夹角$|\theta|$的大小即可，高维几何学中计算向量夹角相对困难，但比较向量夹角大小相对容易，这为使用该指标提供了方便。

此外，和均方误差、相关系数等经典的回归模型评价指标类似，新指标还具有物理意义，即表征预测模型与最优模型的特征系数向量的差异。原因如下，线性回归模型为：

$$\hat{y} = \mathbf{x}^T\boldsymbol{\alpha} + b$$

其中$\mathbf{x}$是样本的特征列向量，每个元素代表一个特征。$\boldsymbol{\alpha}$是系数向量，每个元素代表一个特征的系数，b是偏置项常数，$\hat{y}$是模型对样本$\mathbf{x}$的预测标签值。该模型在样本空间$\mathbb{R}^{d+1}$中对应一个超平面，它可以变形为：

$$\hat{y} - \mathbf{x}^T\boldsymbol{\alpha} - b = 0$$

等价于：

$$[1, -\boldsymbol{\alpha}^T]\begin{bmatrix}\hat{y}\\ \mathbf{x}\end{bmatrix} - b = 0$$

令$\boldsymbol{\gamma}^T = [1, -\boldsymbol{\alpha}^T]$，$\boldsymbol{\gamma}$即为超平面的正法向量（其标签轴$\hat{y}$上的分量大于零）。

可以看到超平面的正法向量$\boldsymbol{\gamma}$完全由各特征对应的系数决定。对两个线性回归模型而言，各特征的系数差距越大，则它们的正法向量的夹角往往越大。那么预测回归超平面$\hat{H}$的正法向量$\hat{\boldsymbol{\gamma}}$与最优回归超平面$H^*$的正法向量$\boldsymbol{\gamma}^*$之间的差距可以衡量预测模型与最优模型的系数向量之间的差距，这便是该指标的物理意义。

该指标的缺陷是没有考虑偏置项b的差异，但不影响其在无监督主动学习回归算法中的使用。其原因如下：若两个平面的正法向量$\boldsymbol{\gamma}$差距很小且偏置b差距很大，则它们近似于平行难以相交。而预测回归超平面和最优回归超平面的训练集中都包含选中的M个待打标样本，这M个样本一般均匀分布在样本池内，因此预测回归超平面$\hat{H}$和最优回归超平面H^*往往会在样本池所在的区域相交，因此一般不会出现正法向量$\boldsymbol{\gamma}$差距很小且偏置b差距很大的情况，因此使用正法向量$\boldsymbol{\gamma}$的差异即可度量两个模型的主要差异。

2.4 分散度、代表性和信息量三个核心指标及其理论解释

2.4.1 分散度、代表性和信息量三个核心指标

Wu 在文献[20]中提出主动学习回归算法中的三个核心指标：分散度、代表性和信息量，它们的具体含义及解释如下：

分散度：主动学习回归算法选择的待打标样本在样本空间中应该尽可能分散，分散度越大越好。该指标的意义是：让被选择的这部分样本更好地代表整个样本空间；

而且可以避免两个或多个待打标样本因太接近而导致所含的信息也相似，从造成打标资源浪费的情况。度量方式举例：待打标样本之间的距离，距离越大则分散度越大。

代表性：主动学习回归算法选择的每一个待打标样本应该尽可能多地代表周围的样本，代表性越大越好。该指标的意义是：被选择的样本能够包含更多样本的信息，从而提升打标的性价比；而且可以避免主动学习回归算法选择到离群点，因为离群点所含有的信息往往噪声很大，打标性价比很低。度量方式举例：待打标样本所在位置的样本密度，样本密度越大，代表性越大。

信息量：主动学习回归算法选择的待打标样本应该包含更多的信息。该指标的意义是：对算法选择的待打标样本打标后，应该对当前的优化目标函数产生最有益的影响，信息量越大越好。度量方式举例：样本的预测标签的置信度，置信度越低信息量越大；对其打标后会给当前回归模型带来的变化程度，变化越大信息量越大。

其中，分散度和代表性两个核心指标的评估不需要标签信息，因此可以直接迁移到无监督主动学习回归算法中。而信息量指标的评估需要当前回归模型的信息，即需要真实标签信息，不再适用于无监督情形，本文将为无监督主动学习回归算法重新定义信息量指标：

信息量（无监督主动学习回归算法）：根据当前回归模型类型推导出的特定指标或采样方法，能够帮助算法选择更适应于该类回归模型的待打标样本，该指标的评估要求使用的回归模型类型已知。该指标的意义在于：针对特定类型的回归模型，更好地度量分散度或代表性，或指导算法选择更好的待打标样本。度量方式举例：针对线性模型，可以使用特征空间中点到流形的距离来更好地度量分散度。

2.4.2 分散度、代表性和信息量的理论解释

2.4.2.1 针对线性回归模型

设待打标样本数为M，特征维数为d。

1. 分散度和代表性的作用

此时的样本空间是二维的，根据平面几何知识，在二维空间中，两个不同的点便可以确定一条直线。当需要采集的待打标样本集合只含有两个样本时，分散度即为：

特征空间中两个特征点的欧氏距离；代表性即为：特征空间中两个特征点所在位置的样本密度。此时最优回归超平面H^*和预测回归超平面$\hat{H}$均退化为了直线，它们的法向量夹角$|\theta|$也等于这两条直线的夹角$|\theta'|$，直线$\hat{H}$与直线H^*的夹角$|\theta'|$越小，对应的待打标样本集合越优。

1) $M = d + 1$，且$d = 1$的情形

由于分散度和代表性并不是独立的，这里采用控制变量法分析两者对于预测回归模型的影响。

a) 控制分散度不变，研究代表性的影响

首先证明：样本$\bar{\mathbf{x}}_i$的代表性越大，其预测样本点$\hat{\mathbf{x}}_i$到最优回归超平面H^*的距离往往越小，即随机变量$\hat{\mathbf{x}}_i$的方差越小。

证明如下：设特征点$\bar{\mathbf{x}}_i$的代表性大于特征点$\bar{\mathbf{x}}_j$，那么特征点$\bar{\mathbf{x}}_i$处的样本密度也大于特征点$\bar{\mathbf{x}}_j$，特征点$\bar{\mathbf{x}}_i$周围的近邻样本数量也多于特征点$\bar{\mathbf{x}}_j$，由于距离相近的特征点真实标签往往也相近，因此与特征点$\bar{\mathbf{x}}_i$真实标签相近的特征点的个数往往也多于特征点$\bar{\mathbf{x}}_j$，那么，特征点$\bar{\mathbf{x}}_i$对回归模型的影响力往往也大于特征点$\bar{\mathbf{x}}_j$。由于最优回归超平面H^*是由样本池中所有预测样本点$\{\hat{\mathbf{x}}_n\}_{n=1}^N$训练得到，因此预测样本点$\hat{\mathbf{x}}_i$往往更接近最优回归超平面$H^*$，即随机变量$\hat{\mathbf{x}}_i$的方差减小，证毕。

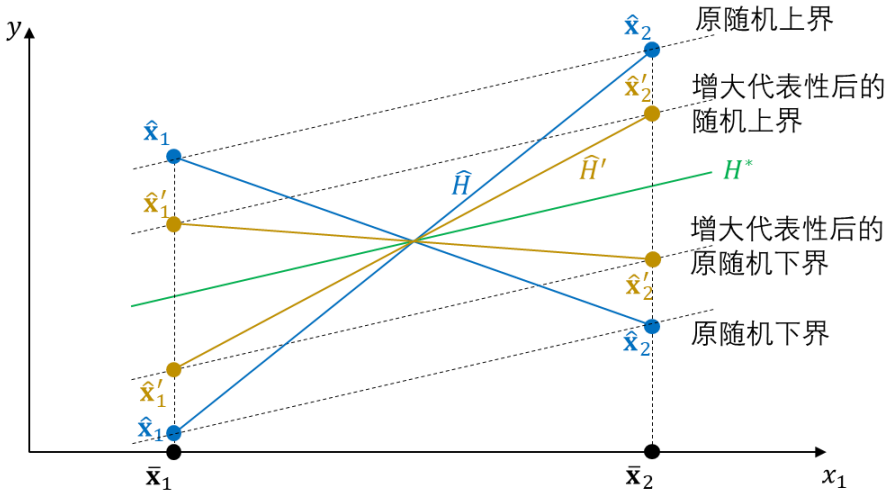

图 2-3 特征空间为一维且分散度不变时，代表性的变化对预测回归模型的影响示意图

如图 2-3 所示，分散度不变即代表特征点$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$的距离$\left|\overrightarrow{\bar{\mathbf{x}}_1\bar{\mathbf{x}}_2}\right|$不变，若$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$的代表性增大，那么预测样本点$\hat{\mathbf{x}}_1$和$\hat{\mathbf{x}}_2$的方差减小，那么显然最优回归直线$H^*$和预测回归直线$\hat{H}$之间的夹角$|\theta'|$的方差减小，即预测回归直线的精度越高。即分散度不变时，代表性越大，待打标样本集越优。

b) 控制代表性不变，研究分散度的影响

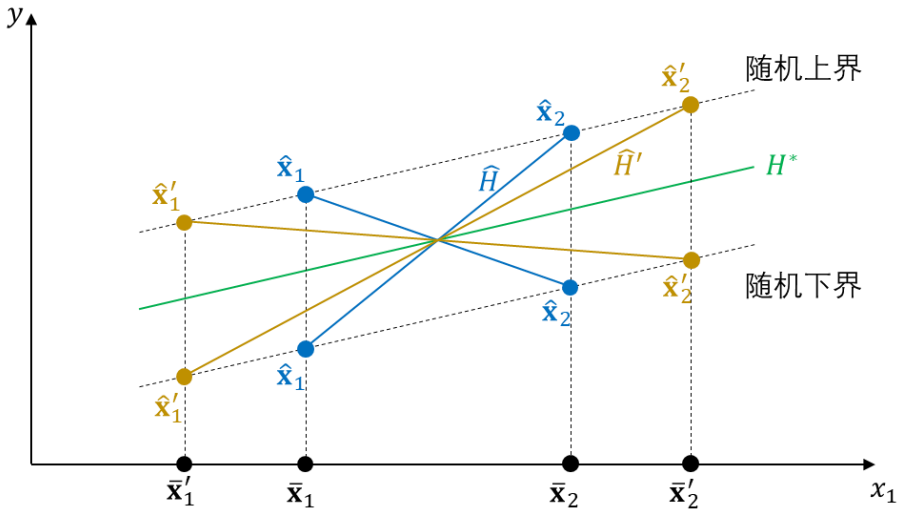

图 2-4 特征空间为一维且代表性不变时，分散度的变化对预测回归模型的影响示意图

如图 2-4 所示，代表性不变即代表$\hat{\mathbf{x}}_1$和$\hat{\mathbf{x}}_2$的方差不变。若分散度增大，那么特征点$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$的距离$\left|\overrightarrow{\bar{\mathbf{x}}_1\bar{\mathbf{x}}_2}\right|$增加，根据平面几何知识，显然最优回归直线$H^*$和预测回归直线$\hat{H}$之间的夹角$|\theta'|$的方差将变小，即预测回归直线的精度越高。即代表性不变时，分散度越大，待打标样本集越优。

2) $M \neq d+1$或$d > 1$的情形

此时预测回归超平面$\hat{H}$和最优回归超平面H^*的正法向量夹角$|\theta|$不能直观表示，可以考虑将它近似地拆分为多个简单问题的叠加。设待打标样本集为$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，$\{\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j\}$为其中任意两点，它们在最优回归超平面上对应的样本点是$\{\mathbf{x}_i^*, \mathbf{x}_j^*\}$，它们的预测样本点为$\{\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j\}$，考察最优回归超平面上的直线$\overrightarrow{\mathbf{x}_i^*\mathbf{x}_j^*}$和预测样本点确定的直线$\overrightarrow{\hat{\mathbf{x}}_i\hat{\mathbf{x}}_j}$之间的差异，则与一维特征空间中的问题一致，能够反映局部地反映预测回归超平面

$\widehat{H}$和最优回归超平面H^*的差异。这样的组合有C_M^2种，于是该问题可以看作C_M^2个一维特征空间中的简单问题的融合。总体来看，如果对于所有的$\{\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j\}$，直线$\overrightarrow{\hat{\mathbf{x}}_i\hat{\mathbf{x}}_j}$与直线$\overrightarrow{\mathbf{x}_i^*\mathbf{x}_j^*}$的差异普遍更小，那么预测回归超平面$\widehat{H}$和最优回归超平面$H^*$的差距也会更小，那么这些简单问题融合之后，分散度和代表性的影响也与 1 中的结论一致。即分散度不变时，代表性越大，待打标样本集越优；代表性不变时，分散度越大，待打标样本集越优。

2. 信息量的作用

分散度和代表性仅通过样本间距离或样本密度等不依赖于回归模型的方式来度量。由 1 中的分析可得，$M \neq d+1$或$d > 1$的情形中，关于分散度指标的分析存在近似处理，这或多或少会影响指标的准确度。如果已知回归模型的类型，便能推导出针对某一类回归模型的信息量指标来更好地度量分散度或代表性，或形成新指标更好地指导待打标样本的选择，它往往能帮助算法针对特定回归模型更好地选择待打标样本，这便是信息量指标的作用。后文将提出一种针对线性回归的信息量指标。

2.4.2.2 针对基于核的非线性回归模型

首先给出结论，只要核函数满足如下条件：任意两点在原始特征空间中的欧式距离增加，它们在映射后的希尔伯特空间中对应的点的欧式距离也增加，那么分散度、代表性和信息量的理论解释在基于核的非线性回归模型中依然成立，下面进行解释。

基于核的非线性回归模型在训练之前，首先使用核函数将原始特征空间映射到高维的希尔伯特空间$\mathbb{R}^k, k > d$，然后在高维的希尔伯特空间上进行线性回归。本文以径向基核函数（Radial basis function，简称：RBF）为例进行分析，该核函数表达式为：

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\lambda\|\mathbf{x}_i - \mathbf{x}_j\|^2}$$

设$\mathbf{x}_i$和$\mathbf{x}_j$是线性空间$\mathbb{R}^d$中的任意两个点，它们的欧式距离是：

$$d_1 = \|\mathbf{x}_i - \mathbf{x}_j\|$$

经过径向基核函数的映射之后，它们在希尔伯特空间中的欧氏距离是：

$$d_2 = \sqrt{k\left((\mathbf{x}_i - \mathbf{x}_j),(\mathbf{x}_i - \mathbf{x}_j)\right)} = \sqrt{2 - 2k(\mathbf{x}_i, \mathbf{x}_j)}$$

因为$k(\mathbf{x}_i, \mathbf{x}_j)$和$\|\mathbf{x}_i - \mathbf{x}_j\|$负相关，所以$d_2$和$d_1$正相关。因此，如果任意两个点$\mathbf{x}_i$和$\mathbf{x}_j$在原线性空间$\mathbb{R}^d$中距离越大，那么它们在希尔伯特空间$\mathbb{R}^k$中的距离也越大；也就是说，如果一群点在原线性空间$\mathbb{R}^d$中越分散，那么它们在希尔伯特空间$\mathbb{R}^k$中也越分散；也就是说，如果任意点$\mathbf{x}_i$在原线性空间$\mathbb{R}^d$中越接近离群点，那么在希尔伯特空间$\mathbb{R}^k$中它也越接近离群点。因此在原线性空间$\mathbb{R}^d$和希尔伯特空间$\mathbb{R}^k$中分散度和代表性的理论解释依然成立。同理，分散度和代表性指标存在近似，若能通过已知的回归模型类型来推导出其他的方式来更好地度量分散度或代表性，或更好地指导待打标样本的选择，那么便是加入了依赖于回归模型类型的信息量指标。

2.5 本章小结

详细对比了无监督主动学习回归算法和有监督主动学习回归算法，指出无监督主动学习回归算法相对于有监督主动学习回归算法具有如下三个核心优势：无需任何真实标签；只需与人工专家交互一次；在训练样本很少时取得更好的性能。

为无监督主动学习回归算法建立了基于回归模型不确定性的数学模型，并提出可以使用原特征空间或希尔伯特空间中的预测回归超平面与最优回归超平面的正法向量夹角大小来评估预测回归模型的精度。

将分散度、代表性和信息量三个核心指标从有监督主动学习回归算法迁移到无监督主动学习回归算法，针对无监督主动学习回归算法重新定义了信息量指标，此时的信息量指标依赖于已知的回归模型类型。而且还为这三个核心指标提供了理论解释。

3 两种新的无监督主动学习回归算法

本章首先提出一种在无监督主动学习回归算法中用于优化待打标样本集的交替优化框架；随后提出两种新的无监督主动学习回归算法：“通过迭代来最大化分散度和代表性的无监督主动学习回归算法”（Iterative Representativeness-Diversity Maximization，简称 iRDM）以及“同时考虑分散度、代表性和信息量的无监督主动学习回归算法”（Diversity, Representativeness and Informativeness，简称 IRD），算法的推导均基于上一章（第 2.2 节）建立的数学模型和提出的预测指标。

3.1 无监督主动学习回归算法中的交替优化框架

从样本池的N个样本中选择出M个待打标样本，有C_N^M种选择方式，这是一个组合爆炸问题，不能通过遍历的方式求解。本文中的思路是，先任意获取M个待打标样本作为初始待打标样本集，然后通过优化的方式，逐渐优化这M个待打标样本，最终获取最优的待打标样本集。

然后介绍如何优化这M个待打标样本。这是一个M目标的优化问题，且目标之间存在耦合，因此很难同时优化。本文借鉴期望最大化算法（Expectation Maximization，简称：EM 算法）的思想，使用一个交替优化框架来优化，每次只优化一个待打标样本，固定其余$M-1$个待打标样本，对每个样本均优化一次即完成一轮迭代，进行若干轮迭代后，若本轮优化的结果在之前某轮优化中出现过，则判定算法收敛。这样M目标的优化问题就被拆分为若干单目标优化问题。当然不排除出现一些不收敛的状况，因此设定最大迭代轮数$c_{\max}$，达到时算法终止并输出当前结果。

本文提出的无监督主动学习回归算法中优化待打标样本集的交替优化框架示意图见图 3-1：

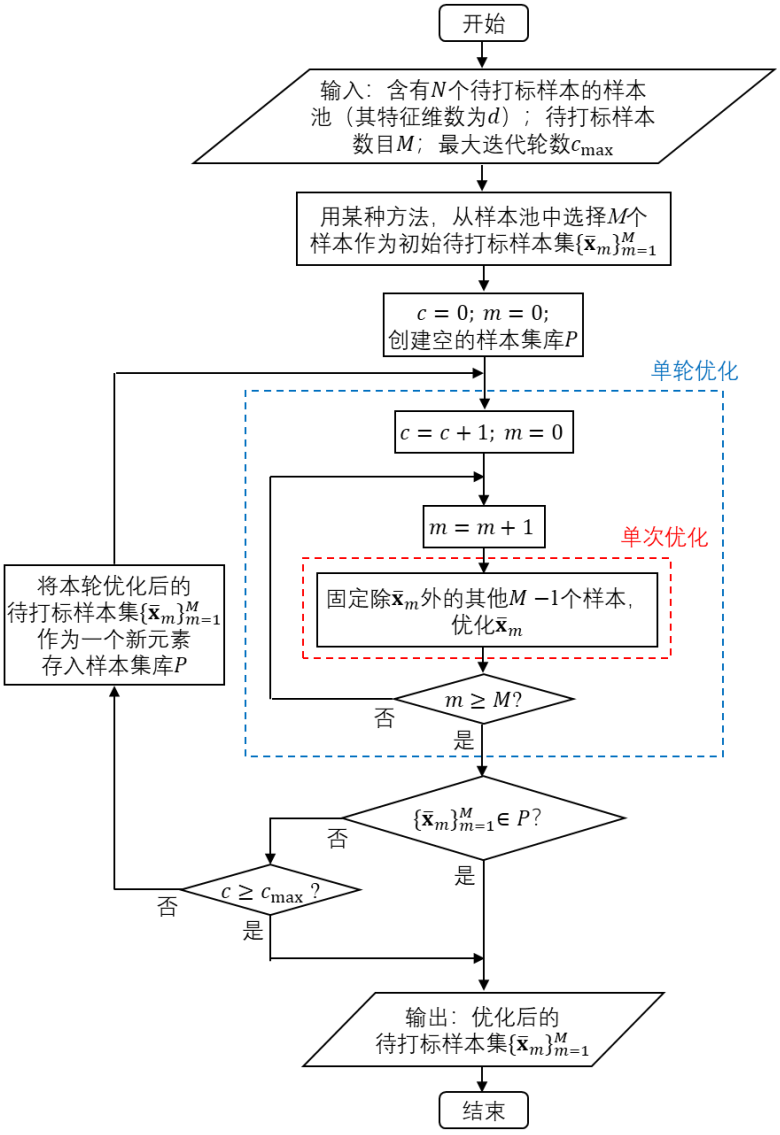

图 3-1 无监督主动学习回归算法中优化待打标样本集的交替优化框架示意图

3.2 通过迭代来最大化分散度和代表性的算法（iRDM）

本节将详细介绍本文提出的一种新的无监督主动学习回归算法：iRDM 算法。该算法将度量分散度和代表性指标，并将它们加入优化目标函数中。iRDM 算法使用上一节（第 3.1 节）提出的交替优化框架来优化待打标样本集，因此只需要介绍如何选择初始的待打标样本集，以及如何进行单次优化即可。

