智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种自训练框架下的三优选半监督回归算法

程康明， 熊伟丽

引用本文:
程康明, 熊伟丽. 一种自训练框架下的三优选半监督回归算法[J]. 智能系统学报, 2020, 15(3): 568–577.
CHENG Kangming, XIONG Weili. Three-optimal semi-supervised regression algorithm under self-training framework[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 568–577.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201905033

DOI: 10.11992/tis.201905033

一种自训练框架下的三优选半监督回归算法

程康明[1], 熊伟丽[2]

（1. 江南大学 物联网工程学院, 江苏 无锡 214122; 2. 江南大学 轻工过程先进控制教育部实验室, 江苏 无锡 214122）

摘　要: 工业生产过程数据由于主导变量分析代价等因素可能出现有标签样本少而无标签样本多的情况, 为提升对无标签样本利用的准确性与充分性, 提出一种自训练框架下的三优选半监督回归算法。对无标签样本与有标签样本进行优选, 保证两类数据的相似性, 以提高无标签样本预测的准确性; 利用高斯过程回归方法对所选有标签样本集建模, 预测所选无标签样本集, 得到伪标签样本集; 通过对伪标签样本集置信度进行判断, 优选出置信度高的样本用于更新初始样本集; 为了进一步提高无标签样本利用的充分性, 在自训练框架下, 进行多次循环筛选提高无标签样本的利用率。通过对脱丁烷塔过程实际数据的建模仿真, 验证了所提方法在较少有标签样本情况下的良好预测性能。

关键词: 工业生产; 无标签样本; 优选; 半监督回归; 相似性; 高斯过程回归; 置信度判断; 自训练; 预测

中图分类号: TP274　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)03−0568−10

中文引用格式: 程康明, 熊伟丽. 一种自训练框架下的三优选半监督回归算法 [J]. 智能系统学报, 2020, 15(3): 568–577.

英文引用格式: CHENG Kangming, XIONG Weili. Three-optimal semi-supervised regression algorithm under self-training framework[J]. CAAI transactions on intelligent systems, 2020, 15(3): 568–577.

Three-optimal semi-supervised regression algorithm under self-training framework

CHENG Kangming[1], XIONG Weili[2]

(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; 2. Key Laboratory of Advanced Process Control for Light Industry, Jiangnan University, Wuxi 214122, China)

Abstract: In industrial production, due to factors such as the cost of analyzing the dominant variable, there may be cases in which there are few labeled and many unlabeled samples. To improve performance and accuracy in the use of unlabeled samples, we propose the use of a three-optimal semi-supervised regression algorithm under a self-training framework. This algorithm first filters unlabeled and labeled samples to ensure similarity between these two types of data and improve the accuracy of predicting the unlabeled samples. Then, a model is established based on the selected labeled samples using Gaussian process regression to predict the unlabeled samples, from which pseudo-label samples are obtained. Then, by determining the confidence levels of the prediction of the pseudo-label samples, samples with higher confidence levels are filtered and used to update the initial samples. Finally, through multiple filtering loops, a self-training framework is applied to improve the utilization of unlabeled samples. By modeling and simulating debutanizer process data, the proposed method was confirmed to have superior prediction performance when there are an insufficient number of labeled samples.

Keywords: industrial production; unlabeled samples; filter; semi-supervised regression; similarity; Gaussian process regression; confidence judgment; self-training; prediction

半监督回归是半监督学习领域一个重要的研究方向, 其基本思想是利用无标签样本信息, 以提高有标签样本的建模效果[1-3]。现有的半监督

收稿日期: 2019−05−15.

基金项目: 国家自然科学基金项目 (61773182); 江苏省自然科学基金项目 (BK20170198).

通信作者: 熊伟丽. E-mail: greenpre@163.com.

回归方法依据其学习方法大致分为两类[4-5]: 基于流行学习的半监督回归算法和协同训练算法。其中, 流行学习方面的研究大都着眼于样本结构信息, 通过挖掘样本蕴含的结构信息, 进而提高算法性能[6-9]。协同训练的主要思路是对有标签样本建立两个学习器, 在假设数据存在双视图的前提下, 通过相互学习, 逐步提高每个学习器的精度, 从而实现对无标签样本信息的准确预测, 进而达到提升模型精度的目的[10-13]。此外, 半监督思想能够与其他典型方法结合, 充分发挥原有方法的优势和半监督思想的特点, 在相关领域也获得了显著成果[14-15]。

自训练 (Self-Training) 是半监督学习的一种重要的学习技术, 一般结合其他算法来实现数据信息的充分挖掘[16-19]。通常用于半监督分类, 与 SVM 结合构成 Self-Training SVM, 通过充分利用无标签样本信息来提高分类器的泛化性能[20]。但其最近也有一些新的发展, 如仝小敏等[21] 将自训练思想应用于回归领域, 充分挖掘并利用无标签样本信息, 提高了回归准确率, 为回归问题提供了一种新的思路。

上述方法有的考虑了有标签样本问题, 有的考虑了无标签样本问题, 有的考虑了置信度判断问题, 有的考虑了利用的充分性等。但是, 当有标签样本集很小时, 上述方法的使用就会出现不足, 如协同训练需要两个冗余的视图, 流行学习大多利用有标签样本信息等。

本文针对有标签样本少的问题, 综合考虑以上各个方面对建模的影响, 为提升对无标签样本利用的准确性与充分性, 提出一种自训练框架下的三优选半监督软测量算法。首先, 对无标签样本与有标签样本进行优选, 保证两类数据的相似性, 以提高无标签样本预测的准确性; 其次, 利用高斯过程回归方法对所选有标签样本集建模, 预测所选无标签样本集, 得到伪标签样本集; 接着, 通过对伪标签样本集置信度进行判断, 优选出置信度高的伪标签样本用于更新初始样本集; 最后, 为了进一步提高无标签样本利用的充分性, 在自训练框架下, 进行多次循环筛选以提高无标签样本的利用率。