首先介绍选择初始的M个待打标样本的方式。这里可以使用随机选择，也可以使用现有的无监督主动学习回归算法。RD 算法的性能普遍优于 GSx，而 P-ALICE 算法的实现较为复杂且需要对样本进行加权，因此 iRDM 算法选择 RD 算法来选择初始的待打标样本集。RD 算法中，k-均值聚类算法已经将样本池划分为M个簇$\{C_m\}_{m=1}^{M}$，每个簇只含有一个待打标样本。为缩小优化范围，我们将每个待打标样本的优化范围限制在其所属的簇内。

3.2.1 分散度与代表性的权衡

iRDM 算法使用了分散度和代表性两个核心指标，因此有必要先分析它们之间的关系。分散度与代表性两个核心指标同时影响着预测回归模型的性能，但是通常无法同时最大化分散度和代表性，这是因为：随着分散度的增加，被选择的待打标样本会越来越接近离群点，而越接近离群点，代表性就越差；相反，如果被选择的待打标样本的代表性最大，那么它们会接近聚类中心，这时它们之间的分散度不一定是最优的，可能还有优化的空间。

因此，无监督主动学习回归算法需要同时考虑分散度和代表性两个核心指标，并权衡它们以找到一个平衡点，使得预测回归模型最接近最优回归模型。特征空间为一维时，分散度与代表性的权衡关系如图 3-2 所示。

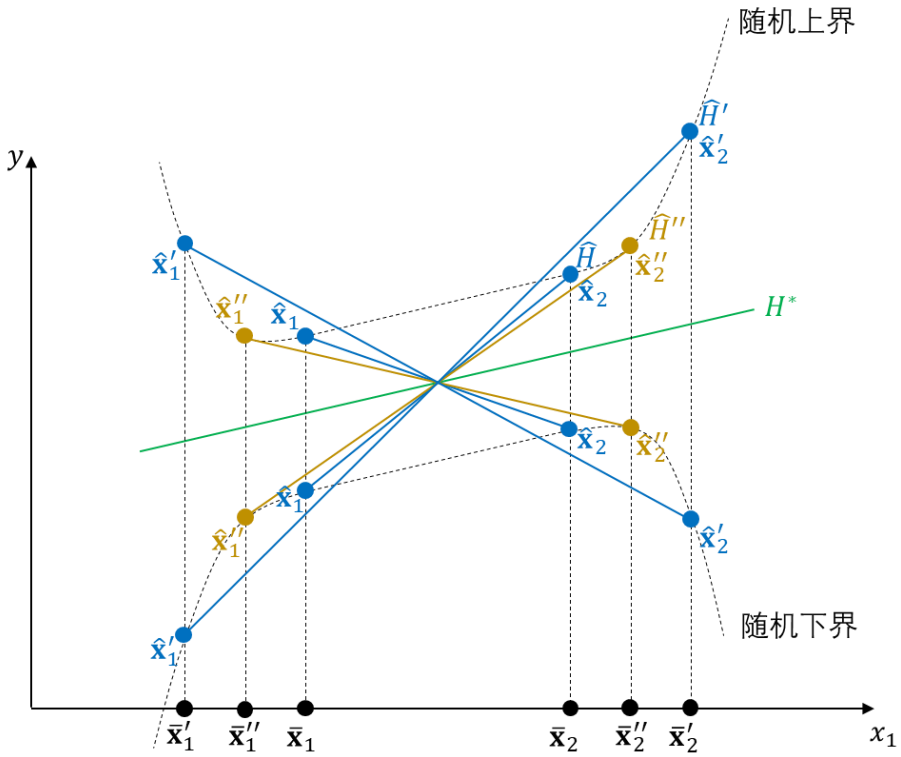

图 3-2 特征空间为一维时，分散度与代表性之间的权衡关系示意图

图中待打标样本集$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2\}$对应的预测样本点$\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2\}$距离最优回归直线$H^*$最近，因此$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2\}$的代表性最大，但它们距离最近，分散度最小；待打标样本集$\{\bar{\mathbf{x}}_1', \bar{\mathbf{x}}_2'\}$对应的预测样本点$\{\hat{\mathbf{x}}_1', \hat{\mathbf{x}}_2'\}$距离最优回归直线$H^*$最远，因此$\{\bar{\mathbf{x}}_1', \bar{\mathbf{x}}_2'\}$代表性最小，但它们距离最远，分散度最大（它们更接近离群点）；待打标样本集$\{\bar{\mathbf{x}}_1'', \bar{\mathbf{x}}_2''\}$的分散度和代表性均适中。很明显，使用待打标样本集$\{\bar{\mathbf{x}}_1'', \bar{\mathbf{x}}_2''\}$训练出的预测回归直线$\widehat{H}''$与最优回归直线$H^*$的夹角$|\theta'|$最小，即预测回归直线$\widehat{H}''$最接近最优回归直线$H^*$。

对比现有的无监督主动学习回归算法，GSx 算法只最大化分散度没有考虑代表性，因此更倾向于选择待打标样本集$\{\bar{\mathbf{x}}_1', \bar{\mathbf{x}}_2'\}$；RD 算法最大化代表性，而没有优化分散度，因此更倾向于选择待打标样本集$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2\}$；iRDM 算法平衡了分散度和代表性并找到了它们的平衡点，因此更倾向于选择最优的待打标样本集$\{\bar{\mathbf{x}}_1'', \bar{\mathbf{x}}_2''\}$。

3.2.2 iRDM 算法中单次优化的目标函数

首先说明需要用到的符号。不妨设当前的待打标样本集是$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，注意这里的符号$\bar{\mathbf{x}}_m$指的是当前的待打标样本集中的第$m$个样本，它会随着优化的进行而发生改变（需要区别于样本池中第m个样本$\mathbf{x}_m$）。设本次优化中待优化样本为$\bar{\mathbf{x}}_m$，其所在的聚类簇是C_m，C_m中的样本数量是N_m，C_m中的样本为$\{\mathbf{x}_n\}_{n=1}^{N_m}$。

在单次优化中，iRDM 算法需要使用目标函数对聚类簇C_m中每个样本进行打分，打分最高的样本$\mathbf{x}_n^*$将被选中，替换原来的样本，成为新的待打标样本$\bar{\mathbf{x}}_m$。优化过程可以表示为：

$$\bar{\mathbf{x}}_m = \mathbf{x}_n^* = \arg\max_{n\in[1,N_m]} Q(\mathbf{x}_n)$$

接下来介绍目标函数$Q(\mathbf{x}_n)$，它由分散度度量和代表性度量融合而成。

首先介绍分散度的度量，用符号D表示。在待打标样本数量$M=2$时，分散度就是两个样本间的距离；而在待打标样本数量$M>2$的情形下，待优化样本与$M-1$个固定样本各有一个欧氏距离，共有$M-1$个距离，借鉴 GSx 算法[39]中贪婪距离的思想，选择这$M-1$个距离中最小的距离作为分散度的度量D，即对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算$D(\mathbf{x}_n)$：

$$D(\mathbf{x}_n) = \min_{\substack{m\in[1,M]\\ m\neq n}} \|\mathbf{x}_n - \mathbf{x}_m\|,\ n=1,\dots,N_m \tag{3-1}$$

$D(\mathbf{x}_n)$越大，代表分散度越大，$\mathbf{x}_n$越值得被选择。

然后介绍代表性的度量，用符号R表示。本文使用待优化样本为$\bar{\mathbf{x}}_m$到聚类簇C_m内其他点的平均距离来度量代表性R，即对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算$R(\mathbf{x}_n)$：

$$R(\mathbf{x}_n) = \frac{1}{N_m-1} \times \sum_{\mathbf{x}_i\in C_m} \|\mathbf{x}_n - \mathbf{x}_i\|,\ n=1,\dots,N_m \tag{3-2}$$

$R(\mathbf{x}_n)$越小，代表代表性越大，$\mathbf{x}_n$越值得被选择。根据k-均值聚类算法的聚类规则，当待优化样本$\bar{\mathbf{x}}_m$最接近聚类簇C_m的类中心时，$R(\bar{\mathbf{x}}_m)$最小，此时代表性最大。而当待优化样本$\bar{\mathbf{x}}_m$接近离群点时，其距离类内很多样本都很远，因此$R(\bar{\mathbf{x}}_m)$较大，因此该项可以限制待优化样本$\bar{\mathbf{x}}_m$成为离群点。

上一节（第 3.2.1 节）提到，分散度和代表性存在权衡关系，因此目标函数中需要同时考虑它们，这里使用加法的方式融合分散度和代表性，即对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算目标函数值$Q(\mathbf{x}_n)$：

$$Q(\mathbf{x}_n) = D(\mathbf{x}_n) \times \frac{1}{R(\mathbf{x}_n)} \tag{3-3}$$

目标函数值$Q(\mathbf{x}_n)$是对分散度和代表性的综合度量：分散度越大，$D(\mathbf{x}_n)$越大；代表性越大，$R(\mathbf{x}_n)$越小，即$\frac{1}{R(\mathbf{x}_n)}$越大。总而言之，分散度的增加或代表性的增加都会增加$Q(\mathbf{x}_n)$，但由于分散度增加时代表性一般会减小，因此使用$Q(\mathbf{x}_n)$可以找到它们的平衡点。

3.2.3 iRDM 算法的流程

本节将总结 iRDM 算法流程，并画出算法流程图。设需要从样本池中选择M个待打标样本。

首先用 RD 算法选择初始的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，即：使用$k$-均值聚类算法（$k = M$）对样本池进行聚类，聚为$M$个簇$\{C_m\}_{m=1}^{M}$，对每一个簇，选择距离类中心最近的样本作为待打标样本。

然后创建一个空的待打标样本集库P，开始迭代过程。每一轮迭代进行M次优化过程，分别优化待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$每一个样本，每次优化只优化一个样本，其余$M-1$个样本固定。设当前单次优化中待优化的待打标样本为$\bar{\mathbf{x}}_m$，它所在的簇为$C_m$，则对$C_m$中的每一个样本$\mathbf{x}_n$，根据公式(3-1)计算其分散度度量$D(\mathbf{x}_n)$；并根据公式(3-2)计算其代表性度量$R(\mathbf{x}_n)$；最后通过公式(3-3)计算其目标函数值$Q(\mathbf{x}_n)$，并选择簇$C_m$中目标函数值最大的样本$\mathbf{x}_n^*$，替换原来的样本作为待打标样本$\bar{\mathbf{x}}_m$。

本轮优化完成后，判断当前待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$是否已经存在于待打标样本集库$P$中，若已经存在则输出当前待打标样本集，算法结束；若不存在则判断是否达到最大迭代轮数，如果达到则输出当前待打标样本集，算法结束，否则将当前的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$作为一个新元素加入待打标样本集库$P$中，开始下一轮迭代。

iRDM 算法的流程图如图 3-3 所示：

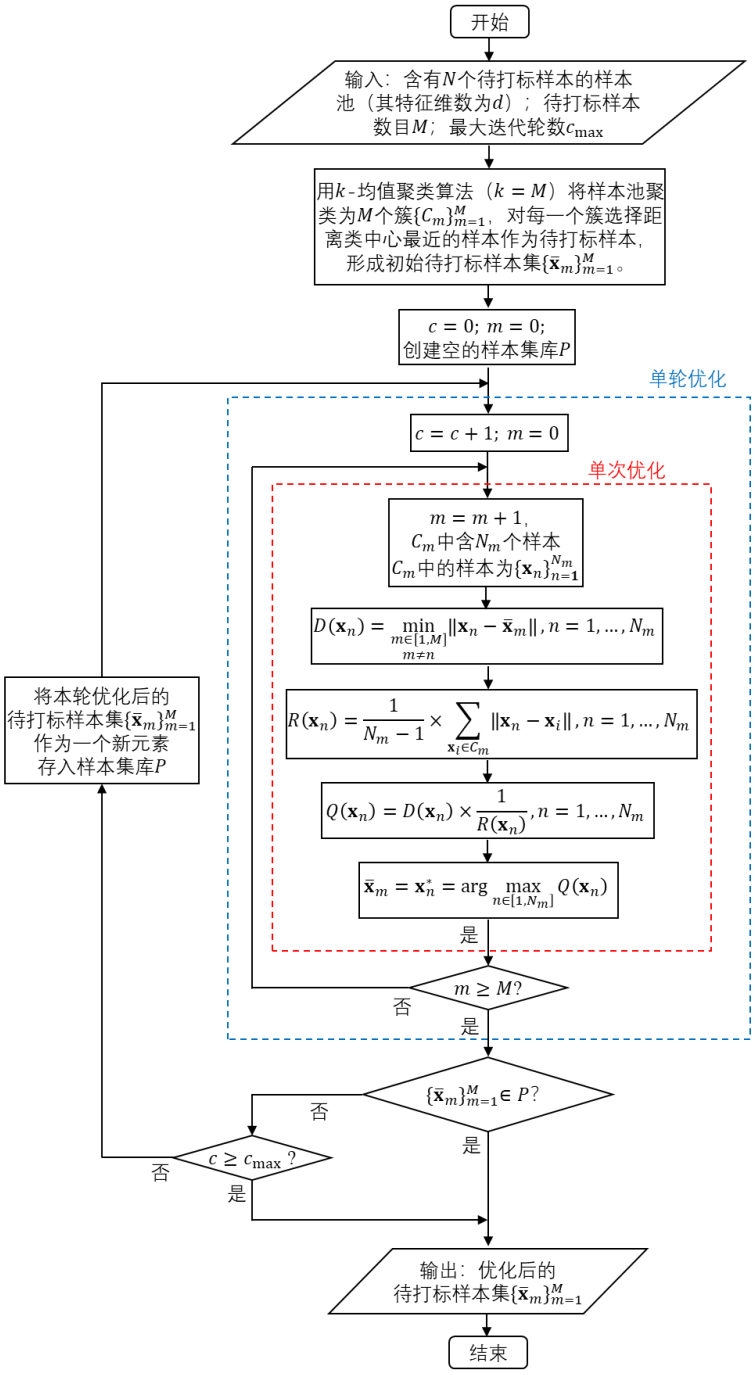

图 3-3 iRDM 算法流程图

3.3 同时考虑分散度、代表性和信息量的算法（IRD）

本节将详细介绍本文提出的另一种新的无监督主动学习回归算法：IRD 算法。该算法将度量分散度-信息量联合指标和代表性指标，并将它们加入优化目标函数中。

IRD 算法同样使用上一节（第 3.1 节）提出的交替优化框架来优化待打标样本集的选择，因此只需要介绍如何选择初始的待打标样本集，以及如何进行单次优化即可。

根据待打标样本数M和特征维数d的关系，IRD 算法分为$M \le d+1$和$M > d+1$两种情形。

3.3.1 $M \le d+1$时的情形

3. 3. 1. 1 $M \le d+1$时待打标样本位置对预测回归模型的影响

当$M < d+1$时，先使用降维算法将特征空间降维到$d' = M-1$维度，便等同于$M = d+1$的情形，因此只分析$M = d+1$的情形。

设待打标样本集为$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，在$M = d+1$时，$d+1$个待打标样本$\{\bar{\mathbf{x}}_m\}_{m=1}^{d+1}$在样本空间$\mathbb{R}^{d+1}$中对应的预测样本点$\{\hat{\mathbf{x}}_m\}_{m=1}^{d+1}$刚好可以确定样本空间$\mathbb{R}^{d+1}$中的一个$d$维的回归超平面，此时可以通过几何学的知识进行理论分析。不妨设本次优化中$\{\bar{\mathbf{x}}_m\}_{m=1}^{M-1}$固定，$\bar{\mathbf{x}}_M$为待优化样本。

1. 特征空间为二维时的情形

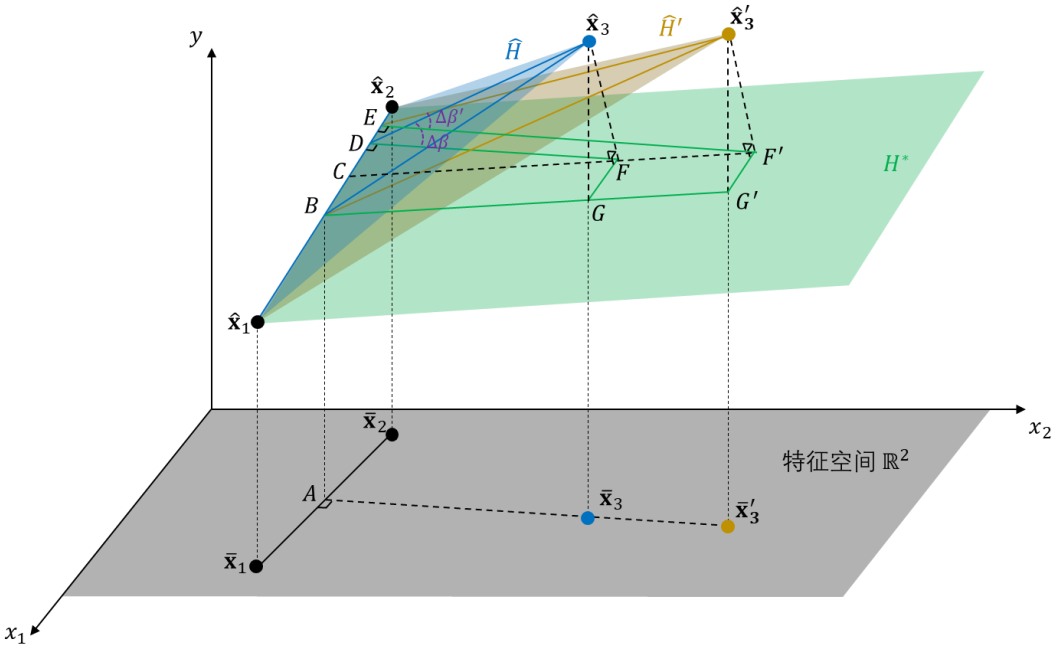

图 3-4 二维特征空间中，待优化样本$\bar{\mathbf{x}}_3$的代表性不变时，其位置对预测回归平面$\hat{H}$的影响

如图 3-4 所示，当特征空间为二维$\mathbb{R}^2$时，样本空间为三维$\mathbb{R}^3$，此时待打标样本集中只有三个样本$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{\mathbf{x}}_3\}$，不妨设$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$固定，$\bar{\mathbf{x}}_3$待优化。绿色平面$H^*$表示最优回归平面（即使用样本池中所有样本训练出的平面，该平面存在但未知），蓝色平面$\widehat{H}$表示优化前的预测回归平面，黄色平面$\widehat{H}'$表示优化后的预测回归平面。连接$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$即获得一条直线，我们用其方向向量$\overrightarrow{\bar{\mathbf{x}}_1\bar{\mathbf{x}}_2}$来表示，$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$在最优回归超平面$H^*$上对应的最优预测样本点是$\mathbf{x}_1^*$和$\mathbf{x}_2^*$，连接它们也可以获得一条直线$\overrightarrow{\mathbf{x}_1^*\mathbf{x}_2^*}$，$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$在预测回归超平面$\widehat{H}$上对应的预测样本点是$\hat{\mathbf{x}}_1$和$\hat{\mathbf{x}}_2$，连接它们也可以获得一条直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$。在特征空间$\mathbb{R}^2$中，过$\bar{\mathbf{x}}_3$作垂线垂直于直线$\overrightarrow{\bar{\mathbf{x}}_1\bar{\mathbf{x}}_2}$于点$A$，设$\bar{\mathbf{x}}_3'$在直线$\overrightarrow{\bar{\mathbf{x}}_3A}$的反向延长线上，$\bar{\mathbf{x}}_3$与$\bar{\mathbf{x}}_3'$在最优回归平面$H^*$上对应的最优预测样本点分别是点$G$和$G'$，其预测标签分别是$\hat{\mathbf{x}}_3$与$\hat{\mathbf{x}}_3'$，待打标样本集$\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{\mathbf{x}}_3'\}$对应的预测回归平面是$\widehat{H}'$。下面给出一个关键定理并证明之：

定理：若$\left|\overrightarrow{\hat{\mathbf{x}}_3G}\right| = \left|\overrightarrow{\hat{\mathbf{x}}_3'G'}\right|$，且直线$\overrightarrow{\mathbf{x}_1^*\mathbf{x}_2^*}$与直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$重合，那么$\widehat{H}'$的正法向量与$H^*$的正法向量的夹角大小$|\theta'|$小于$\widehat{H}$的正法向量与$H^*$的正法向量的夹角大小$|\theta|$。接下来证明之：

要满足条件“$\left|\overrightarrow{\hat{\mathbf{x}}_3G}\right| = \left|\overrightarrow{\hat{\mathbf{x}}_3'G'}\right|$”需要假设“$\mathbf{x}_3$与$\mathbf{x}_3'$的代表性一致”。第 2.4.2 节提到，代表性的作用是让待优化样本不要接近离群点，因此可以先令$\mathbf{x}_3$与$\mathbf{x}_3'$的代表性一致，在确定优化目标函数后，再将代表性的度量作为正则项加入即可。令$\mathbf{x}_3$与$\mathbf{x}_3'$的代表性一致，那么根据第 2.4.2 节中关于无监督主动学习中代表性的理论解释，此时线段$\left|\overrightarrow{\hat{\mathbf{x}}_3G}\right|$与$\left|\overrightarrow{\hat{\mathbf{x}}_3'G'}\right|$长度相同。

要满足条件“直线$\overrightarrow{\mathbf{x}_1^*\mathbf{x}_2^*}$与直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$重合”需要假设“固定变量$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$已经达到最优”。若将候选样本中的$M$个待打标样本看作$M$个待优化的变量，则$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$是固定的变量，$\bar{\mathbf{x}}_3$是待优化的变量。这里借鉴期望最大化算法（EM 算法）的假设，即固定变量$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$已经达到最优（或者接近最优，差距可以忽略不计）。与期望最大化算法一样，在最初的迭代中该条件可能很难满足，但随着迭代过程的进行，会逐渐满足该条

件，因此这个假设是合理的。若$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$优化地越好，则直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$越接近最优回归超平面$H^*$，若已经优化到足够好，则可以认为$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$就在最优回归超平面$H^*$上，即直线$\overrightarrow{\mathbf{x}_1^*\mathbf{x}_2^*}$与直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$重合。

于是定理有了另一种表述：若$\bar{\mathbf{x}}_3$和$\bar{\mathbf{x}}_3'$的代表性一样，且$\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$已经优化到足够好，那么$\bar{\mathbf{x}}_3'$相对于$\bar{\mathbf{x}}_3$是更优的选择。该定理将用于度量 IRD 算法中的分散度和信息量。

以上的两个假设分别令定理的两个条件成立，接下来为了证明定理，需要证明如下三引理：

引理一：过$\hat{\mathbf{x}}_3$与$\hat{\mathbf{x}}_3'$分别作垂线垂直于最优回归平面H^*于点F和F'，那么线段$\left|\overrightarrow{\hat{\mathbf{x}}_3F}\right|$与线段$\left|\overrightarrow{\hat{\mathbf{x}}_3'F'}\right|$长度相同。

证明：显然直线$\overrightarrow{\hat{\mathbf{x}}_3F}$平行于直线$\overrightarrow{\hat{\mathbf{x}}_3'F'}$，三角形$\Delta\hat{\mathbf{x}}_3GF$与$\Delta\hat{\mathbf{x}}_3'G'F'$均为直角三角形，又因为线段$\left|\overrightarrow{\hat{\mathbf{x}}_3G}\right|$与$\left|\overrightarrow{\hat{\mathbf{x}}_3'G'}\right|$长度相同，直线$\overrightarrow{\hat{\mathbf{x}}_3G}$与$\overrightarrow{\hat{\mathbf{x}}_3'G'}$的方向相同（平行于标签轴$y$），于是三角形$\Delta\hat{\mathbf{x}}_3GF$与$\Delta\hat{\mathbf{x}}_3'G'F'$全等，即线段$\left|\overrightarrow{\hat{\mathbf{x}}_3F}\right|$与线段$\left|\overrightarrow{\hat{\mathbf{x}}_3'F'}\right|$长度相同，证毕。

引理二：过F和F'分别作垂线垂直于直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$于$D$和$E$，那么线段$\left|\overrightarrow{DF}\right|$的长度小于线段$\left|\overrightarrow{EF'}\right|$的长度。

证明：首先三角形DCF和三角形ECF'都是直角三角形，又因为$\overrightarrow{DF}//\overrightarrow{EF'}$，所以$\angle DFC=\angle EF'C$，所以三角形$DCF$相似于三角形$ECF'$；又因为$\left|\overrightarrow{\bar{\mathbf{x}}_3A}\right|<\left|\overrightarrow{\bar{\mathbf{x}}_3'A}\right|$，所以$\left|\overrightarrow{CF}\right|<\left|\overrightarrow{CF'}\right|$，所以$\left|\overrightarrow{DF}\right|<\left|\overrightarrow{EF'}\right|$，证毕。

引理三：优化前的预测回归平面$\widehat{H}$与最优回归平面H^*的正法向量的夹角大小$|\theta|$等于直线$\overrightarrow{D\hat{\boldsymbol{x}}_3}$和直线$\overrightarrow{DF}$的夹角大小$|\beta|$；优化后的预测回归平面$\widehat{H}'$与最优回归平面$H^*$的正法向量的夹角大小$|\theta'|$等于直线$\overrightarrow{E\hat{\boldsymbol{x}}_3'}$和直线$\overrightarrow{EF'}$的夹角$|\beta'|$。

证明：因为直线$\overrightarrow{\hat{\mathbf{x}}_3F}\perp H^*$，所以$\overrightarrow{\hat{\mathbf{x}}_3F}\perp\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$，又因为$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{D\hat{\mathbf{x}}_3}$，所以$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp$平面$D\hat{\mathbf{x}}_3F$，所以$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{DF}$；因为直线$\overrightarrow{\hat{\mathbf{x}}_3'F'}\perp H^*$，所以$\overrightarrow{\hat{\mathbf{x}}_3'F'}\perp\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$，又因为$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{E\hat{\mathbf{x}}_3'}$，所以$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp$平面$D\hat{\mathbf{x}}_3'F'$，所以$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{EF'}$；因为$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{D\hat{\mathbf{x}}_3}$，$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{DF}$，所以$|\theta|$等于直线$\overrightarrow{D\hat{\mathbf{x}}_3}$和直线$\overrightarrow{DF}$的夹角$|\beta|$；因为$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{E\hat{\mathbf{x}}_3'}$，$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}\perp\overrightarrow{EF'}$，所以$|\theta'|$等于直线$\overrightarrow{E\hat{\mathbf{x}}_3'}$和直线$\overrightarrow{EF'}$的夹角$|\beta'|$。

由引理一和引理二，$\tan|\beta'|=\left|\overrightarrow{\hat{\mathbf{x}}_3'F'}\right|/\left|\overrightarrow{EF'}\right|<\tan|\beta|=|\hat{\mathbf{x}}_3F|/\left|\overrightarrow{DF}\right|$，即$|\beta'|<|\beta|$，又由引理三，$|\theta'|<|\theta|$，定理证毕。

有了该定理可知，在$\bar{\mathbf{x}}_3$代表性不变的情况下，$|\theta|$与$\left|\overrightarrow{\bar{\mathbf{x}}_3\bar{A}}\right|$呈负相关关系，即：

$$|\theta|\propto\frac{1}{\left|\overrightarrow{\bar{\mathbf{x}}_3\bar{A}}\right|}$$

2. 特征空间为高维时的情形

此时单次优化中，不妨设前$d=M-1$个样本$\{\bar{\mathbf{x}}_m\}_{m=1}^{d}$固定，第$d+1$个样本$\mathbf{x}_{d+1}$待优化。

此时最优回归模型H^*、优化前的预测回归平面$\hat{H}$和优化后的预测回归平面$\hat{H}'$均对应的是样本空间$\mathbb{R}^{d+1}$中的d维超平面。直线$\overrightarrow{\bar{\mathbf{x}}_1\bar{\mathbf{x}}_2}$进化为特征空间$\mathbb{R}^d$中的$d-1$维的线性流形$\bar{F}$，该流形正好可以由$d$个无共线性特征点$\{\bar{\mathbf{x}}_m\}_{m=1}^{d}$唯一确定（备注：二维空间中共线性即指：三个点在同一条一维直线上，M维空间中的共线性指：所有样本都在同一个$M-1$维或更低维度的线性流形上。这里假设固定的$M-1$个特征点无共线性，即使出现了共线性，则会出现若干满足条件的“线性流形簇”，此时任选其中一个线性流形即可，不影响推导和结论）。直线$\overrightarrow{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}$进化为样本空间$\mathbb{R}^{d+1}$中的$d-1$维的线性流形$\hat{F}$，该流形正好可以由$d$个固定的预测样本点$\{\hat{\mathbf{x}}_m\}_{m=1}^{M-1}$唯一确定。

在特征空间$\mathbb{R}^d$中，过待优化的特征点$\bar{\mathbf{x}}_{d+1}$作垂线垂直于d个固定特征点$\{\mathbf{x}_m\}_{m=1}^{d}$唯一确定的$d-1$维的线性流形$\bar{F}$于点$A$，在线段$\overrightarrow{\mathrm{A}\bar{\mathbf{x}}_{d+1}}$的延长线上找点$\bar{\mathbf{x}}'_{d+1}$。这与特征空间为二维时类似。

推导过程与二维空间类似，受篇幅限制，这里不再详述，最终结论与特征空间为二维时一致：在$\bar{\mathbf{x}}_{d+1}$代表性不变的情况下，优化前的预测回归平面$\widehat{H}$与最优回归平面H^*的正法向量的夹角大小$|\theta|$大于优化后的预测回归平面$\widehat{H}'$于最优回归平面H^*的正法向量的夹角大小$|\theta'|$，即$\bar{\mathbf{x}}'_{d+1}$相对于$\bar{\mathbf{x}}_{d+1}$是更优的选择。即在$\bar{\mathbf{x}}_{d+1}$代表性不变的情况下，$|\theta|$与待优化样本$\bar{\mathbf{x}}_{d+1}$到d个固定的特征点确定的流形$\bar{F}$的距离$dist(\bar{\mathbf{x}}_{d+1},\bar{F})$呈负相关关系，即：

$$|\theta| \propto \frac{1}{dist(\bar{\mathbf{x}}_{d+1},\bar{F})}$$

3.3.1.2 $M \le d+1$时，IRD 算法单次优化的目标函数

首先说明需要用到的符号。不妨设当前的待打标样本集是$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，注意这里的符号$\bar{\mathbf{x}}_m$指的是当前的待打标样本集中的第$m$个样本，它会随着优化的进行而发生改变（需要区别于样本池中第m个样本$\mathbf{x}_m$）。设本次优化中待优化样本为$\bar{\mathbf{x}}_m$，其所在的聚类簇是C_m，C_m中的样本数量是N_m，C_m中的样本为$\{\mathbf{x}_n\}_{n=1}^{N_m}$。

在单次优化中，iRDM 算法需要使用目标函数对聚类簇C_m中每个样本进行打分，打分最高的样本$\mathbf{x}_n^*$将被选中，替换原来的样本，成为新的待打标样本$\bar{\mathbf{x}}_m$。优化过程可以表示为：

$$\bar{\mathbf{x}}_m = \mathbf{x}_n^* = \arg\max_{n\in[1,N_m]} Q(\mathbf{x}_n)$$

接下来介绍目标函数$Q(\mathbf{x}_n)$，它融合了分散度与信息量度量以及代表性度量。

1. 分散度-信息量联合度量

首先介绍分散度-信息量联合度量，用符号$dist(\bar{\mathbf{x}}_m,\bar{F})$表示。设固定的$M-1$个样本在特征空间$\mathbb{R}^d$中确定的$d-1$维线性流形是$\bar{F}$，根据第 3.3.1 节，分散度和信息量的度量$dist(\bar{\mathbf{x}}_m,\bar{F})$即为待优化待打标样本$\bar{\mathbf{x}}_m$到线性流形$\bar{F}$的距离。

为了计算$dist(\bar{\mathbf{x}}_m, \bar{F})$，首先需要确定在特征空间$\mathbb{R}^d$中线性流形$\bar{F}$的方程（由于特征空间$\mathbb{R}^d$的维度只比线性流形$\bar{F}$的维度高一维，因此线性流形$\bar{F}$也是特征空间$\mathbb{R}^d$中的超平面，其正法向量方向唯一，用一个方程即可表示它），即确定方向系数$\mathbf{w}$和偏置项b，满足：

$$\mathbf{x}^T\mathbf{w} + b = 0$$

已知$M-1$个固定特征点均在线性流形$\bar{F}$上，将这些特征点的坐标代入方程：

$$\begin{bmatrix} \mathbf{x}_1{}^T & 1 \\ \mathbf{x}_2{}^T & 1 \\ \cdots & \cdots \\ \mathbf{x}_{M-1}{}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = \mathbf{0} \tag{3-4}$$

固定的样本有$M-1$个，由于$M = d+1$，因此正好有d个方程，根据之前的假设，这固定的$M-1$个特征向量线性无关，因此方程(3-4)只存在一个基础解系：

$$\begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = k \begin{bmatrix} \tilde{\mathbf{w}} \\ \tilde{b} \end{bmatrix}, k \neq 0$$

那么线性流形$\bar{F}$的方程为：

$$\mathbf{x}^T\tilde{\mathbf{w}} + \tilde{b} = 0$$

对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算其到线性流形$\bar{F}$的距离：

$$dist(\mathbf{x}_n, \bar{F}) = \frac{\mathbf{x}_n^T\tilde{\mathbf{w}}+\tilde{b}}{\|\tilde{\mathbf{w}}\|},\ n = 1, \dots, N_m \tag{3-5}$$

$dist(\mathbf{x}_n, \bar{F})$越大，代表分散度和信息量越大，$\mathbf{x}_n$越值得被选择。

2. 代表性度量

然后介绍代表性的度量，用符号R表示。这里与 iRDM 算法的代表性度量一致，使用待优化样本为$\bar{\mathbf{x}}_m$到聚类簇C_m内其他点的平均距离来度量代表性R，即对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算$R(\mathbf{x}_n)$：

$$R(\mathbf{x}_n) = \frac{1}{N_m-1} \times \sum_{\mathbf{x}_i \in C_m} \|\mathbf{x}_n - \mathbf{x}_i\| \tag{3-6}$$

$R(\mathbf{x}_n)$越小，代表代表性越大，$\mathbf{x}_n$越值得被选择。根据k-均值聚类算法的聚类规则，当$\mathbf{x}_n$最接近聚类簇C_m的类中心时，$R(\mathbf{x}_n)$最小，此时代表性最大。而当$\mathbf{x}_n$接近离群点时，其距离类内很多样本都很远，因此$R(\mathbf{x}_n)$较大，因此该项可以限制待优化样本$\bar{\mathbf{x}}_m$选择到接近离群点的样本$\mathbf{x}_n$。

代表性的作用是限制待优化样本接近离群点，因此作为正则项加入优化目标函数即可，这里使用乘法的方式加入。优化目标函数确定了，即对聚类簇C_m中的每一个样本$\mathbf{x}_n$，计算目标函数值$Q(\mathbf{x}_n)$：

$$Q(\mathbf{x}_n) = dist(\mathbf{x}_n, \bar{F}) \times \frac{1}{R(\mathbf{x}_n)} \tag{3-7}$$

3.3.1.3 $M \leq d+1$时，IRD 的算法流程

IRD 算法与 iRDM 算法使用相同的优化框架。

如果$M < d+1$，则使用主成分分析算法将特征空间降维到$M-1$维，然后执行后面的操作。

首先用 RD 算法选择初始的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$，即：使用$k$-均值聚类算法（$k=M$）对样本池进行聚类，聚为$M$个簇$\{C_m\}_{m=1}^{M}$，对每一个簇，选择距离类中心最近的样本作为待打标样本。

然后创建一个空的待打标样本集库P，开始迭代过程。每一轮迭代进行M次优化过程，分别优化待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$每一个样本，每次优化只优化一个样本，其余$M-1$个样本固定。设当前单次优化中优化待打标样本$\bar{\mathbf{x}}_m$，它所在的簇为$C_m$，其余的$M-1$个样本固定，根据公式(3-5)计算这$M-1$个样本固定在特征空间中确定的线性流形$\bar{F}$的方程。对$C_m$中的每一个样本$\mathbf{x}_n$，根据公式(3-5)计算$\mathbf{x}_n$到线性流形$\bar{F}$的距离$dist(\mathbf{x}_n, \bar{F})$（分散度-信息量联合度量）；并根据公式(3-6)计算其代表性度量$R(\mathbf{x}_n)$；最后通过公式(3-7)计算其目标函数值$Q(\mathbf{x}_n)$，并选择簇C_m中目标函数值最大的样本$\mathbf{x}_n^*$，替换原来的样本作为待打标样本$\bar{\mathbf{x}}_m$。

当本轮优化完成后，判断当前的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$是否已经存在于待打标样本集库$P$中，若已经存在则输出当前待打标样本集，算法结束；若不存在则判断是否达到最大迭代轮数，如果达到则输出当前待打标样本集，算法结束，否则将当前的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=1}^{M}$作为一个新元素加入待打标样本集库$P$中，进行下一轮迭代。