1　相关知识

1.1　高斯过程回归

GPR 是一种基于统计学习理论的非参数概率模型, 适合处理高维度、小样本及非线性等数据的建模问题[22]。高斯过程回归算法流程: 给定训练样本集 $\{\boldsymbol{X},\boldsymbol{y}\}$, 其中 $\boldsymbol{X}=\{\boldsymbol{x}_i\in R^D\}$, $\boldsymbol{y}=\{\boldsymbol{y}_i\in\mathbf{R}\}$, $i=1,2,\cdots,n$, 分别代表 D 维的输入和输出数据, 通常观测值 $\boldsymbol{y}_i$ 与输入 $\boldsymbol{x}_i$ 应满足 $\boldsymbol{y}_i=f(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}$, 其中 f 为未知函数形式, 高斯噪声 $\boldsymbol{\varepsilon}\sim N(0,\sigma^2)$。对于一个新的输入 $\boldsymbol{x}^*$, 相应的预测值 y^* 也满足高斯分布, 其均值和方差如式 (1) 和 (2) 所示:

$$\mu(\boldsymbol{y}^*)=\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x}^*)\boldsymbol{K}^{-1}\boldsymbol{y} \tag{1}$$

$$\sigma^2(\boldsymbol{y}^*)=\boldsymbol{C}(\boldsymbol{x}^*,\boldsymbol{x}^*)-\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x}^*)\boldsymbol{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^*) \tag{2}$$

式中: $\boldsymbol{k}(\boldsymbol{x}^*)$ 为训练样本与测试样本之间的协方差矩阵; $\boldsymbol{K}$ 为训练样本的自协方差矩阵; $\boldsymbol{C}$ 为测试样本自协方差。

本文选择常用的高斯协方差函数, 定义如下:

$$\boldsymbol{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)=v\exp\left[-\frac{1}{2}\sum_{d=1}^{D}\boldsymbol{\omega}_d\left(\boldsymbol{x}_i^d-\boldsymbol{x}_j^d\right)\right] \tag{3}$$

式中: v 为先验知识的总体度量; $\boldsymbol{\omega}_d$ 表示每个成分 $\boldsymbol{x}^d$ 的重要性程度。

根据式 (1)~(3) 确定模型的均值与方差, 对任意输入 $\boldsymbol{x}^*$, 即可预测其输出 $\boldsymbol{y}^*$。详细的 GPR 建模过程可参考文献 [23]。

1.2　半监督回归置信度判断

置信度判断是半监督回归中一个至关重要的问题, 其决定了能否选出可信的伪标签样本更新有标签样本集, 直接影响算法预测性能。

现有的算法一般借鉴 COREG 算法中的假设: 若利用了置信度最高的无标签样本, 则对更新后的有标签样本集建模能获得最好的预测效果。换言之, 置信度最高的样本是指样本加入有标签样本集后使标签样本的预测值与实际值最一致的样本[24]。该类算法已取得不错的效果。如文献 [14] 中, 通过将式 (4) 最大化取出置信度最高的无标签样本:

$$\Delta_u=\sqrt{\sum_{\boldsymbol{x}_i\in\boldsymbol{\Omega}_u}(\boldsymbol{y}_i-h(\boldsymbol{x}_i))^2/k}-\sqrt{\sum_{\boldsymbol{x}_i\in\boldsymbol{\Omega}_u}(\boldsymbol{y}_i-h'(\boldsymbol{x}_i))^2/k} \tag{4}$$

式中: Δ_u 为预测误差的差值; $\boldsymbol{\Omega}_u$ 为无标签样本 $\boldsymbol{x}_u$ 的 k 个近邻有标签样本组成的集合; y_i 为验证集的真值; h 为 $\boldsymbol{\Omega}_u$ 训练所得初始学习器; h' 为 $\boldsymbol{\Omega}_u$ 加上 $\boldsymbol{x}_u$ 及其标签 $h(\boldsymbol{x}_u)$ 后的集合训练所得新学习器。

1.3　自训练算法

自训练是半监督学习中一种重要的学习技术, 一般结合其他算法用于半监督分类上, 其作

用是提高无标签样本的利用率，以提高模型的泛化性能[25]。其工作流程：首先，根据已有的少量有标签样本训练出一个初始的分类器；然后，利用分类器预测无标签样本的标签，选出置信度高的一部分伪标签样本加入有标签样本集；最后，重复执行以上步骤，若达到预期性能或循环次数限制，则跳出循环。

将自训练框架融入本文算法，能实现对无标签样本充分准确地利用，算法思想如图 1 所示。

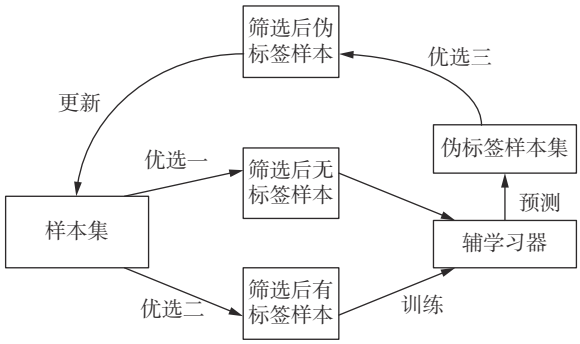

图 1　算法的基本思想

Fig. 1　Basic concept of the algorithm

2　本文算法

2.1　算法分析

半监督学习的基本思想是假设数据分布服从三大假设，建立学习器以预测无标签样本标签值[1]。半监督学习的基本设置：给定一个未知分布的有标签样本集 $\boldsymbol{S}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_{|S|},y_{|S|})\}$ 以及一个无标签样本集 $\boldsymbol{M}=\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_{|M|}\}$，期望得到模型 $f:\boldsymbol{X}\to\boldsymbol{y}$，能准确预测无标签样本 $\boldsymbol{x}'$ 的标签。此处，$\boldsymbol{x}_i,\boldsymbol{x}'_i\in\boldsymbol{X}$，$y_i\in\boldsymbol{y}$ 为样本 $\boldsymbol{x}_i$ 的标签，针对回归问题，一般假设 $\boldsymbol{y}$ 服从流行假设，即局部邻域输出具有相似性；$|\boldsymbol{S}|$ 与 $|\boldsymbol{M}|$ 分别为 $\boldsymbol{S}$ 与 $\boldsymbol{M}$ 集的势，即其所包含的样本数。