$M \leq d+1$时，IRD 算法流程图如图 3-5 所示：

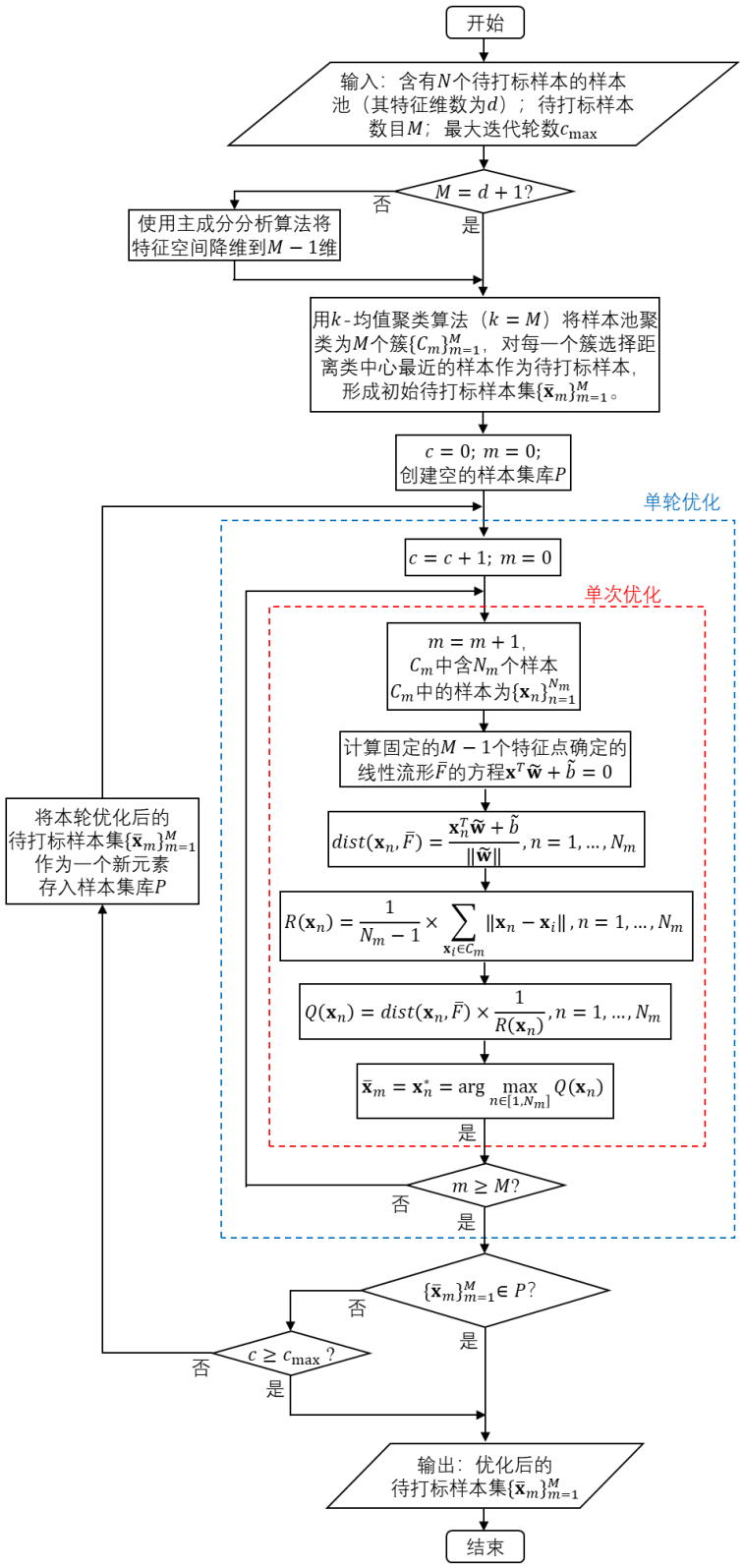

图 3-5 $M \leq d + 1$时，IRD 算法流程图

3.3.2 $M > d+1$时的情形

3.3.2.1 $M > d+1$时 IRD 算法的分段采样方案

这种情形下，IRD 先使用$M = d+1$时的方法选择$d+1$个待打标样本，然后再通过其他无监督主动学习算法选择剩下的$M-d-1$个样本。这样做的合理性在于，使用$M = d+1$时的方法选择$d+1$个待打标样本已经具有良好的性能，后续的样本是对性能的进一步提升。本文使用 iRDM 方法（在第 3.2 节中介绍）来选择剩下的$M-d-1$个待打标样本，这个过程同时考虑分散度和代表性，具体过程如下：

使用第 3.1 节提出的交替优化框架来选择剩下的$M-d-1$个待打标样本。首先对样本池中的非待打标样本进行k-均值聚类（$k = M-d-1$）生成$M-d-1$个簇，对每个簇，选择距离该簇中心最近的样本作为初始的待打标样本集$\{\bar{\mathbf{x}}_m\}_{m=d+2}^{M}$，然后迭代地优化之。在单次优化中，有$M-1$个待打标样本固定（已经选好的$d+1$样本和待优化集合中的$M-d-2$个样本），1 个待打标样本$\bar{\mathbf{x}}_m$待优化，设它所在的簇为$C_m$。使用公式(3-1)计算待优化样本的分散度度量$D$，使用公式(3-2)计算待优化样本的代表性度量$R$，使用公式(3-3)对$C_m$中所有样本计算目标函数值$Q$，选择最大的替换$\bar{\mathbf{x}}_m$成为新的待打标样本。

3.3.2.2 $M > d+1$时 IRD 的算法流程

$M > d+1$时，IRD 算法流程图如图 3-6 所示。

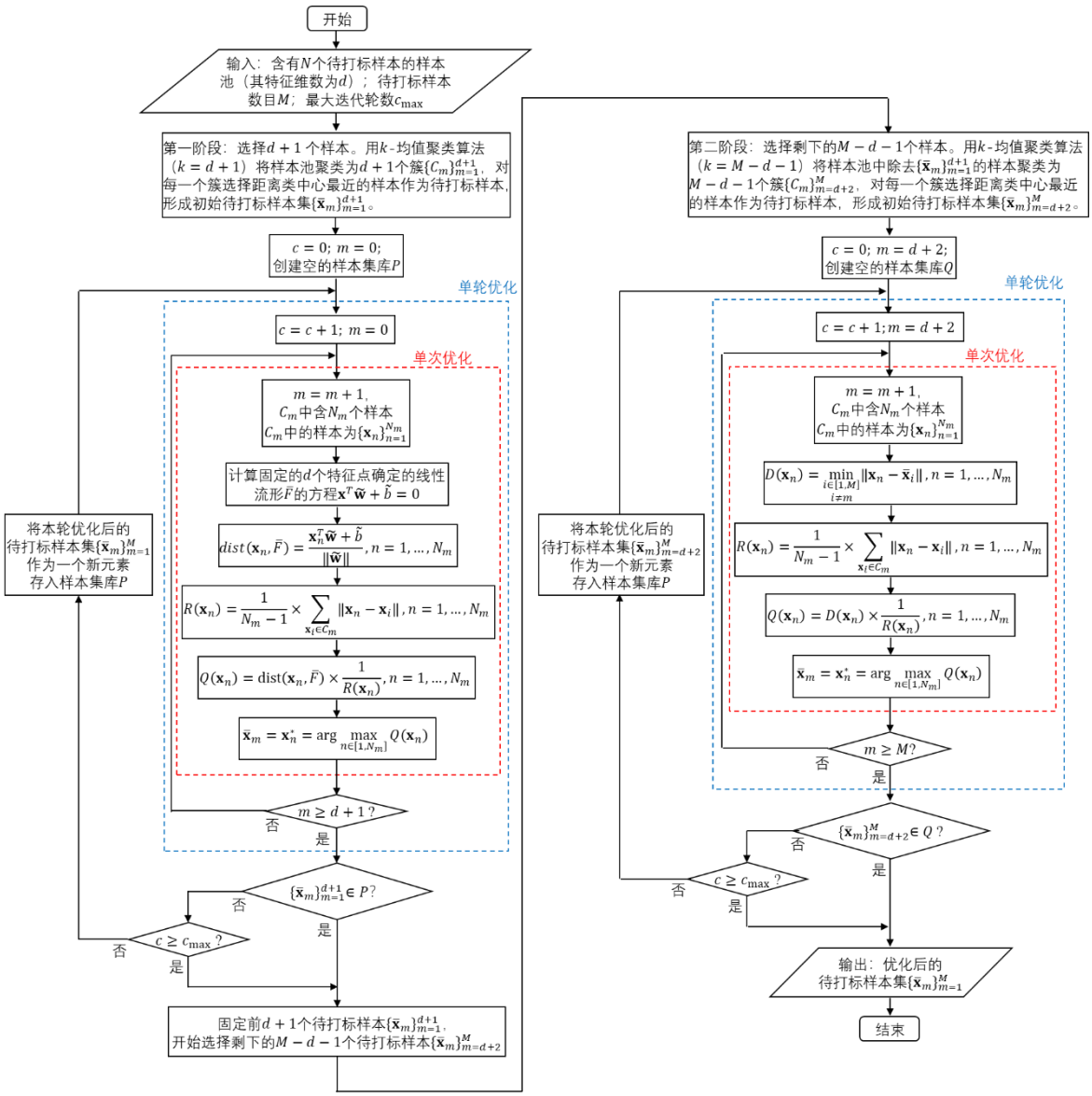

图 3-6 $M>d+1$时，IRD 算法流程图

3.3.3 iRDM 算法与 IRD 算法的比较

iRDM 算法和 IRD 算法使用了相同的交替优化框架，区别在于单次优化时使用的目标函数不同。

iRDM 算法的目标函数中分散度的度量D的计算类似于 GSx 中的贪婪距离，只考虑了分散度一个核心指标；而 IRD 算法中的$dist(\mathbf{x}_n,\bar{F})$项的计算使用的是点到线性流形的距离，拥有基于线性回归模型的理论推导，因此同时融合了分散度和信息量两个核心指标。

iRDM 算法和 IRD 算法的R项相同，均是对代表性指标的度量。

当待打标样本数量M和特征维度d满足$M = d + 1$时，iRDM 在一维特征空间中有理论解释，但是在特征空间为高维时使用的是启发式算法。而 IRD 算法在一维特征空间中与 iRDM 算法一致，在二维和高维特征空间中依然拥有理论解释。

当待打标样本数量M和特征维度d满足$M > d + 1$时，iRDM 算法为启发式算法，IRD 算法需要使用 iRDM 算法来选择剩下的$M - d - 1$个样本。

iRDM 算法和 IRD 算法是无监督主动学习回归算法，均可以同时适用于线性回归和基于核的非线性回归。IRD 算法选择$d + 1$个待打标样本时的理论推导虽是基于线性回归的，但$dist(\mathbf{x}_n, \bar{F})$项与分散度也往往是正相关关系，因此也可以适用于基于核的非线性回归，但效果不一定优于 iRDM（这一点将在后文实验中验证）。

在特征空间为二维时，图 3-7 和图 3-8 分别直观地展示了 iRDM 算法和 IRD 算法的一轮迭代过程（包含 3 个单次优化过程）。

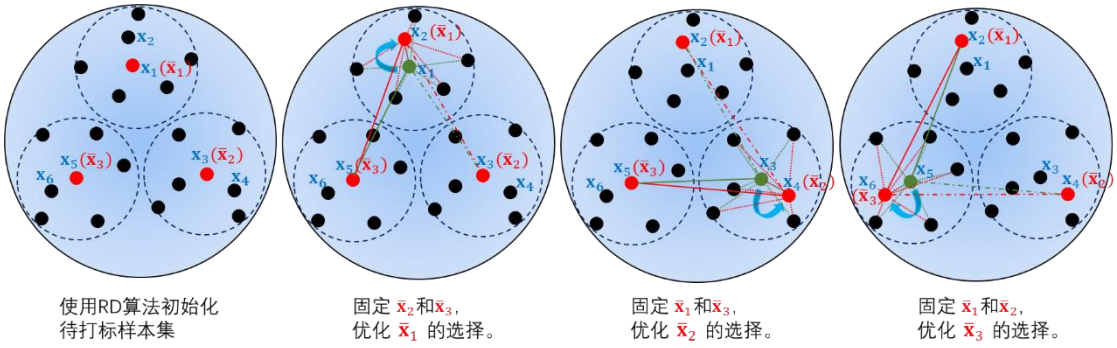

图 3-7 特征空间维数$d = 2$，待打标样本数$M = 3$时，iRDM 算法单轮优化过程示意图

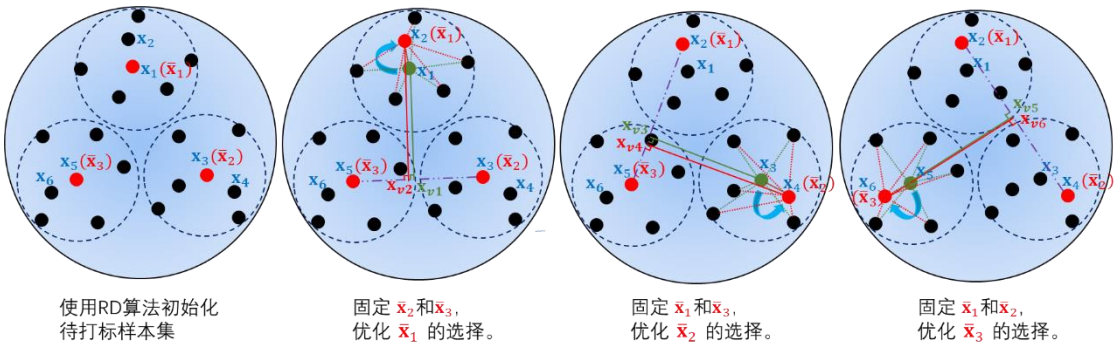

图 3-8 特征空间维数$d = 2$，待打标样本数$M = 3$时，IRD 算法单轮优化过程示意图

3.4 本章小结

提出一种交替优化框架，用于在无监督主动学习回归中优化待打标样本集，可以将多目标优化问题拆分为若干单目标优化问题迭代进行。

提出一种新的无监督主动学习回归算法：通过迭代来最大化分散度和代表性的算法（Iterative Representativeness-Diversity Maximization，简称 iRDM），该算法使用提出的交替优化框架，度量了分散度与代表性指标，并将它们融合到了单次优化的目标函数中。

提出另一种新的无监督主动学习回归算法：同时考虑分散度、代表性和信息量的算法（Diversity, Representativeness and Informativeness，简称 IRD），它与 iRDM 算法的区别是，加入针对线性回归模型的信息量指标。

对比了 iRDM 算法与 IRD 算法，并给出了优化过程的示意图。

4 实验和结果

本章介绍实验部分。该部分实现了第 1.2 章中介绍的八种已有的主动学习回归算法，与第 3 章中提出的两种新的无监督主动学习回归算法（iRDM 和 IRD），在 12 个涵盖多个实际应用领域的公开回归数据集上，使用线性回归模型和基于核的非线性回归模型进行了多次实验，验证了本文提出的两种算法的有效性和稳定性，以及 IRD 算法在线性回归中相对于 iRDM 算法的优越性；同时验证了无监督主动学习回归中分散度、代表性和信息量的作用。

本文的实验是基于 MATLAB2019 软件平台来实现的。

4.1 回归数据集及其预处理

4.1.1 回归数据集

本课题使用 12 个真实的回归数据集来进行实验，这些数据集涵盖很多应用领域。其中，9 个数据集来自 UCI 机器学习数据库（UCI Machine Learning，网址：http://archive.ics.uci.edu/ml/index.php），2 个数据集来自 CMU StatLib Datasets Archive（网址：http://lib.stat.cmu.edu/datasets/），这些数据集在之前的研究[20][22][23][38][39]中曾用到。我们还使用了一个公开的情感计算数据集：德国视听即兴演讲数据库：The Vera am Mittag（德语，简称 VAM。英语是：Vera at Noon）[4]，该数据集曾在多个研究中被用到[17][41][42][43][44][48]。它包含 947 条语音样本，来自于 47 位演讲者。数据集已经进行了声音信号的特征提取，从中提取了 46 个声学特征[43][44]，其中 9 个特征是关于音调的，5 个特征是关于持续时间的，6 个特征是是关于能量的，还有 26 个特征是梅尔倒频谱参数特征（Mel-Frequency Cepstral Coefficients，简称：MFCC）。其情感标签分为三个维度：效价，唤醒度和优势度，本文只使用唤醒度作为标签。

关于数据集的详细信息见表 4-1。

表 4-1 实验中使用的 12 个回归数据集的详细信息

数据集	来源	样本数	原始特征数	数值类型特征数	类别类型特征数	独热编码后的特征数	降维后的特征数
Concrete-CS	UCI	103	7	7	0	7	7
Yacht	UCI	308	6	6	0	6	6
autoMPG	UCI	392	7	6	1	9	7
NO2	StatLib	500	7	7	0	7	7
Housing	UCI	506	13	13	0	13	13
CPS	StatLib	534	10	7	3	19	10
EE-Cooling	UCI	768	7	7	0	7	7
VAM-Arousal	ICME	947	46	46	0	46	10
Concrete	UCI	1030	8	8	0	8	8
Airfoil	UCI	1503	5	5	0	5	5
Wine-red	UCI	1599	11	11	0	11	11
Wine-white	UCI	4898	11	11	0	11	11

4.1.2 数据预处理

首先划分训练集和测试集。对每个数据集，均随机抽取 50%的样本作为训练集（样本池），将另外 50%的样本当作测试集。本实验不划分验证集，其一是因为训练集是不带任何标签的样本池，因此无法划分验证集；其二是因为本文的实验中没有进行超参数调节，因此也不需要验证集。然后按如下步骤进行数据预处理：

1. 类别特征转化为数值特征。

实验用到的 12 个数据集中，autoMPG 和 CPS 含有类别特征，类别特征的各个类别取值是没有大小顺序的。本实验使用独热编码将类别特征转化为数值特征，独热编码是一种简单且应用广泛的特征预处理方法。例如 autoMPG 数据集含有一个类别特征“来源”，其三个分类取值为：“美国”、“日本”和“德国”，我们将“美国”编码为：$[1,0,0]$，将“日本”编码为：$[0,1,0]$，将“德国”编码为：$[0,0,1]$，于是便使用了三个数值特征来替代这一个类别特征。

2. 特征归一化。

同一数据集的不同特征的量纲不同，因此数值差异可能较大，但数值小的特征往往可能对结果有很大的影响，因此应该使用特征归一化，去除各特征的量纲对结果的影响。

对每个数据集的训练集，我们将各个特征的分布归一化为均值为 0，标准差为 1 的分布，具体操作过程为：求取每个特征的均值μ_{train}和标准差σ_{train}，对训练集中的每个样本在该特征上的取值x_{train}减去该特征的均值μ_{train}再除以标准差σ_{train}，得到归一化后的特征取值x'_{train}，用公式表示为：

$$x'_{\text{train}} = \frac{x_{\text{train}} - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

使用的函数是 MATLAB2019 的 zscore 函数。

对每个数据集的测试集，由于在实际应用中我们并不知道测试集的分布，因此不能够使用测试集的均值和标准差来对测试集进行归一化，本文中我们使用训练集的均值和标准差作为测试集均值和标准差的估计值，用于归一化测试集，具体操作过程为：对测试集中的每个样本在该特征上的取值x_{test}减去训练集的均值μ_{train}再除以训练集的标准差σ_{train}，得到归一化后的特征取值x'_{test}，用公式表示为：

$$x'_{\text{test}} = \frac{x_{\text{test}} - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

3. 特征降维。

对于维度较高的数据集，往往需要更多训练样本才能训练出可靠的回归模型，然而需要用到主动学习的回归问题往往因为打标难度大而只能使用较少的训练样本，难以适应高维数据集，因此实验中使用主成分分析算法[45]对高维数据集进行降维，具体的做法如下：

1）对 VAM-Arousal 数据集，使用主成分分析将其从 46 维降低到 10 维。

2）对 CPS 和 autoMPG 这两个含有类别特征的数据集，由于在进行独热编码之后，其特征维度提升，我们也使用主成分分析将其降维到其原始的特征维度数，例如 CPS 数据集原始特征维度数是 10，进行独热编码之后的特征维度是 19，我们使用主成分分析将其从 19 维降低到 10 维。

对上述三个数据集的训练集可以直接使用主成分分析算法降维，但对于它们的测试集，由于其分布未知，不能直接使用主成分分析算法。本实验使用训练集的投影系数矩阵来对测试集进行降维，从而确保没有数据泄露。

4.2 对比的算法

4.2.1 基线：随机采样

即随机从样本池中选择指定数量的样本来打标。任何主动学习回归算法的效果都不应该低于基线，如果低于基线则算法对该数据集无效。但是在某些特殊情况下，主动学习回归算法的性能可能会差于随机采样，因此研究主动学习时一定要和随机采样这个基线进行对比，效果差于随机采样的主动学习回归算法的性能排名是无效的。

4.2.2 有监督主动学习回归算法：

基于委员会分歧的有监督主动学习回归算法（QBC）：随机重采样形成 4 组副本。同时适用于线性回归和非线性回归。

基于模型变化的期望最大化的有监督主动学习回归算法（EMCM）：随机重采样形成 4 组副本。仅适用于线性回归。

基于残差的主动学习回归方法（RSAL）:无超参数。仅适用于基于核函数的非线性回归。

将 RD 算法与 EMCM 算法结合的主动学习回归算法（RD-EMCM）：随机重采样形成 4 组副本。仅适用于线性回归。

改进的基于贪婪采样的有监督主动学习回归（iGS）：无超参数。仅适用于线性回归。

4.2.3 无监督主动学习回归算法：

基于泛化误差条件期望的重要性加权最小二乘的无监督主动学习回归算法（P-ALICE）：参数λ的寻优范围是：$\{0, .1, .2, .3, .4, .41, .42, \ldots, .59, .6, .7, .8, .9, 1\}$。仅适用于线性回归。

基于贪婪采样的无监督主动学习回归算法（GSx）：无超参数。同时适用于线性回归和非线性回归。

同时考虑待打标样本的分散度和代表性的无监督主动学习回归算法（RD）：无超参数。同时适用于线性回归和非线性回归。

通过迭代来最大化分散度和代表性的无监督主动学习回归算法（iRDM）：最大迭代轮数$c_{max}=5$。同时适用于线性回归和非线性回归。

同时考虑分散度、代表性和信息量的无监督主动学习回归算法（IRD）：最大迭代轮数$c_{max}=5$。同时适用于线性回归和非线性回归。

4.3 使用的回归模型

将主动学习回归算法用于不同的回归模型中时，其性能可能会不同，部分主动学习回归算法仅适用于某种类型的回归模型（例如 P-ALICE 算法只适用于线性回归模型）。目前大多数已有文献只考虑了线性回归模型，本文不仅考虑了线性回归模型，还会考虑基于核函数的非线性回归模型。

实验中使用的线性回归模型为岭回归（Ridge Regression，简称 Ridge）[49][50][51]。岭回归是应用十分广泛的线性回归模型，它在最小二乘回归的基础上加上 L2 正则化项来减小模型方差，非常适合主动学习回归中训练样本很少的情况。本实验中岭回归使用三个不同的 L2 正则化系数$r=[0.01,0.1,1]$。代码实现时所使用的函数是 MATLAB2019 的库函数“ridge”。

实验中使用的非线性回归模型是基于径向基核函数（Radial basis function，简称：RBF）的支持向量机回归（RBF Support Vector Machine，简称：RBF SVR）[52]。使用三种框约束系数（Box constraint）$C=[500,50,5]$（该系数等效于 RBF SVR 中的正则化系数的倒数，C越小对应的正则化系数越大），使用的 RBF 核参数是$\lambda=0.01$（RBF 核公式：$k(\mathbf{x}_i,\mathbf{x}_j)=e^{-\lambda\|\mathbf{x}_i-\mathbf{x}_j\|^2}$），不敏感带的宽度的一半$\varepsilon=0.1\times\mathrm{std}(\mathbf{y}_{\mathrm{train}})$，$\mathbf{y}_{\mathrm{train}}$表示训练集的真实标签组成的向量，$\mathrm{std}(\mathbf{y}_{\mathrm{train}})$表示其标准差。代码实现时所使用的函数是 MATLAB2019 的库函数“fitrsvm”。

4.4 主动学习回归算法性能的评价方式及评价指标

4.4.1 评价方式

评价主动学习回归算法性能的常用方法有两种：

第一种：比较不同的主动学习回归算法在采集相同数量的待打标样本时（即训练回归模型的训练集的大小相同时），打标后训练出的回归模型的性能好坏。训练出的回归模型越好，则代表在当前数量待打标样本的条件下，该主动学习回归算法的性能越好。

第二种：比较在训练出相同性能的回归模型的情况下，使用不同主动学习回归算法采样所需要的样本数量。所需要的样本数量越少，则该主动学习回归算法性能越好。

本文采用第一种方法，因为第一种方法中“相同数量的待打标样本”控制起来很容易，而第二种方法中“相同性能的回归模型”控制起来较为困难，其一是因为实际中往往只能通过调整待打标样本数目实现“性能相近”，而很难实现“性能相同”，其二是因为回归模型的评价指标也会影响其性能评价。

4.4.2 评价指标

实验中使用的回归模型的评价指标是均方误差（Root Mean Squared Error，简称 RMSE）和相关系数（Correlation Coefficient，简称 CC）。本章将依据 100 次重复实验的结果，通过均方误差和相关系数的“均值”对各主动学习回归算法的性能进行排序，通过均方误差和相关系数的“标准差”对各主动学习回归算法的稳定性进行排序，并进行统计测试，得出更具有说服力的结论。

4.5 实验结果

本实验实现的十种采样方法中（九种主动学习回归算法加上基线随机采样（RS）），适用于 Ridge 的有：RS、QBC、EMCM、iGS、RD-EMCM、P-ALICE、GSx、RD、iRDM 和 IRD 九种算法；适用于 RBF SVR 的有 RS、QBC、RSAL、GSx、RD、iRDM 和 IRD 七种算法。

4.5.1 各主动学习回归算法在使用不同数量的训练样本（不同M）时的性能对比

各算法在 12 个数据集上，在使用不同数量的训练样本（不同M）时，在 100 次实验中的平均 RMSE 和平均 CC 如图 4-1（Ridge）和图 4-2（RBF SVR）所示：

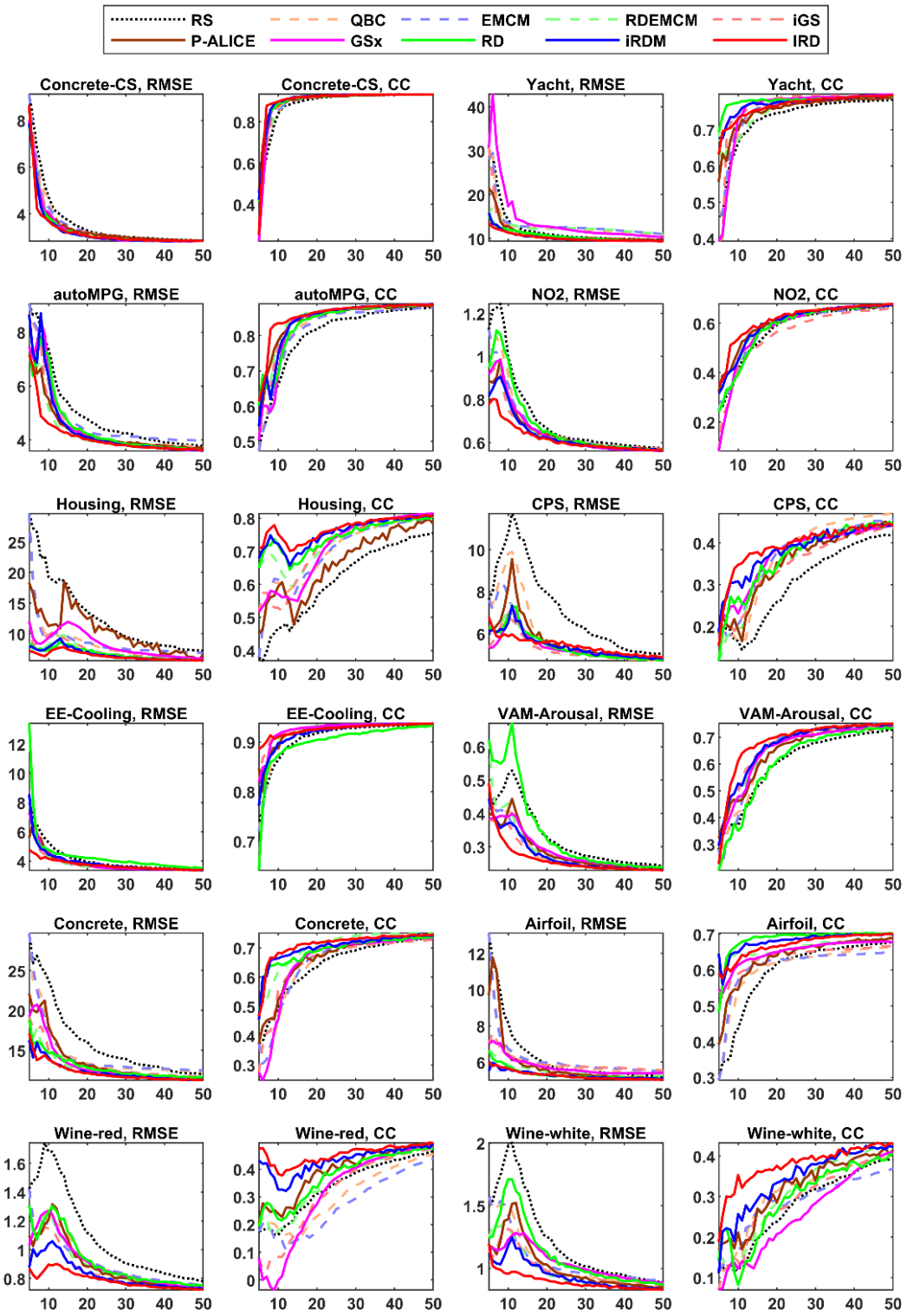

图 4-1 在 12 个数据集上的平均 RMSE 和平均 CC，Ridge（$r = 0.1$）

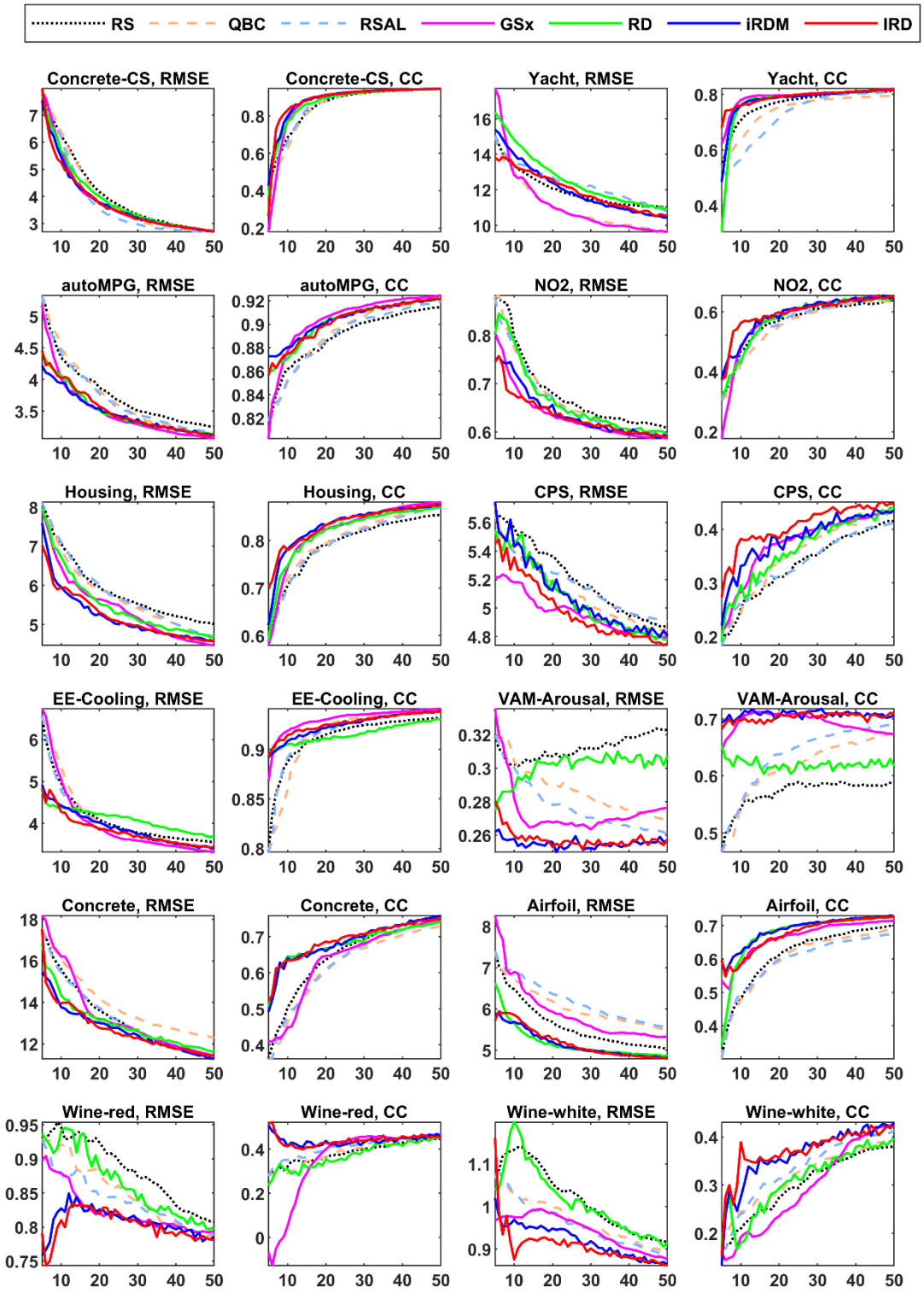

图 4-2 在 12 个数据集上的平均 RMSE 和平均 CC，RBF SVR（$C = 50$）

由图 4-1 和图 4-2 得知：

无论对于 Ridge 还是 RBF SVR，随着M增大，算法性能会有波动，这是因为训练样本很少导致很多不确定性。但对绝大多数数据集，随着 M 增大，各算法对应的平均 RMSE 呈减小趋势，平均 CC 呈增大趋势。这符合预期，因为相同条件下训练样本数量越多，训练出的回归模型性能一般会更好。

对绝大多数数据集，对绝大多数M，无论是考量平均 RMSE 还是平均 CC，八种用于 Ridge 的主动学习回归算法的性能均优于随机采样，六种用于 RBF SVR 的主动学习回归算法的性能均优于随机采样或与随机采样相当。

对绝大多数数据集，对绝大多数M，本文提出的 iRDM 算法和 IRD 算法性能优于大多数其他主动学习回归算法。对 Ridge，IRD 算法的性能优于 iRDM 算法；对 RBF SVR，两者性能不相上下。

4.5.2 各主动学习回归算法在不同数据集上综合性能的对比

由于在同一数据集上，不同主动学习回归算法的性能排序在不同M处可能不同，为了考察不同主动学习回归算法对某一特定数据集的整体性能，本实验计算各算法在$M \in [5,20]$时对应的曲线下面积（Aera Under the Curve，简称 AUC 值），作为该算法在该数据集上的整体表现。选择$M \in [5,20]$是因为主动学习回归算法在训练样本较少时相对于随机采样性能提升较大，此时更能体现主动学习回归算法的价值。

在 12 个数据集上，各主动学习回归算法的平均 RMSE 和平均 CC 的 AUC 值相对于基线随机采样的比值见图 4-3（Ridge）和图 4-4（RBF Ridge）。由于不同数据集的结果差异较大，图中所有结果都除以随机采样的结果，因此随机采样的结果总是“1”。

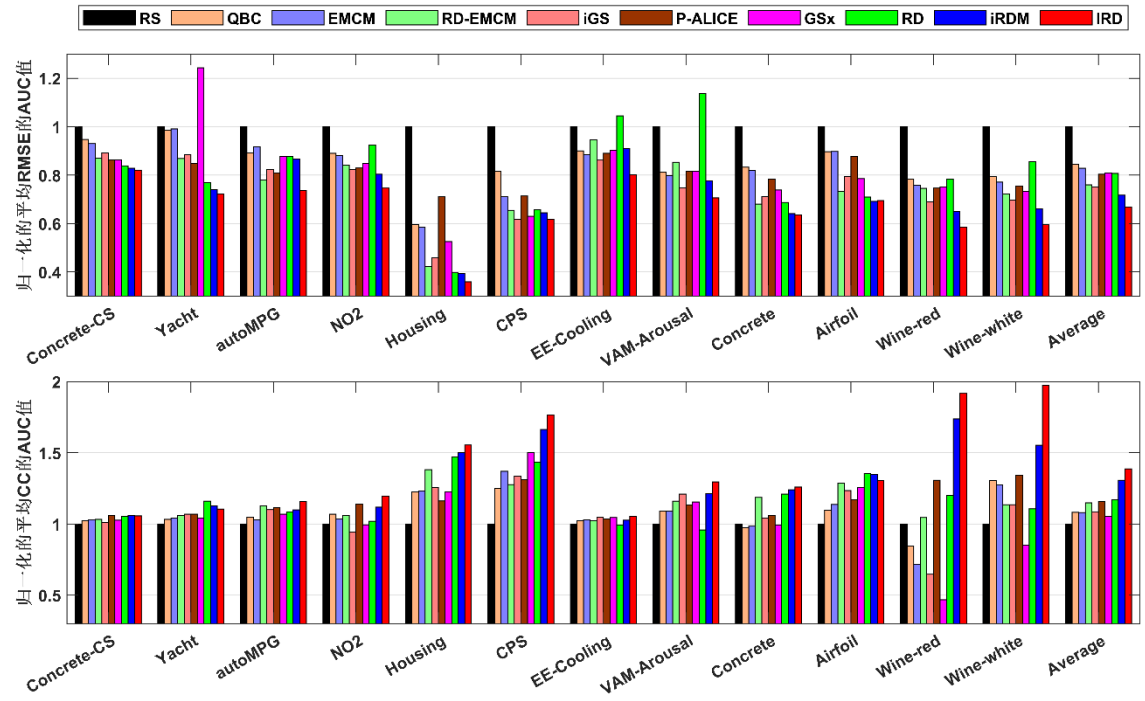

图 4-3 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，Ridge（$r = 0.1$）

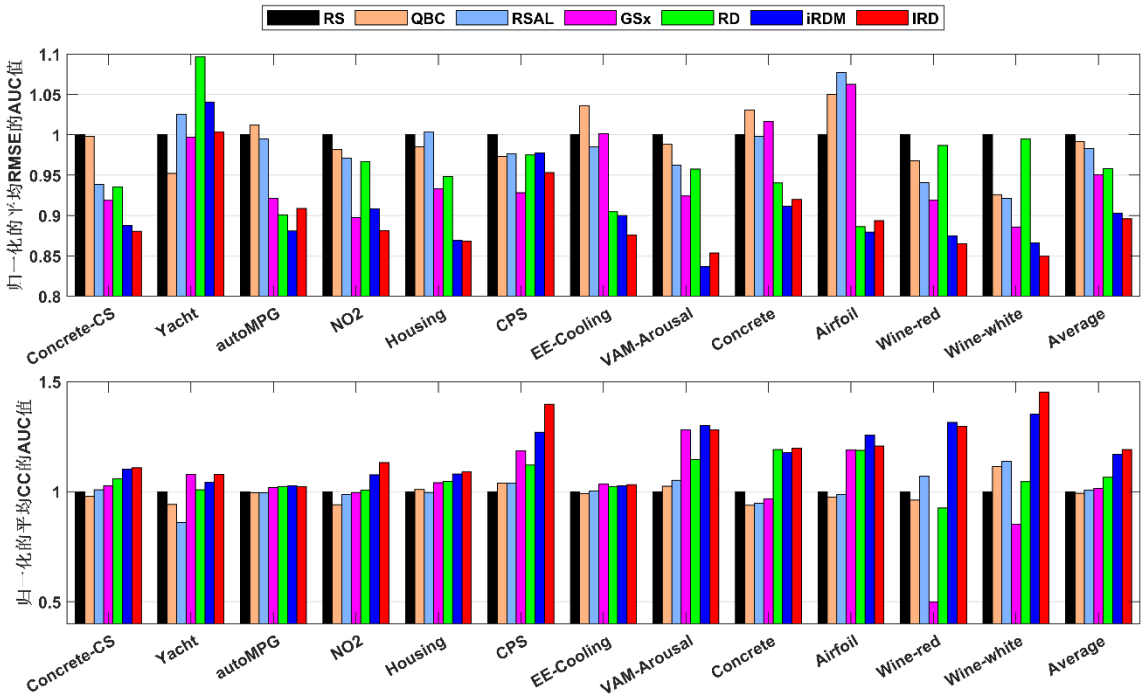

图 4-4 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，RBF SVR（$C = 50$）

在图 4-3 和图 4-4 中 12 个数据集的平均结果的基础上，计算各主动学习回归算法的性能相对于随机采样的提升百分比（平均 RMSE 的 AUC 值相对于随机采样的减小百分比和平均 CC 的 AUC 值相对于随机采样的增大百分比），见表 4-2 和表 4-3。

表 4-2 各算法性能相对于随机采样的提升百分比，12 个数据集取平均，Ridge（$r = 0.1$）

	QBC	EMCM	RD-EMCM	iGS	P-ALICE	GSx	RD	iRDM	IRD
RMSE	15.45	17.16	24.04	24.96	19.65	19.05	19.35	28.28	33.20
CC	8.18	8.08	14.78	8.56	15.88	5.20	17.01	30.69	38.65

表 4-3 各算法性能相对于随机采样的提升百分比，12 个数据集取平均，RBF SVR（$C = 50$）

	QBC	RSAL	GSx	RD	iRDM	IRD
RMSE	0.83	1.72	4.96	4.23	9.74	10.41
CC	-0.78	0.67	1.41	6.50	16.87	19.11

由图 4-3，图 4-4，表 4-2 和表 4-3 得知，对绝大多数数据集，以及 12 个数据集的平均结果，无论是考量平均 RMSE 的 AUC 值还是平均 CC 的 AUC 值，对 Ridge，八种主动学习回归算法的性能均优于随机采样，本文提出的 iRDM 算法和 IRD 算法性能优于其他主动学习回归算法；IRD 算法优于 iRDM 算法；对 RBF SVR，两种有监督主动学习回归算法的性能与随机采样相当，四种无监督主动学习回归算法的性能均优于随机采样；本文提出的 iRDM 算法和 IRD 算法性能优于其他主动学习回归算法；IRD 算法的性能与 iRDM 算法相当，仅略有提升。