图 2 为自训练框架下的三优选半监督回归算法的流程图，主要包含自训练框架，无标签样本筛选，有标签样本筛选和置信度判断 4 个部分。由于初始有标签样本数量少，仅通过现存有标签样本无法对大多数无标签样本进行准确预测，为了实现无标签样本的准确预测，依据有标签样本的分布特点，先对无标签样本进行筛选，选出能够被准确预测的无标签样本。同时，由于有标签样本少，可能存在的有标签样本的离群点会对辅学习器的性能产生较大影响，故根据有标签样本的分布特点，剔除离群点，然后建立辅学习器，实现对无标签样本更准确地预测。

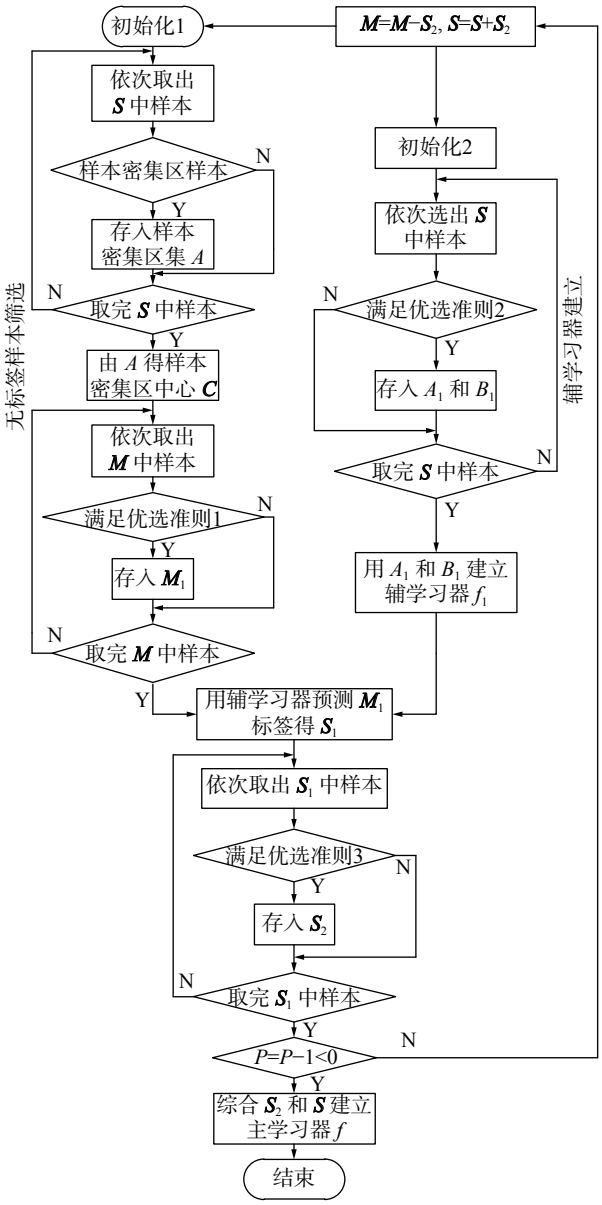

图 2　算法流程图

Fig. 2　Flow chart of the algorithm

通过以上两步筛选过程，准确预测了部分无标签样本，但是也导致无标签样本利用率不高，故引入了自训练的框架，通过多次循环提高无标签样本利用率。由于自训练对预测准确度要求很高，一旦达不到精度要求，易造成误差的累积，大幅度降低模型预测精度，故通过置信度判断，对预测的无标签样本进一步筛选，进一步提高无标签样本预测的准确性。

在本文算法学习过程中，首先根据无标签样本筛选过程，对无标签样本集 $\boldsymbol{M}$ 筛选得到 $\boldsymbol{M}_1$；同时根据辅学习器建立过程，通过筛选有标签样本集 $\boldsymbol{S}$ 得到 $\boldsymbol{A}_1$ 和 $\boldsymbol{B}_1$，建立辅学习器 f_1；然后利用 f_1 预测 $\boldsymbol{M}_1$ 得到伪标签样本集 $\boldsymbol{S}_1$；最后利用置信度判断筛选 $\boldsymbol{S}_1$ 得到 $\boldsymbol{S}_2$。重复上述过程 P 次，将 $\boldsymbol{S}_2$ 结合

$\boldsymbol{S}$ 建立主学习器, 所得主学习器将拥有良好的预测精度和泛化能力。

2.2　三优选策略

本文提出的三优选策略, 综合考虑无标签样本筛选、有标签样本筛选以建立学习器, 以及基于置信度判断的伪标签样本筛选等三方面问题, 进而达到准确预测无标签样本的目的, 以防止自训练过程误差累积情况的发生。优选策略中对无标签样本和有标签样本进行的优选, 不仅保证初期仅预测有标签样本分布范围内的无标签样本, 而且排除了有标签样本中可能的离群点的影响, 因此初始预测误差不会过大, 避免了在有标签样本少的情况下自训练算法初期可能产生过大震荡。

2.2.1　三优选准则

优选准则 1: 给定一个阈值 θ_1, 利用马氏距离[26]度量无标签样本 $\boldsymbol{x}'_i$ 与有标签样本密集区中心 $\boldsymbol{C}$ 的相似度, 若 $\boldsymbol{x}'_i$ 与 $\boldsymbol{C}$ 的距离小于 θ_1, 则 $\boldsymbol{x}'_i$ 满足优选条件。

优选准则 2: 给定一个阈值 θ_2, 利用马氏距离度量有标签样本间的相似度, 统计有标签样本 $\boldsymbol{x}_i$ 与其余有标签样本的马氏距离小于 θ_2 的有标签样本数量 m, 若 m 不小于 2, 则 $\boldsymbol{x}_i$ 满足优选条件。