两种有监督主动学习回归算法表现不好的原因可能是训练样本过少（$M \in [5,20]$），少量的标签可能会误导采样过程，这也间接证实了第 2.1 节中提到的无监督主动学习回归算法的优势：在训练样本很少时，采样过程不会受到已有少量标签的误差的干扰。

4.5.3 各算法对回归模型超参数变化的鲁棒性

一个好的主动学习回归算法需要对回归模型的超参数具有鲁棒性，对 Ridge，本实验考察 L2 正则化系数r；对 RBF SVR，本实验考察框约束系数C（C也可以看作正则化系数，C减小时，等效的正则化系数增加）。由于训练样本较少，RBF SVR 中 RBF 核的参数λ应该设置一个较小的数值，否则容易过拟合，因此这里直接设置$\lambda = 0.01$而不使用其他数值。

对 Ridge，$r=[0.01,0.1,1]$时，八种主动学习回归算法在 12 个数据集上对应的平均 RMSE 的 AUC 值和平均 CC 的 AUC 值相对于随机采样的比值分别见图 4-5，图 4-3 和图 4-6；对 RBF SVR，$C=[500,50,5]$时，六种主动学习回归算法在 12 个数据集上对应的平均 RMSE 的 AUC 值和平均 CC 的 AUC 值相对于随机采样的比值分别见图 4-7，图 4-4 和图 4-8。

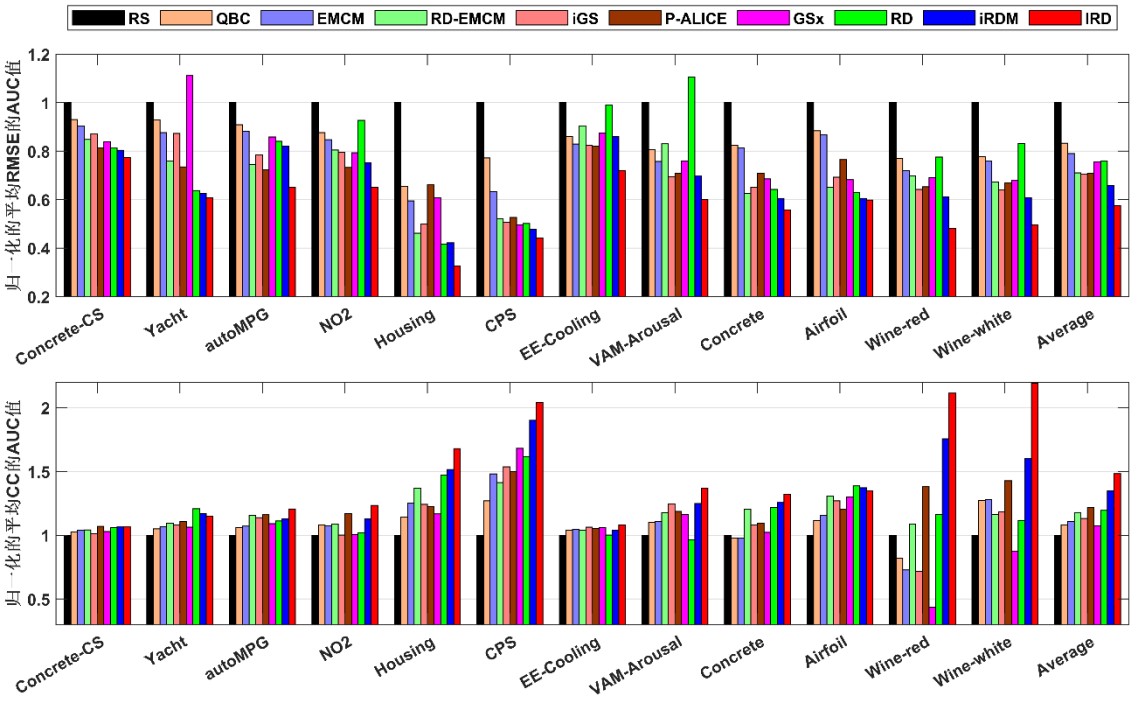

图 4-5 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，Ridge（$r=0.01$）

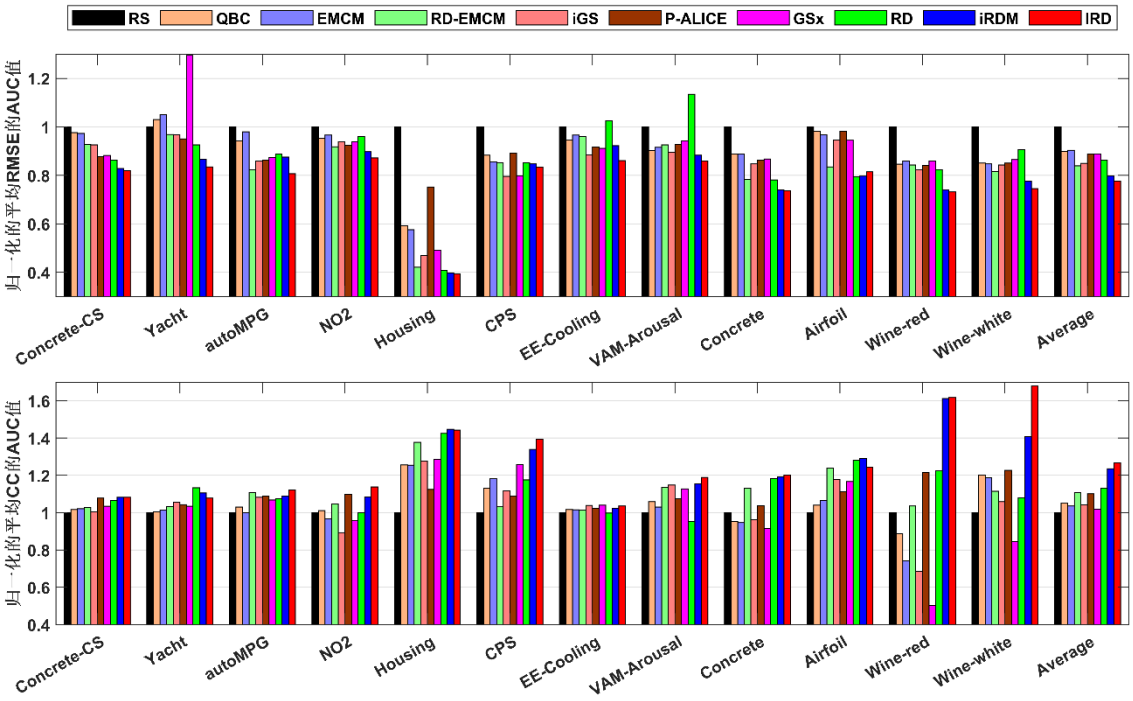

图 4-6 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，Ridge（$r = 1$）

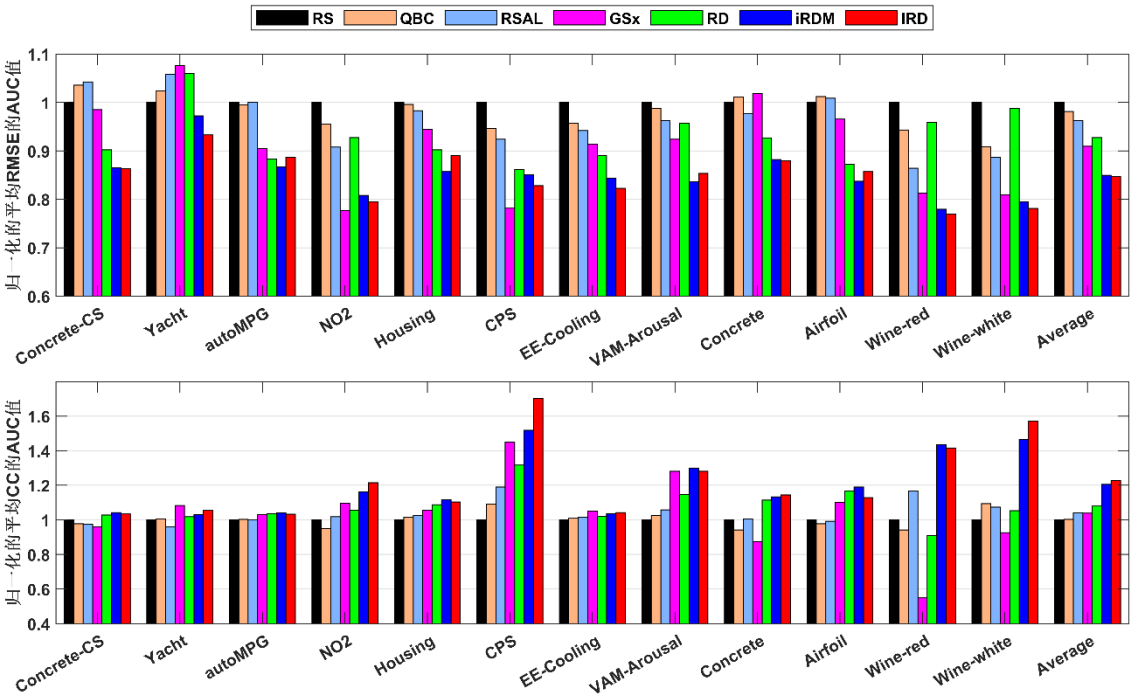

图 4-7 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，RBF SVR（$C = 500$）

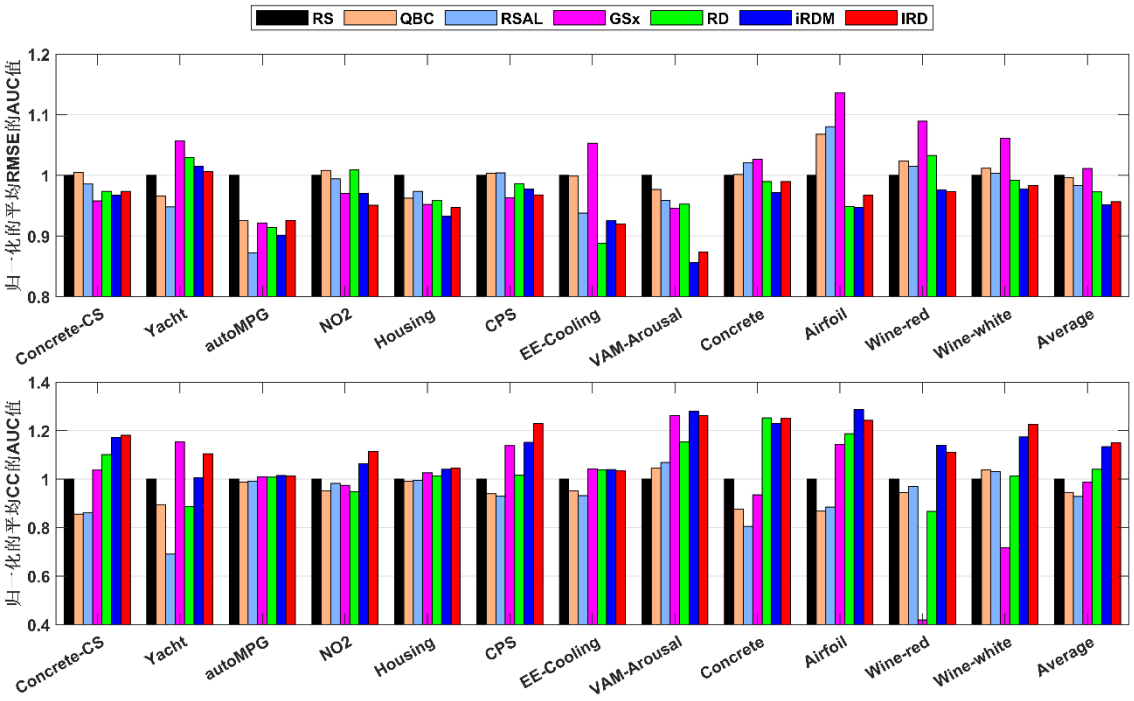

图 4-8 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，RBF SVR（$C = 5$）

由图 4-5、图 4-6、图 4-7 和图 4-8 得知，无论对 Ridge 中不同的r还是对 RBF SVR 中不同的C，无论是考量平均 RMSE 的 AUC 值还是平均 CC 的 AUC 值，算法性能排序基本不变，本文提出的 iRDM 算法和 IRD 算法性能依然优于其他算法。此外还能得到一个新结论：随着正则化系数的增加（r的增加和C的减小），各主动学习回归算法的性能相对于随机采样的提升变小，这符合预期，因为正则化系数增加时回归模型方差减小，基线“随机采样”的性能提升。

4.5.4 各算法对样本池分布的扰动的鲁棒性

由于在实际应用中，样本池的分布不会一成不变，因此好的主动学习回归算法需要对样本池分布的扰动具有一定的鲁棒性。本实验 100 次重复结果中，每次都随机从数据集中选择 50%的样本当作样本池，因此每次实验使用的样本池的分布都不相同。通过考量算法 100 次实验结果中算法性能的标准差，即可评估算法对于样本池分布扰动的鲁棒性。

在 12 个数据集上，在 Ridge 中使用的八种主动学习回归算法和在 RBF SVR 中使用的六种主动学习回归算法，在 100 次实验中对应的 RMSE 的标准差的 AUC 值和 CC 的标准差的 AUC 值相对于随机采样的比值分别如图 4-9（Ridge）和图 4-10（RBF SVR）所示：

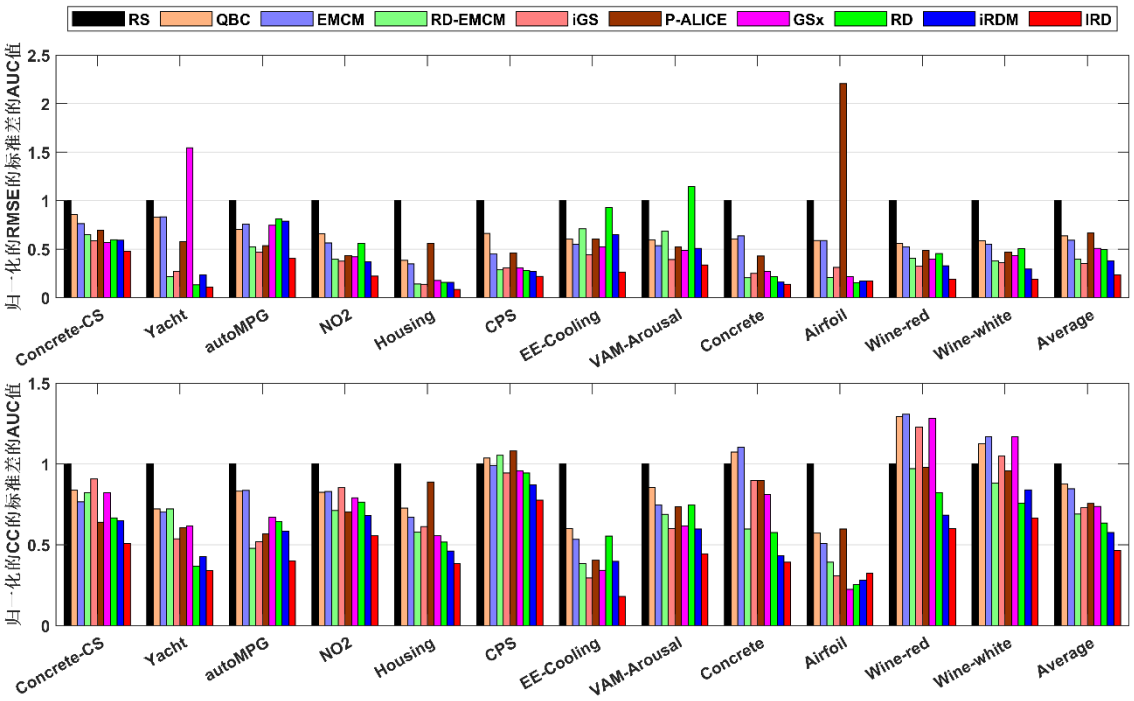

图 4-9 在 12 个数据集上的归一化的 RMSE 和 CC 的标准差的 AUC 值，Ridge（$r = 0.1$）

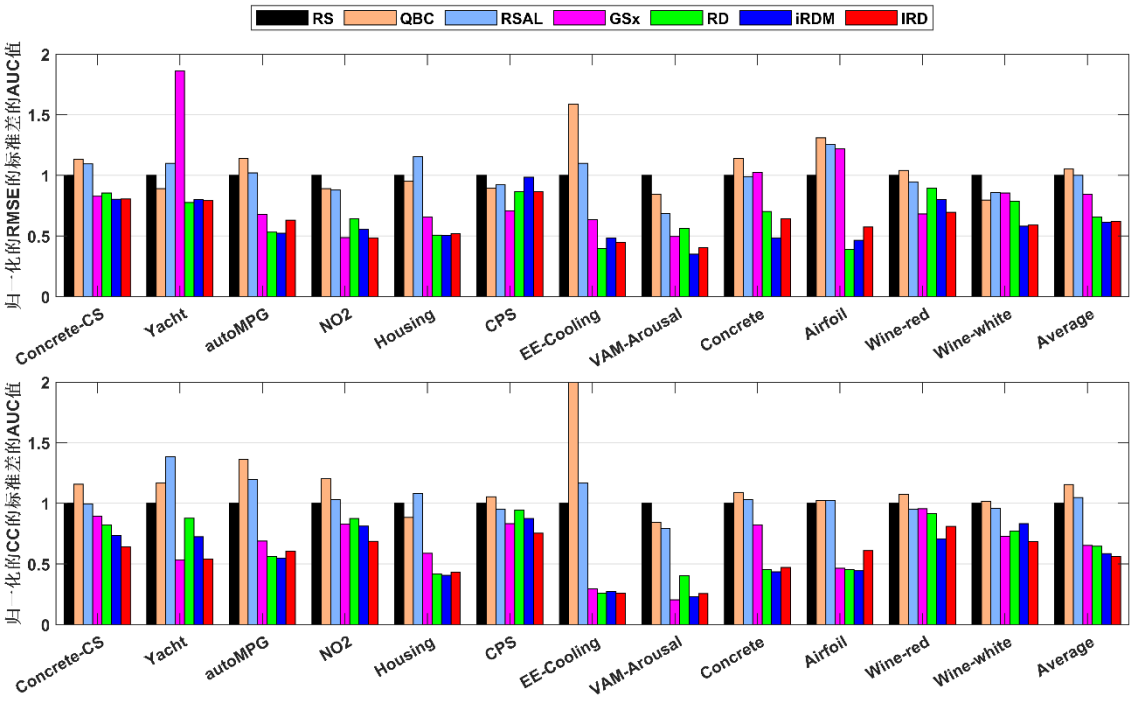

图 4-10 在 12 个数据集上的归一化的 RMSE 和 CC 的标准差的 AUC 值，RBF SVR（$C = 50$）

在图 4-9 和图 4-10 中 12 个数据集的平均结果的基础上，计算各算法对应的 RMSE 的标准差和 CC 的标准差的 AUC 值相对于随机采样的减小百分比，见表 4-4（Ridge）和表 4-5（RBF SVR）。

表 4-4 标准差相对于随机采样的减小百分比，12 个数据集取平均，Ridge（$r = 0.1$）

	QBC	EMCM	RD-EMCM	iGS	P-ALICE	GSx	RD	iRDM	IRD
RMSE	36.42	40.89	60.03	64.86	33.48	49.24	50.57	62.36	76.64
CC	12.53	15.37	31.02	27.13	24.55	26.23	36.59	42.49	53.60

表 4-5 标准差相对于随机采样的减小百分比，12 个数据集取平均，RBF SVR（$C = 50$）

	QBC	RSAL	GSx	RD	iRDM	IRD
RMSE	-5.15	-0.05	15.63	34.16	38.92	37.89
CC	-15.64	-4.71	34.78	35.44	41.48	43.77

由图 4-9、图 4-10、表 4-4 和表 4-5 得知，对绝大多数数据集，以及 12 个数据集的平均结果，无论是考量 RMSE 的标准差的 AUC 值还是 CC 的标准差的 AUC 值，对 Ridge，八种主动学习回归算法的稳定性均优于随机采样，本文提出的 iRDM 算法和 IRD 算法的稳定性均优于其他主动学习回归算法，IRD 算法的稳定性优于 iRDM 算