优选准则 3: 给定一个阈值 θ_3, 判断每一个伪标签样本加入建模过程后对模型预测效果的影响, 若模型对测试样本预测方差 var 小于阈值 θ_3, 则该伪标签样本可信, 能够用于更新有标签样本集。

上述 3 个优选准则, 优选准则 1 通过马氏距离选出以有标签样本密集区中心为球心 θ_1 为半径的超球体内全部的无标签样本点, 该准则有利于选出能被准确利用的无标签样本。优选准则 2 采用了 Knorr 等[27]提出的离群点的定义, 利用马氏距离度量相似度, 剔除了阈值 θ_2 限定下的离群点, 从而提升辅学习器的预测精度。准则 1 和准则 2 的阈值一般由工程经验设定, 也可通过枚举法获取, 即先固定阈值 θ_2, 更改阈值 θ_1, 由其对模型预测结果的影响确定 θ_1 最优质阈值范围, 同理可得阈值 θ_2 范围, 然后利用枚举法, 在范围内枚举 θ_1 与 θ_2, 根据模型预测结果得 θ_1 与 θ_2。优选准则 3 通过对伪标签样本的筛选, 选出其中可信的伪标签样本用于更新有标签样本集与无标签样本集, 其基本思想是假设当前样本集为最优样本集, 根据当前有标签样本数据集计算阈值, 利用该阈值通过优选准则 3 判断某伪标签样本的加入对算法预测效果的影响, 若算法的预测误差变小, 则该伪标签样本可信, 用于更新样本集; 另外, 不可信的伪标签样本不采用其标签, 等待更新后的样本集对其重新预测, 根据该准则能够显著提高无标签样本集预测的准确性。有标签样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间马氏距离 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 计算公式为

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{sqrt}\left[(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)\right] \tag{5}$$

$$\boldsymbol{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} \tag{6}$$

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{7}$$

式中: $\boldsymbol{S}$ 是有标签样本的协方差矩阵; n 为变量个数; $\bar{\boldsymbol{x}}$ 为有标签样本均值。

另外, 优选准则 1 中 $\boldsymbol{x}'_i$ 与 $\boldsymbol{C}$ 的距离 d_i 同样可用式 (5)~(7) 获得。优选准则 3 中的阈值 θ_3 与预测方差 var 计算为

$$f_{\mathrm{gpr}}=\mathrm{gprtrain}(\boldsymbol{B}_1,\boldsymbol{A}_1) \tag{8}$$

$$\boldsymbol{y}_{\mathrm{predict}}=f_{\mathrm{gpr}}(\boldsymbol{x}_{\mathrm{test}}) \tag{9}$$

$$\theta_3=\sum_{i=1}^{N}(\boldsymbol{y}_{\mathrm{predict}}(i)-\boldsymbol{y}_{\mathrm{test}}(i))^2 \tag{10}$$

从 $\boldsymbol{S}_1$ 中依次选择样本加入 $\boldsymbol{B}_1$ 和 $\boldsymbol{A}_1$ 中, 组成 $\boldsymbol{B}'_1$ 和 $\boldsymbol{A}'_1$, 再用其建模。

$$f'_{\mathrm{gpr}}=\mathrm{gprtrain}(\boldsymbol{B}'_1,\boldsymbol{A}'_1) \tag{11}$$

$$\hat{\boldsymbol{y}}_{\mathrm{predict}}=f'_{\mathrm{gpr}}(\boldsymbol{x}_{\mathrm{test}}) \tag{12}$$

$$\mathrm{var}=\sum_{i=1}^{N}(\hat{\boldsymbol{y}}_{\mathrm{predict}}(i)-\boldsymbol{y}_{\mathrm{test}}(i))^2 \tag{13}$$

式中: f'_{gpr} 和 f_{gpr} 表示所建立模型; $\mathrm{gprtrain}(\boldsymbol{B}_1,\boldsymbol{A}_1)$ 表示根据 $\boldsymbol{B}_1$ 和 $\boldsymbol{A}_1$ 建立 GPR 模型; $\boldsymbol{B}_1$ 和 $\boldsymbol{A}_1$ 分别为有标签样本集的辅助变量和标签值; $\boldsymbol{x}_{\mathrm{test}}$ 和 $\boldsymbol{y}_{\mathrm{test}}$ 分别为测试样本集的辅助变量和标签值; $\boldsymbol{y}_{\mathrm{predict}}$ 为模型对 $\boldsymbol{x}_{\mathrm{test}}$ 的预测值; $\boldsymbol{B}'_1$ 和 $\boldsymbol{A}'_1$ 分别为加入伪标签样本后有标签样本集的辅助变量和标签值; $\hat{\boldsymbol{y}}_{\mathrm{predict}}$ 为更新后的有标签样本集对 $\boldsymbol{x}_{\mathrm{test}}$ 的预测值。

2.2.2　无标签样本筛选

为了选出能够被准确预测的无标签样本, 利用优选准则 1 对无标签样本进行筛选, 选出满足条件的无标签样本, 保证所选出的无标签样本分布于有标签样本所分布的范围内, 以降低无标签样本预测错误的可能性, 同时避免自训练算法初期可能存在过大误差导致算法初期过大的震荡。而准则 1 能否做出有效筛选, 取决于能否获得准确的有标签样本分布范围, 尤其是有标签样本数目比较少的情况, 此时有标签样本分布范围对离群点更敏感。针对此类情况, 提出一种寻找有标签样本密集区中心的方法。算法流程如图 3 所示。