法；对 RBF SVR，两种有监督主动学习回归算法的稳定性略低于随机采样，四种无监督主动学习回归算法的稳定性优于随机采样，本文提出的 iRDM 算法和 IRD 算法的稳定性均优于其他主动学习回归算法，IRD 算法的稳定性与 iRDM 算法相当，在 CC 上略有提升。

4.5.5 各算法性能差异的统计测试

为评估本文提出的 iRDM 算法和 IRD 算法相对于其他现有算法性能提升是否具有统计显著性，以及 IRD 算法相对于 iRDM 算法的性能提升是否具有统计显著性，本实验还进行了统计测试。本实验使用的统计测试为 Dunn 检验[46]，这种方法非常适合于对多算法在多个数据集上多次实验的结果进行统计检验，它将在 12 个数据集上对各算法的 RMSE 的 AUC 值和 CC 的 AUC 值进行非参数的多重比较检验，并使用错误发现率（False Discovery Rate）的方法[47]进行p值的矫正。关于 Ridge 的结果见表 4-6，关于 RBF SVR 的结果见表 4-7，其中具有统计显著性差异的结果用粗体标出。假设检验的原假设H_0是：表行索引中的算法与列索引中的算法性能没有显著差异，备择假设H_1是：表行索引中的算法与列索引中的算法性能具有显著差异。

表 4-6 非参数多重检验的p值（阈值$\alpha = 0.05$，如果$p < \alpha/2$则拒绝H_0），Ridge（$r = 0.1$）

		RS	QBC	EMCM	RD-EMCM	iGS	P-ALICE	GSx	RD	iRDM
iRDM	RMSE	**.0000**	**.0000**	**.0000**	.0885	.1067	**.0004**	**.0034**	**.0051**	-
	CC	**.0000**	**.0000**	**.0000**	**.0002**	**.0000**	**.0000**	**.0000**	**.0008**	-
IRD	RMSE	**.0000**	**.0000**	**.0000**	**.0001**	**.0002**	**.0000**	**.0000**	**.0000**	**.0118**
	CC	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0152**

表 4-7 非参数多重检验的p值（阈值$\alpha = 0.05$，如果$p < \alpha/2$则拒绝H_0），RBF SVR（$C = 50$）

		RS	QBC	RSAL	GSx	RD	iRDM
iRDM	RMSE	**.0000**	**.0001**	**.0002**	**.0162**	**.0151**	-
	CC	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	-
IRD	RMSE	**.0000**	**.0000**	**.0001**	**.0076**	**.0072**	.3867
	CC	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	.2222

由表 4-6 和表 4-7 得知：对 Ridge，除了 iRDM 算法在 RMSE 指标上与两种有监督主动学习回归算法 RD-EMCM 和 iGS 无显著性差异外，无论考量 RMSE 还是 CC 指标，

iRDM 和 IRD 算法相对随机采样和其他八种现有算法的性能提升均具有显著性；无论考量 RMSE 还是 CC 指标，IRD 算法对 iRDM 算法的性能提升具有显著性。对 RBF SVR，无论考量 RMSE 还是 CC 指标，iRDM 和 IRD 算法相对随机采样和其他六种现有算法的性能提升均具有显著性；IRD 算法与 iRDM 算法的性能则没有显著性差异。

4.5.6 无监督主动学习回归算法中分散度、代表性和信息量三个核心指标的作用

1. iRDM 算法中分散度和代表性的作用

iRDM 算法使用了分散度和代表性两个指标。首先验证分散度的作用，若将 iRDM 算法中单次优化的目标函数中的$D(\mathbf{x}_n)$项去掉，则其目标函数为：

$$Q(\mathbf{x}_n) = \frac{1}{R(\mathbf{x}_n)}$$

根据$R(\mathbf{x}_n)$的计算公式，此时算法将选择每个簇中距离中心最近的点，与 RD 算法等价，因此对比 RD 算法即可。然后验证代表性的作用，构建如下对照算法 iDM：即将 iRDM 算法中单次优化的目标函数中的$R(\mathbf{x}_n)$项去掉，即目标函数为：

$$Q(\mathbf{x}_n) = D(\mathbf{x}_n)$$

因此实验将对比 RD、iDM 和 iRDM 三种算法，与之前的实验设置相同，100 次实验中，在 12 个数据集上，这三种算法的平均 RMSE 的 AUC 值和平均 CC 的 AUC 值相对于随机采样的比值如图 4-11（Ridge）和图 4-12（RBF SVR）所示：

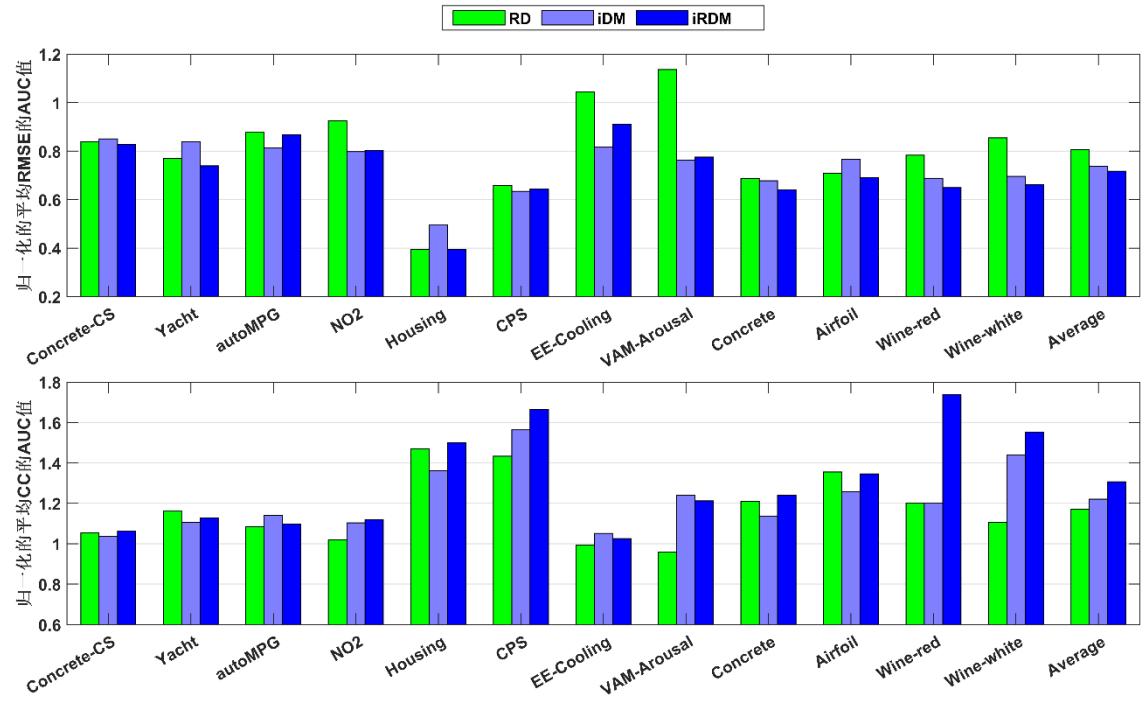

图 4-11 iRDM 算法中分散度和代表性的作用，Ridge（$r = 0.1$）

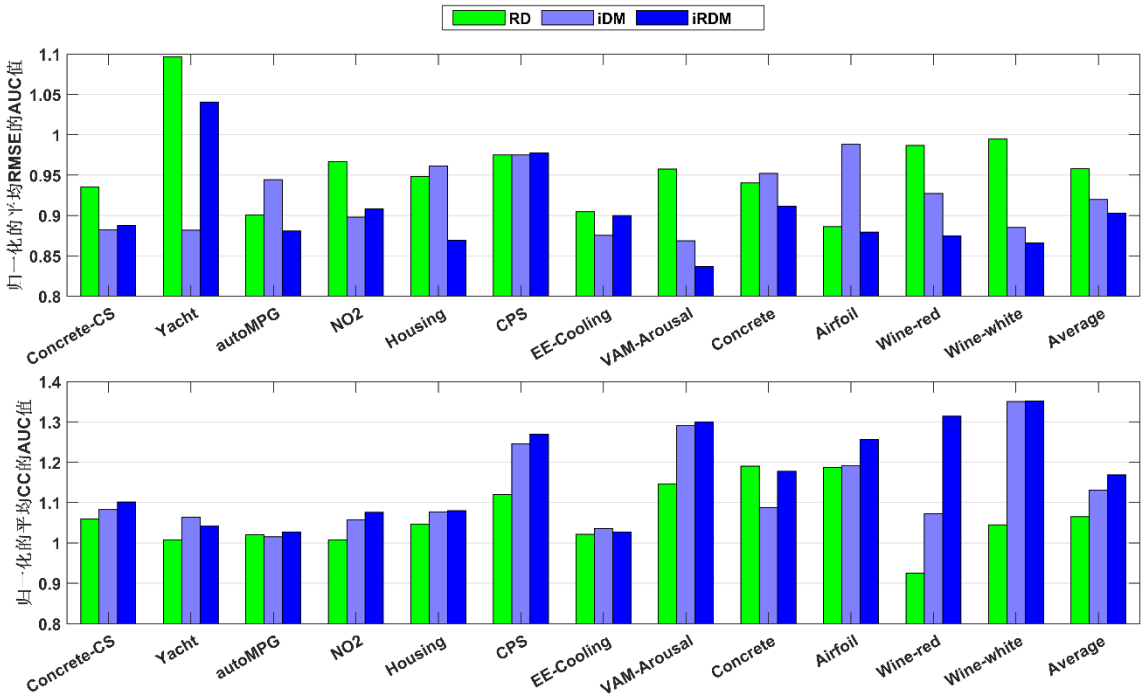

图 4-12 iRDM 算法中分散度和代表性的作用，RBF SVR（$C = 50$）

由图 4-11 和图 4-12 得知，无论对 Ridge 还是 RBF SVR，对绝大多数数据集，以及 12 个数据集的平均结果，无论是考察平均 RMSE 的 AUC 值还是平均 CC 的 AUC 值，三种算法的性能排序是：iRDM>iDM>RD，这说明分散度和代表性在 iRDM 算法中均发挥了作用。

2. IRD 算法中分散度、代表性和信息量的作用

IRD 算法使用了分散度-信息量联合指标和代表性指标。首先验证分散度的作用，与 RD 算法对比即可；然后验证信息量的作用，对比 iRDM 算法即可；最后验证代表性的作用，构建如下对照算法 ID：即将 IRD 算法中单次优化的目标函数中的 R 项去掉，即目标函数为：

$$Q(\mathbf{x}_n) = dist(\mathbf{x}_n, \bar{F})$$

因此实验将对比 RD、ID、iRDM 和 IRD 四种算法，与之前的实验设置相同，100 次实验中，在 12 个数据集上，这四种算法的平均 RMSE 的 AUC 值和平均 CC 的 AUC 值相对于随机采样的比值如图 4-13（Ridge）和图 4-14（RBF SVR）所示：

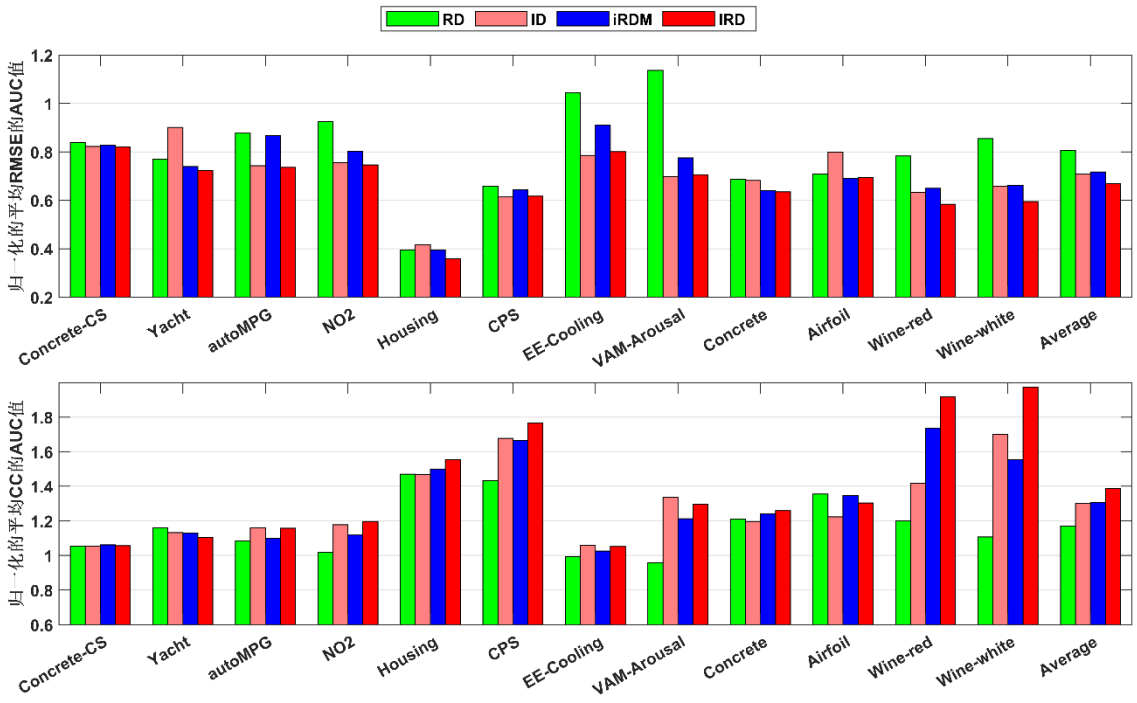

图 4-13 IRD 算法中分散度、代表性和信息量的作用，Ridge（$r = 0.1$）

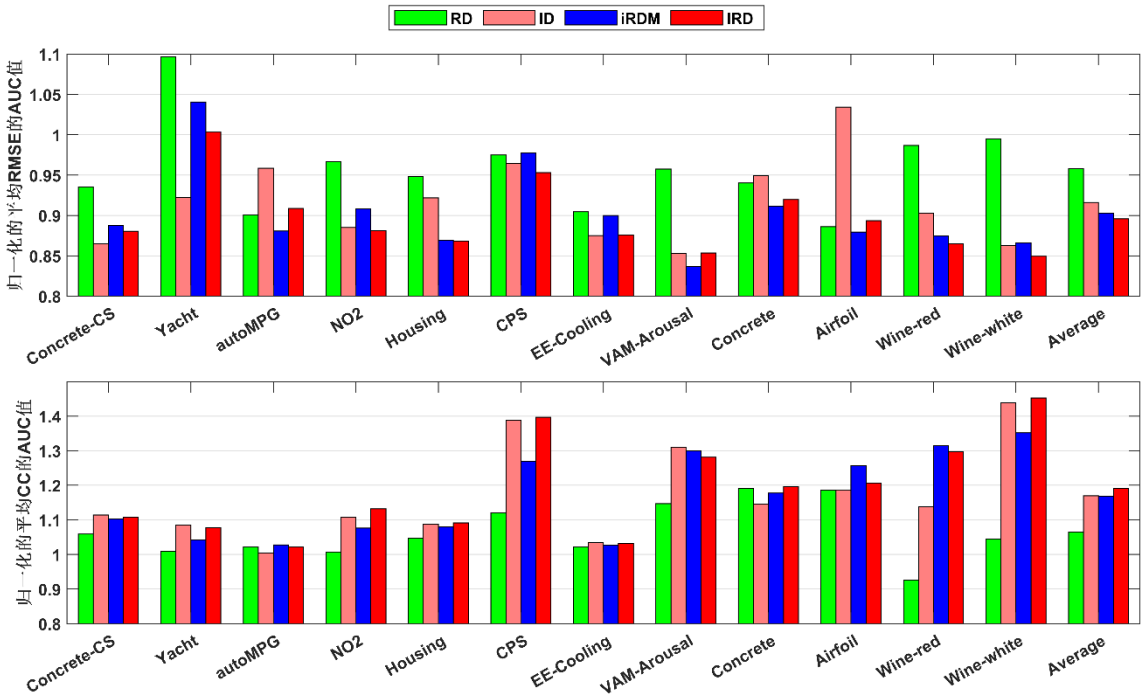

图 4-14 IRD 算法中分散度、代表性和信息量的作用，RBF SVR（$C = 50$）

由图 4-13 和图 4-14 得知对绝大多数数据集，以及 12 个数据集的平均结果，对 Ridge，无论是考察平均 RMSE 的 AUC 值还是平均 CC 的 AUC 值，四种算法的排序是：IRD>iRDM≈ID>RD，这说明分散度、代表性和信息量指标在 IRD 算法中均发挥了作用；对 RBF SVR，四种算法的排序依然是：IRD>iRDM≈ID>RD，不过 IRD 相对于 iRDM 的提升较小，这说明分散度和代表性在 IRD 算法中均发挥了作用，但由于回归模型类型不再是线性模型，信息量发挥的作用较小。

4.5.7 使用无监督主动学习回归算法来提升有监督主动学习回归算法的性能

第 2.1 节提到，可以使用无监督主动学习回归算法来为有监督主动学习回归算法采集初始的少量样本，从而提升有监督主动学习回归算法的性能，本实验将验证之。使用的回归模型是 Ridge，使用的算法是：分别使用随机采样和五种无监督主动学习回归算法（P-ALICE、GSx、RD、iRDM 和 IRD 算法）为有监督主动学习回归算法 RD-EMCM 从完全无标签的样本池中选择最初的$M_0 = 5$个待打标样本，打标并训练初始回归模型，之后使用 RD-EMCM 算法来选择剩下的样本（算法符号：例如使

用 iRDM 算法选择初始样本时，算法表示为：“iRDM+RD-EMCM”）。其他实验设置与之前相同，实验同样重复 100 次。各算法在 12 个数据集上的平均 RMSE 和平均 CC 在$M \in [5,20]$的 AUC 值相对于“RS+RD-EMCM”的比值见图 4-15。

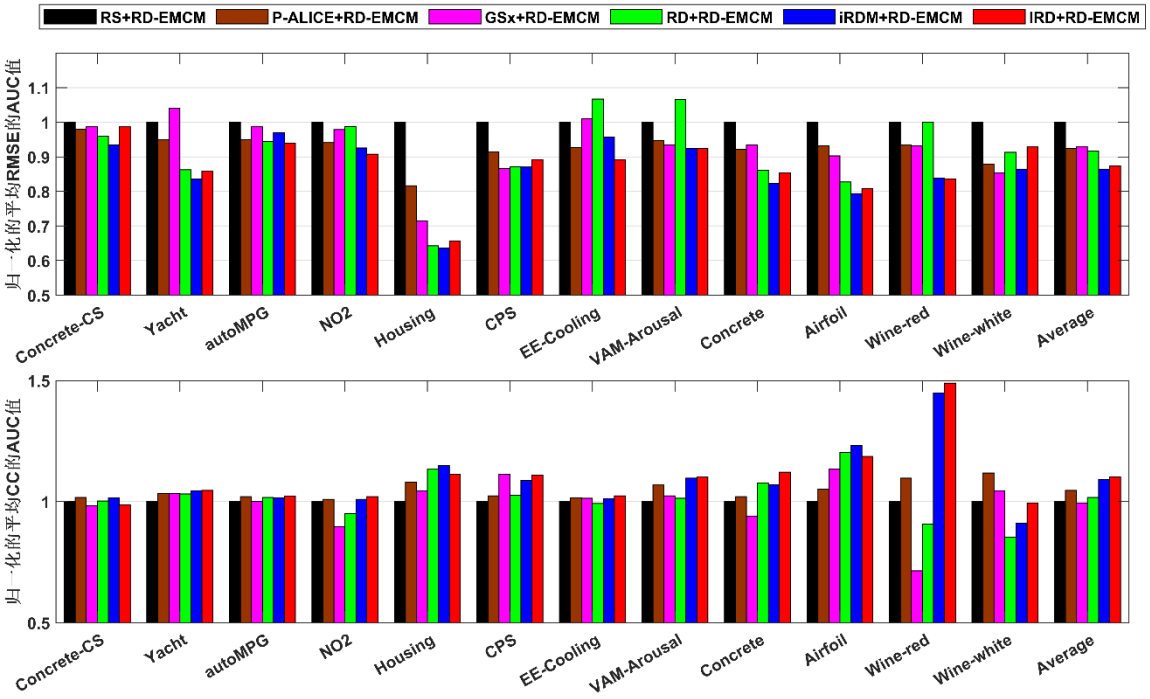

图 4-15 在 12 个数据集上的归一化的平均 RMSE 和平均 CC 的 AUC 值，Ridge（$r = 0.1$）

由图 4-15 得知，对 Ridge，无论考量 RMSE 还是 CC，使用无监督主动学习回归算法替代随机采样来为有监督主动学习回归算法 RD-EMCM 选择初始的少量样本能有效提升 RD-EMCM 的性能；且使用本文提出的 iRDM 和 IRD 算法相比使用其他已有算法提升更多，这是因为好的初始化能够提升初始回归模型的性能。使用 IRD 算法与使用 iRDM 算法的性能提升幅度相当，可能原因是 IRD 算法的信息量指标与 RD-EMCM 中的信息量指标评估方式不一致。