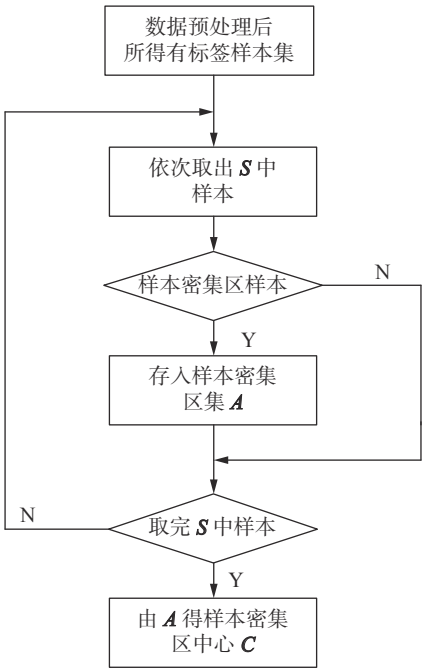

图 3　样本密集区中心获取

Fig. 3　Obtaining the center of the samples in the dense area

获得样本密集区中心后，准则 1 的筛选效果得以保证，再利用其筛选无标签样本。算法流程如图 4 所示。

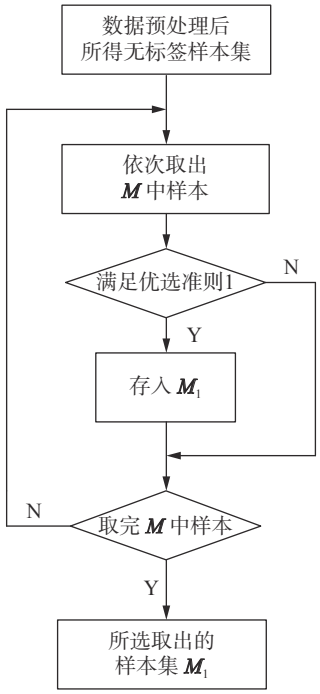

图 4　无标签样本优选

Fig. 4　Filtering of unlabeled samples

以上算法流程图中，图 3 为获取样本密集区中心的流程，图 4 为无标签样本优选，按照流程，首先进行样本密集区中心的计算，然后根据所得密集区中心优选无标签样本，具体算法步骤：

1) 初始化参数。$\boldsymbol{A}$、$\boldsymbol{M}_1$ 初始赋空集，初始化阈值 θ_1 和 θ_4 及有标签样本集 $\boldsymbol{S}$ 和无标签样本集 $\boldsymbol{M}$；

2) 依次取出 $\boldsymbol{S}$ 中样本 $\boldsymbol{x}_i$；

3) 由式 (5)~(7) 计算得 $\boldsymbol{x}_i$ 与其余有标签样本间的马氏距离 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$；

4) 统计满足条件 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)<\theta_4$ 的次数，若次数不小于 2，则 $\boldsymbol{x}_i$ 为样本密集区样本，将其存入 $\boldsymbol{A}$；

5) 判断是否取完 $\boldsymbol{S}_1$ 中样本，若取完，根据 $\boldsymbol{A}$ 计算样本密集区中心 $\boldsymbol{C}$，跳转至 6)，开始优选无标签样本，否则，跳转至 2)，继续寻找样本密集区样本；

6) 依次取出 $\boldsymbol{M}$ 中样本 $\boldsymbol{x}'_j$；

7) 由式 (5)~(7) 计算得 $\boldsymbol{x}'_j$ 与样本密集区中心 $\boldsymbol{C}$ 的马氏距离 d_j；

8) 判断是否满足优选准则 1，若满足，将 $\boldsymbol{x}'_j$ 存入 $\boldsymbol{M}_1$，跳转至 9)，否则，直接跳转至 9)；

9) 判断是否取完 $\boldsymbol{M}$ 中样本，若取完，跳出循环，否则，跳转至 6)，继续优选无标签样本。

2.2.3　辅学习器建立

为获得富有针对性的辅学习器，已知 $\boldsymbol{M}_1$ 分布于 $\boldsymbol{C}$ 周围的前提下，利用优选准则 2 筛选出与 $\boldsymbol{M}_1$ 特性更相符的有标签样本，即可根据所选有标签样本建立辅学习器，实现对 $\boldsymbol{M}_1$ 更准确的预测。辅学习器建立流程图如图 5 所示。

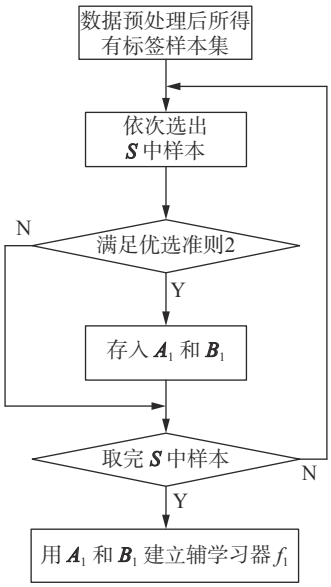

图 5　辅学习器的建立

Fig. 5　Building the auxiliary learning machine

如图 5 所示，利用马氏距离度量样本间的相似度，判断每一个有标签样本是否满足优选准则 2，选出其中满足条件的样本，分别存入 $\boldsymbol{A}_1$ 和 $\boldsymbol{B}_1$，$\boldsymbol{A}_1$ 中存辅助变量，$\boldsymbol{B}_1$ 中存对应的标签值，然后利用 GPR 建立辅学习器 f_1。

2.2.4　置信度判断

在自训练框架下，若加入不准确的伪标签样本，容易造成主学习器误差累积，从而导致预测

效果差, 故必须对伪标签样本进行置信度判断。现有的置信度判断方法大都利用文献 [19] 所提的方法, 其每次选出置信度列表上最高的伪标签样本加入有标签样本集, 随着过程的进行, 部分置信度不高的伪标签样本升到置信度列表的最高位置, 现有的置信度算法无法避免这部分置信度不高的伪标签样本加入有标签样本集。其次, 在有标签样本很少的情况下, 某些无标签样本集很可能拥有相同的 k 个近邻有标签样本组成的集 Ω_u, 从而导致部分伪标签样本具有相同的置信度而无法选出最高置信度的伪标签样本; 最后, 由于无标签样本少, k 个近邻有标签样本组成的集 Ω_u 较小, 导致所有的无标签样本本身无法得到准确的预测, 即使每次选出置信度列表上最高的伪标签样本, 也无法显著提高主学习器预测效果。

综上, 为了对有标签样本少的情况进行置信度判断, 结合所提优选准则 3, 提出了一种新的半监督回归置信度判断方法, 具体算法步骤如下:

1) 初始化, 伪标签样本集 $\boldsymbol{S}_1$, 阈值 θ_3;

2) 根据式 (8)~(10) 计算得阈值 θ_3;

3) 依次取出 $\boldsymbol{S}_1$ 中样本 $\boldsymbol{x}_i$;

4) 根据式 (12)~(14) 计算得 var;

5) 判断是否满足优选准则 3, 若满足, 将 $\boldsymbol{x}_i$ 存入 $\boldsymbol{S}_2$, 跳转至 6), 否则, 直接跳转至 6);

6) 判断是否取完 $\boldsymbol{S}_1$ 中样本, 若取完, 跳出循环, 否则, 跳转至 2), 继续优选。

2.3　算法总体步骤

自训练框架下的三优选半监督回归算法步骤如下:

1) 初始化参数, 包括有标签样本集 $\boldsymbol{S}$, 无标签样本集 $\boldsymbol{M}$, 循环次数 P;

2) 筛选无标签样本, 得到无标签样本集 $\boldsymbol{M}_1$;

3) 筛选有标签样本, 得到 $\boldsymbol{B}_1$ 和 $\boldsymbol{A}_1$, 并利用其建立辅学习器 f_1;

4) 利用 f_1 预测 $\boldsymbol{M}_1$ 的标签值, 得到伪标签样本集 $\boldsymbol{S}_1$;

5) 对伪标签样本集 $\boldsymbol{S}_1$ 进行置信度判断, 筛选出其中可信的伪标签样本组成 $\boldsymbol{S}_2$;

6) 根据 $\boldsymbol{S}_2$ 更新有标签样本集 $\boldsymbol{S}$ 与无标签样本集 $\boldsymbol{M}$;

7) 重复进行 2)~6), 循环 P 次;

8) 根据更新后的有标签样本集 $\boldsymbol{S}$ 建立主学习器 f。

本文算法使用了自训练算法的框架, 本文算法的收敛性能分析如下: 假设第一次循环初始样本建模预测误差为 h, rmse_0^k 为第 k 次循环学习器初始误差, θ_3^k 为第 k 次循环优选准则 3 置信度评判指标, rmse_1^k 为第 k 次循环结束时学习器预测误差。每次循环初始误差为上次循环结束时的误差, 即 $\mathrm{rmse}_0^k = \mathrm{rmse}_1^{k-1}$, 根据优选准则 3 的定义, 每次寻找使初始预测误差更小的伪标签样本, 用于更新样本集, 每一轮的最终预测误差均会不大于初始预测误差, 即 $\mathrm{rmse}_1^k \leqslant \mathrm{rmse}_0^k$, 且由式 (8)~(10) 计算 θ_3^k, 有 $\theta_3^k = \mathrm{rmse}_0^k$, 若算法共循环了 p 次, 则有 $\mathrm{rmse}_1^p \leqslant \mathrm{rmse}_0^p = \mathrm{rmse}_1^{p-1} \leqslant \mathrm{rmse}_0^{p-1} = \cdots = \mathrm{rmse}_1^1 \leqslant \mathrm{rmse}_0^1 = h$, 所以, 本文算法是不发散的。

3　脱丁烷塔仿真研究

为了验证本文所提方法的实际效果, 以脱丁烷塔过程[28] 作为对象。脱丁烷塔装置的产品丁烷的浓度对石油产量和质量影响很大, 但丁烷浓度难以直接测量, 如何建立准确模型预测丁烷浓度是石油炼制领域一直以来的一个重要研究课题。脱丁烷塔过程样本数据有 7 个辅助变量: 塔顶温度 x_1、塔顶压力 x_2、塔顶回流量 x_3、塔顶产品流出 x_4、第 6 层塔板温度为 x_5、x_6 为塔底温度 1、x_7 为塔底温度 2。主导变量是塔底丁烷浓度。详细的工艺过程描述见文献 [29]。

该实验过程数据共 2 394 组样本。为模拟有标签样本少而无标签样本多的情况, 对数据进行截选, 有标签样本和无标签样本分别为 150 组与 500 组。

为了详细展现本文算法性能, 纵向比较了几点创新对模型跟踪效果的影响, 具体比较如下:

1)GPR 方法[30]。仅利用已有的有标签样本建立 GPR 模型。

2) 无优选半监督 GPR(none-optimal semi-supervised GPR, NS-GPR) 方法[30]。直接利用有标签样本建模, 获取辅学习器, 然后预测无标签样本的标签, 进而利用伪标签样本更新初始有标签样本集, 进而建模。

3) 双优选半监督 GPR(double-optimal semi-supervised GPR, DS-GPR) 方法[30]。利用优选准则 1 筛选无标签样本, 同时利用优选准则 2 筛选有标签样本, 接着对其建模得到辅学习器, 后续同方法 2。

4) 自训练双优选半监督 GPR(double-optimal semi-supervised GPR from self-training framework, SFDS-GPR) 方法。在双优选的基础上, 引入自训练框架, 循环执行双优选建模过程。

5) 三优选半监督 GPR(three-optimal semi-Supervised GPR, TS-GPR) 方法。在双优选的基础上, 利用优选准则 3, 对预测的无标签样本进行置信度的判断, 筛选出其中可信的伪标签样本。