4.6 实验结论及分析

结论 1：无论是考量 RMSE 还是 CC，对 Ridge，实验中实现的所有九种主动学习回归算法的性能和稳定性均优于随机采样；对 RBF SVR，实验中实现的所有七种

主动学习回归算法的性能和稳定性均优于随机采样或与随机采样相当。这证明了实验中各主动学习回归算法的有效性。

结论 2：无论是考量 RMSE 还是 CC，无论对 Ridge 还是对 RBF SVR，本文提出的 iRDM 算法和 IRD 算法的性能和稳定性均优于其他无监督主动学习回归算法，且在训练样本数$M \in [5,20]$时优于有监督主动学习回归算法，且绝大多数性能差异具有统计显著性。这证明了本文提出的 iRDM 和 IRD 相对于现有算法，在 Ridge 和 RBF SVR 中均具有优越性。

结论 3：无论是考量 RMSE 还是 CC，对 Ridge，IRD 算法的性能和稳定性优于 iRDM 算法，且性能提升具有统计显著性差异；对 RBF SVR，IRD 算法的性能和稳定性略优于 iRDM 算法，但是性能差异没有统计显著性。这验证了线性回归中 IRD 算法信息量指标的作用，且信息量指标必须依赖于已知的回归模型类型，对其他类型的回归模型无效或效果有限。

结论 4：对 Ridge，iRDM 算法中分散度和代表性均发挥了作用，IRD 算法中分散度、代表性和信息量也均发挥了作用；对 RBF SVR，iRDM 算法中分散度和代表性均发挥了作用，而 IRD 算法中仅分散度和代表性发挥了作用，信息量没有作用或作用很小。这验证了无监督主动学习回归算法中三种核心指标的作用，以及本文提出的三种核心指标的度量方法的合理性，也再次验证了信息量指标依赖于特定的回归模型类型。

结论 5：在线性回归中，使用更好的无监督主动学习回归算法来为有监督主动学习回归算法选择初始的少量待打标样本，能够有效提升有监督主动学习回归算法的性能。这验证了第 2.1 节阐述的无监督主动学习回归算法的第 1 项实用价值，也指出了一种新的提升有监督主动学习回归算法性能的思路。

结论 6：在训练样本数$M \in [5,20]$时，QBC 算法和 EMCM 算法在 Ridge 中的表现，以及 QBC 算法和 RSAL 算法在 RBF SVR 中的表现不是很理想，性能弱于大多数无监督主动学习回归算法，其原因是在训练样本很少时，少量的真实标签信息可能会误导有监督主动学习回归算法的采样过程。这验证了第 2.1 节阐述的无监督主动

学习回归算法的第 3 项实用价值（第 2.1 节阐述的无监督主动学习回归算法的第 2 项实用价值“只需与人工专家交互一次”为其本身属性，无需实验验证）。

补充分析：在训练样本数$M \in [5,20]$时，本文提出的两种无监督主动学习回归算法性能均优于有监督主动学习回归算法，而在大多数机器学习算法中，往往无监督学习的性能会弱于有监督学习，这体现了“主动学习回归算法”具有一定的特殊性。不过无监督主动学习回归算法并不能完全替代有监督主动学习回归算法，因为两者的使用方式存在一定的差异，如图 2-1 所示，无监督主动学习回归算法能且只能一次性选择所有的待打标样本，而有监督主动学习回归算法能够不断迭代地选择新样本来打标并加入训练集，从而不断地更新回归模型，直到回归模型达到了预期的精度。因此，在打标成本预算确定，即能够打标的样本数量确定的情况下，直接使用无监督主动学习回归算法可能取得比使用有监督主动学习回归算法更好的效果。但若实际应用中只确定了回归模型的预期精度，那么无监督主动学习回归算法只能用于选择初始的少量待打标样本来打标并训练初始回归模型，随后还是需要使用有监督主动学习回归算法来更新模型直到模型达到预期精度。不过无论哪一种情况，无监督主动学习回归算法都能发挥作用，这说明本文提出的算法的适用性较广。

4.7 本章小结

本章实现了 11 种采样算法，包括八种现有主动学习回归算法，本文提出的两种无监督主动学习回归算法以及基线随机采样，并在涵盖多个实际应用领域的 12 个公开回归数据集上，使用线性回归（Ridge）和基于核的非线性回归（RBF SVR）两种回归模型，进行了多次重复实验，然后通过大量图表分析实验数据，并使用 Dunn 算法进行了统计检验。

这些实验和结果验证了实验中实现的主动学习回归算法的有效性；验证了本文提出的无监督主动学习回归算法 iRDM 和 IRD 的优越性；验证了无监督主动学习回归算法中分散度、代表性和信息量三个核心指标的有效性及其度量方法的合理性；验证了信息量指标对回归模型类型的依赖性；验证了使用无监督主动学习回归算法来选择初始的少量待打标样本能有效提升有监督主动学习回归算法的性能；还验证了

在训练样本很少时使用无监督主动学习回归算法能取得比使用有监督主动学习回归算法更好的性能。

5 总结与展望

5.1 总结

本文研究了主动学习在回归问题中的应用，目的是解决某些回归问题中样本的真实标签难以获取的问题。

理论部分，本文首先将主动学习回归算法拆分为有监督情形和无监督情形，并详细对比了无监督主动学习回归算法和有监督主动学习回归算法，指出无监督主动学习回归算法相对于有监督主动学习回归算法的三个优势：“采样过程中无需任何真实标签信息”，“只需要与人工专家交互一次”以及“可以替代随机采样为有监督主动学习回归算法选择初始的少量样本来打标，从而提升其性能”；本文还为无监督主动学习回归算法建立了基于回归模型不确定性的数学模型，并提出一种用于在无监督主动学习回归算法中预测回归模型性能的新指标；本文还将有监督主动学习回归算法中的分散度、代表性和信息量三个核心指标迁移到无监督主动学习回归算法中，并在线性回归和基于核的非线性回归中使用提出的数学模型和预测指标为这三个核心指标提供了理论解释；本文还提出了一种用于在无监督主动学习回归算法中优化待打标样本集的交替优化框架，以及两种新的无监督主动学习回归算法：融合分散度和代表性的 iRDM 算法，以及融合分散度、代表性和信息量的 IRD 算法，其中包含本文提出的三种核心指标的度量方式：“分散度度量”、“代表性度量”和针对线性回归模型的“分散度-信息量联合度量”。本文为无监督主动学习回归算法建立的数学模型，提出的预测指标，总结的三个核心指标及其对应的三种度量方式，能为后续的主动学习回归算法的研究提供理论依据和新思路。

实验部分，本文在 12 个公开的回归数据集上进行了大量实验，使用线性回归（Ridge）和基于核的非线性回归（RBF SVR）两种回归模型，证明了本文提出的无监督主动学习回归算法的性能比现有的无监督主动学习回归算法更优越也更稳定；实验还证明了新算法性能的提升是得益于对分散度、代表性和信息量的合理度量；实验还详细对比了本文提出的 iRDM 算法和 IRD 算法的性能表现，发现 IRD 算法只在线性模型中性能优于 iRDM 算法，从而证明了信息量指标对回归模型的类型具有依

赖性；实验还证明了使用无监督主动学习回归算法代替随机采样过程为有监督主动学习回归算法选择初始的少量待打标样本，能够有效提升有监督主动学习回归算法的性能；此外实验还发现：在训练样本很少的条件下，本文提出的无监督主动学习回归算法性能优于现有的有监督主动学习回归算法。

5.2 展望

本文的研究只考虑了主动学习在线性回归和基于核函数的非线性回归中的应用，未来可以研究主动学习在其他功能更强大的回归模型中的应用，例如决策树回归，神经网络回归等。

本文中研究的主动学习回归算法在采样过程中特征空间是固定的，未来或许可以在采样过程中引入特征变换，让特征空间更加适合当前的主动学习回归算法，从而使算法更容易采集到有价值的样本。

目前的工作只将主动学习与有监督回归算法结合，未来可以考虑将主动学习与半监督回归算法结合，这样可以充分利用无标签样本的信息，进一步提升回归模型的性能，或进一步减少打标成本。

目前的 IRD 算法中，具有理论依据的信息量指标只在采集最初的$d+1$个样本（d为样本特征维数）时用到，当训练样本数量M和样本特征维度数d满足$M>d+1$时需要使用其他算法采集剩下的样本。这使得 IRD 算法被分成了两个阶段，从而降低了算法的易用性，未来可以思考如何将它们融合为一个阶段，并提出更完善的理论解释。

目前的 IRD 算法在$M<d+1$时需要先对特征空间降维，而降维会丢失部分信息，未来可以考虑如何在$M<d+1$时度量分散度和信息量，同时不丢失信息或尽可能减少信息丢失。

目前的 IRD 算法中分散度-信息量联合度量需要计算特征空间中的超平面方程，若特征空间维度较高则会比较耗时，未来可以考虑如何在高维数据集中更高效地度量分散度和信息量。

在使用无监督主动学习回归算法来为有监督主动学习回归算法选择初始的M_0个样本时，本文仅设置$M_0 = 5$，但这不一定是最优的，今后的研究可以考虑如何根据具体问题确定最优的M_0。

致谢

日月如梭，光阴似箭，一转眼，三年的研究生旅程就要结束了。

首先，感谢我的导师。他是一位学识渊博、科研作风严谨、对学生严格要求的好导师。遇见恩师，三生之幸。记得我刚进入实验室时基础较为薄弱，是他不厌其烦地一遍又一遍地指导我，甚至亲自指导我如何使用代码来实现自己的算法，将实验技能一项项地传授给我；在我的科研遇到瓶颈的时候，是他将自己的经历分享给我，一直鼓励我前行；在我迷茫的时候，他的谆谆教诲更是让我受益良多，也让我逐渐坚定自己的方向。此外，他在生活中也给了我很多关心和支持，他的体谅、包容和耐心帮助我一步步地成长，真的十分感动，感谢恩师！

其次，感谢实验室的各位同学们，难忘和大家一起开组会，一起分享自己的工作，一起探讨遇到的困难，一起奋斗时的快乐和幸福，感谢大家一路走来对我的支持和帮助，很高兴能和大家相伴三年！

同时，感谢我的父亲和母亲，他们一直都是我坚实的后盾。感谢他们在我二十余年寒窗苦读期间无微不至的关心与不求回报的付出！

感谢我的女友，她在我困难的时候一直支持着我，在我失落的时候一直鼓励着我，在我孤单的时候一直陪伴着我。

2020 年是不平凡的一年，疫情给我们带来阻挠，但我们终会胜利！感谢在疫情期间依然关心我们的学校领导和各位老师们，感谢各位医护人员和志愿者们，是你们的努力为我营造安全的环境让我能够顺利地完成毕业论文。

祝大家平安幸福！

参考文献

[1] Picard R. *Affective Computing* [M]. Cambridge, MA, USA: MIT Press, 1997

[2] Russell J A. A circumplex model of affect [J]. *Journal of Personality and Social Psychology*, 1980, 39(6): 1161-1178

[3] Mehrabian A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies* [M]. Cambridge, MA, USA: Oelgeschlager, Gunn & Hain, 1980

[4] Grimm M, Kroschel K, and Narayanan S S. The Vera Am Mittag German audio-visual emotional speech database [C]. In Proc. Int'l Conf. on Multimedia & Expo (ICME). Hannover, German, June 2008. 865-868

[5] Koelstra S, et al. DEAP: A database for emotion analysis; Using physiological signals [J]. *IEEE Trans. on Affective Computing*, 2012, 3(1): 18-31

[6] Bradley M M, Lang P J. The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual [R]. University of Florida, Gainesville, FL, Technical Report B-3, 2007

[7] Wu D, Lawhern V J, Gordon S, et al. Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR) [J]. *IEEE Trans. on Fuzzy Systems*, 2017, 25(6): 1522–1535

[8] Wu D, Lawhern V J, Gordon S, et al. Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression [C]. in Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics. Budapest, Hungary, October 2016. 730–736

[9] Wu D, Lawhern V J, Gordon S, et al. Spectral meta-learner for regression (SMLR) model aggregation: Towards calibrationless brain-computer interface (BCI) [C]. in Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics. Budapest, Hungary, October 2016. 743–749

[10] Chuang S -W, Ko L -W, Lin Y -P, et al. Co-modulatory spectral changes in independent brain processes are correlated with task performance [J]. Neuroimage, 2012, 62: 1469-1477

[11] Chuang C -H, Ko L -W, Jung T -P, et al. Kinesthesia in a sustained-attention driving task [J]. Neuroimage, 2014, 91: 187-202

[12] Joo J, Wu D, Mendel J M, et al. Forecasting the post fracturing response of oil wells in a tight reservoir [C]. In Proc. SPE Western Regional Meeting, San Jose, CA, March 2009

[13] Settles B. Active learning literature survey [R]. University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009

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

[15] Sugiyama M. Active learning in approximately linear regression based on conditional expectation of generalization error [J]. *Journal of Machine Learning Research*, 2006, 7: 141–166

[16] Willett R, Nowak R, Castro R M. Faster rates in regression via active learning [C]. In Proc. Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada, December 2006. 179-186

[17] Wu D, Huang J. Affect estimation in 3D space using multitask active learning for regression [J]. *IEEE Trans. on Affective Computing*, 2020, in press

[18] Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data [C]. in Proc. 34th Int'l. Conf. on Machine Learning (ICML). Sydney, Australia, August 2017. 1183—1192

[19] Huang S, Jin R, Zhou Z. Active Learning by Querying Informative and Representative Examples [J]. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2014, 36(10): 1936-1949

[20] Wu D. Pool-based sequential active learning for regression [J]. *IEEE Trans. on Neural Networks and Learning Systems*, 2019, 30(5): 1348-1359

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

[22] Cai W, Zhang M, Zhang Y. Batch mode active learning for regression with expected model change [J]. *IEEE Trans. on Neural Networks and Learning Systems*, 2017, 28(7): 1668-1681

[23] Cai W, Zhang Y, Zhou J. Maximizing expected model change for active learning in regression [C]. In Proc. IEEE 13th Int'l. Conf. on Data Mining (ICDM). Dallas, TX, USA, December 2013. 51–60

[24] RayChaudhuri T, Hamey L G C. Minimisation of data collection by active learning [C]. in Proc. IEEE Int'l. Conf. on Neural Networks. WA, Australia, Nov./Dec. 1995. 1338–1341

[25] Douak F, Melgani F, and Benoudjit Nabil. Kernel ridge regression with active learning for wind speed prediction [J]. *Applied Energy*, 2013, 103: 328–340

[26] Abe N, Mamitsuka H. Query learning strategies using boosting and bagging [C]. in Proc. 15th Int'l Conf. on Machine Learning (ICML). Madison, WI, USA, July 1998. 1-9

[27] Freund Y, Seung H, Shamir E, et al. Selective sampling using the query by committee algorithm [J]. *Machine Learning*, 1997, 28(2-3): 133-168

[28] Seung H S, Opper M, Sompolinsky H. Query by committee [C]. in Proc. ACM Workshop on Computational Learning Theory. Pittsburgh, PA, USA, July 1992. 287–294

[29] Marathe A D, Lawhern V J, Wu D, et al. Improved neural signal classification in a rapid serial visual presentation task using active learning [J]. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 2016, 24(3): 333-343

[30] Burbidge R, Rowland J J, King R D. Active learning for regression based on query by committee [R]. Lecture Notes in Computer Science, 2007, 4881: 209-218

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

[32] Demir B, Bruzzone L. A multiple criteria active learning method for support vector regression [J]. *Pattern Recognition*, 2014, 47: 2558-2567

[33] Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning [C]. in Proc. Neural Information Processing Systems (NIPS). Denver, CO, USA, November 1995: 231–238

[34] Cai W, Zhang Y, Zhou S, et al. Active learning for support vector machines with maximum model change [R]. in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2014. 211–216

[35] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks [C]. in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, HI, USA, October 2008. 1070–1079

[36] Settles B, Craven M, Ray S. Multiple-instance active learning [C]. in Proc. Advances in Neural

Information Processing Systems (NIPS). Vancouver, BC, Canada, December 2008. 1289-1296
[37] Donmez P, Carbonell J G. Optimizing estimated loss reduction for active sampling in rank learning [C]. in Proc. 25th Int'l Conf. on Machine Learning (ICML). Helsinki, Finland, July 2008. 248–255
[38] Yu H, Kim S. Passive sampling for regression [C]. In Proc. IEEE Int'l. Conf. on Data Mining (ICDM). Sydney, Australia, December 2010. 1151-1156
[39] Wu D, Lin C T, Huang J. Active learning for regression using greedy sampling [J]. *Information Sciences*, 2019, 474: 90-105
[40] 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, 2020, 审稿中
[41] Grimm M, Kroschel K. Emotion estimation in speech using a 3D emotion space concept [M]. In Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech, 2007. 281-300
[42] Grimm M, Kroschel K, Mower E, et al. Primitives-based evaluation and estimation of emotions in speech [J]. *Speech Communication*, 2007, 49: 787-800
[43] Wu D, Parsons T D, Mower E, et al. Speech emotion estimation in 3D space [C]. In Proc. IEEE Int'l Conf. on Multimedia & Expo (ICME), Singapore, July 2010. 737-742
[44] Wu D, Parsons T D, Narayanan S S. Acoustic feature analysis in speech emotion primitives estimation [C]. In Proc. InterSpeech, Makuhari, Japan, September 2010
[45] Ilin A, Raiko T. Practical Approaches to Principal Component Analysis in the Presence of Missing Values [J]. *Journal of Machine Learning Research*, 2010, 11: 1957–2000
[46] Dunn O J. Multiple comparisons among means [J]. *Journal of the American Statistical Association*, 1961, 56: 62-64
[47] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing [J]. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1995, 57: 289-300
[48] 伍冬睿, 刘子昂. 一种用于语音情感计算的无监督主动学习方法 [P]. CN110782876A, 2020-02-11
[49] Hoerl A E, Kennard R W. Ridge Regression: Biased Estimation for Nonorthogonal Problems [J]. *Technometrics*, 1970, 12(1): 55–67
[50] Hoerl A E, Kennard R W. Ridge Regression: Applications to Nonorthogonal Problems [J]. *Technometrics*, 1970, 12(1): 69–82
[51] Marquardt D W, Snee R D. Ridge Regression in Practice [J]. *The American Statistician*, 1975, 29(1): 3–20
[52] Fan R, Chen P, Lin C. Working set selection using second order information for training support vector machines [J]. *Journal of Machine Learning Research*, 2005, 6: 1889–1918

附录 1　攻读学位期间发表的论文及专利

[1] Liu Z, Wu D. Integrating Informativeness, Representativeness and Diversity in Pool-Based Sequential Active Learning for Regression [C]. in Proc. International Joint Conference on Neural Networks. Glasgow, UK, July 2020, accepted

[2] 伍冬睿, 刘子昂. 一种用于语音情感计算的无监督主动学习方法 [P]. CN110782876A, 2020-02-11

[3] 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, 2020, 审稿中

附录 2　攻读学位期间参与的科研项目

1. 参与国家自然科学基金面上项目——机器学习模型泛化能力启发的区间二型鲁棒模糊控制研究（No. 61873321）。