6) 本文方法。

图 6 为几种方法对丁烷浓度真实值的跟踪效

果，其中，基准线上的点表示预测值与真实值一致，故样本点离基准线越近，表示预测效果越好。

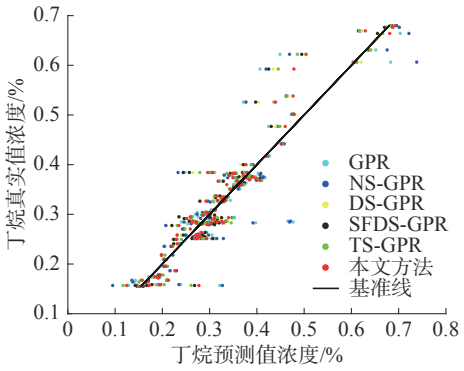

图 6　不同方法的纵向对比

Fig. 6　Longitudinal comparison of different methods

由图 6 可知，本文方法的效果更好，尤其在丁烷浓度高的区域。为了更直观地对比各方法的预测效果，图 7 展现了对真实值的跟踪误差。

综合图 6 和图 7 可知，NS-GPR 虽然考虑了无标签样本信息，但其无差别的利用将会带来大量噪声，而 DS-GPR 由于经过了双优选过程，提升了对初始无标签样本预测的准确性，主学习器对丁烷浓度预测效果有了明显的提高。SFDS-GPR 在 DS-GPR 的基础上加入了自训练框架，期望达到提高无标签样本利用率的目的，但是，由于自训练框架对初始无标签样本预测准确性要求较高，DS-GPR 达不到精度要求，故引入自训练框架无法提高主学习器预测效果，反而会由于误差累积带来大量噪声，降低主学习器预测效果。TS-GPR 在 DS-GPR 的基础上加入了置信度判断，对预测所得伪标签样本进行置信度判断，选出其中可信的伪标签样本更新有标签样本集，对主学习器预测效果有了更明显的提升。

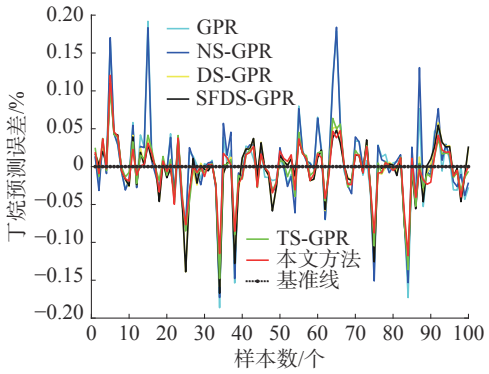

图 7　不同方法的预测误差对比

Fig. 7　Comparison of the prediction errors of different methods

本文方法在考虑无标签样本利用准确性的同时，引入了置信度判断，进一步提升伪标签样本的准确性，引入了自训练框架提高无标签样本的利用率，综合考虑了无标签样本利用的准确性和充分性，充分利用了无标签样本所包含的信息，获得了良好的效果。

为了展现方法在每个区间的预测效果，统计真实值与各方法预测值的分布情况如图 8 所示。

分析图 8 各方法在不同区域的预测效果，发现在 0.1%~0.15% 内 GPR 和 NS-GPR 对丁烷浓度真实值的预测效果较差，在 0.4%~0.5% 内，DS-GPR、SFDS-GPR 和 TS-GPR 都有一定的预测误差，而本文方法在上述区域中的预测误差都相对较小。

以上对比证明了本文方法的优越性，为了更加直观地体现优选过程，选取了样本中塔顶压力、塔顶产品流出及塔底温度 1 等 3 个维度的量，利用三维图表现出了这些量在三优选过程中分布的变化。

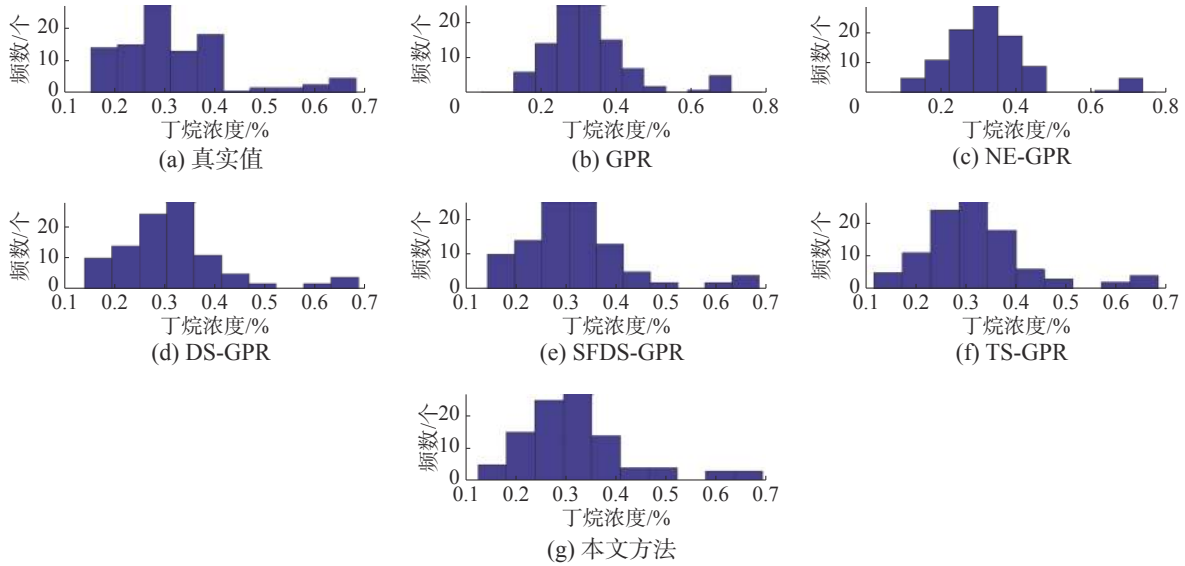

图 8　多种方法预测值与真实值的直方图统计

Fig. 8　Histogram of the predicted and true values of different methods

由图 9 的过程发现, 第 1 次优选过程中无标签样本集与有标签样本集有明显差异, 可知初始无标签样本集无法准确预测所有的无标签样本, 仅能够准确预测少量无标签样本, 经过三优选过程后, 发现筛选后的无标签样本集主要集中在筛选后的有标签样本集附近, 样本集的相似度显著提升, 故筛选后的无标签样本被准确预测的概率更高; 在第 2 次优选过程中, 经过样本集更新, 有标签样本集对无标签样本的预测能力加强, 在一次三优选过程后, 筛选后的无标签样本集与筛选后的有标签样本集相似度进一步提升, 筛选所得无标签样本被进行更准确地预测; 一直到第 4 次优选过程, 有标签样本集经过更新, 所蕴含信息已经能够对这一过程进行准确解释, 本着充分挖掘无标签样本所有信息的目的, 再一次进行三优选过程, 发现此次筛选前后无标签样本集基本没有差异, 即已经完成对无标签样本信息的充分利用, 其余少数未被选到的无标签样本是无法进行准确预测的, 强行预测并利用其信息只会带来明显的误差, 对建模有害无益。最终充分准确地利用了无标签样本所包含的有用信息, 使得建模更准确可靠。

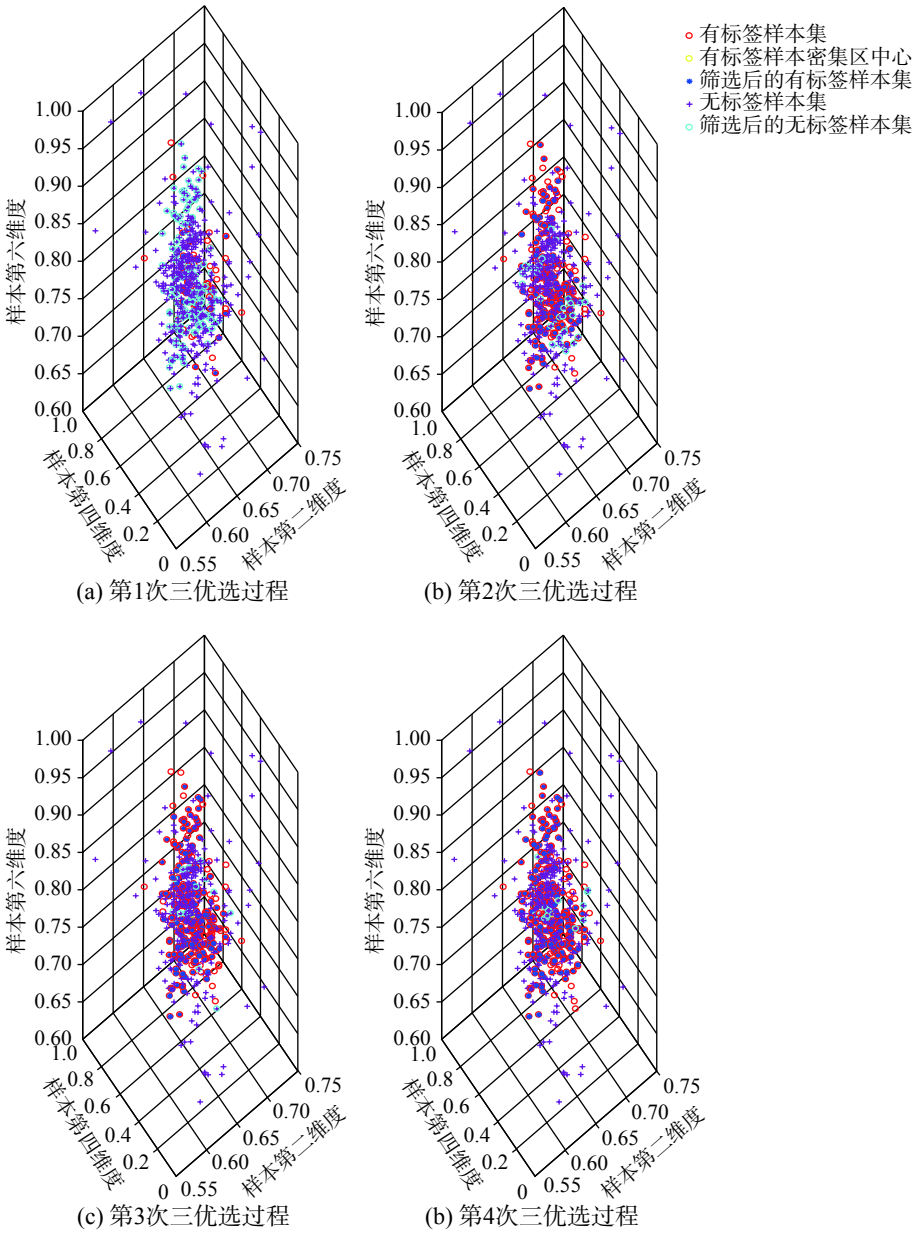

图 9　本文方法变量信息变化

Fig. 9　Changes in the values' information of the methods used in this study

上述分析对比验证了本文方法的实际效果。发现在对无标签样本进行准确充分地利用后, 即使初始有标签样本较少, 依然能够取得很好的预测效果。各方法具体跟踪效果如表 1 所示。

表 1　不同方法预测效果对比

Table 1　Comparison of the predictive effects of different methods

方法	均方根误差
GPR	0.055 1
NS-GPR	0.055 6
DS-GPR	0.040 5
SDS-GPR	0.041 3
TS-GPR	0.036 7
本文方法	0.033 5

4　结束语

本文所提算法，通过所定义的 3 个优选准则和自训练框架，从无标签样本、有标签样本、置信度判断与样本利用率等角度，实现了对样本信息充分准确地利用。充分利用有标签样本已有信息，准确估计无标签样本缺失信息，进而利用所估计的无标签样本信息，扩充有标签样本已有信息，从而大幅度提升了主学习器的预测效果。用脱丁烷塔过程数据验证本文算法实际效果，实验结果表明，所提方法在有标签样本较少时，具有良好的预测效果，为半监督回归提供了一种新思路。

参考文献：

[1] SHAHSHAHANI B M, LANDGREBE D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE transactions on geoscience and remote sensing, 1994, 32(5): 1087–1095.

[2] LIU Huizeng, SHI Tiezhu, CHEN Yiyun, et al. Improving spectral estimation of soil organic carbon content through semi-supervised regression[J]. Remote sensing, 2017, 9(1): 29.

[3] 姜婷, 袭肖明, 岳厚光. 基于分布先验的半监督 FCM 的肺结节分类 [J]. 智能系统学报, 2017, 12(5): 729–734.
JIANG Ting, XI Xiaoming, YUE Houguang. Classification of pulmonary nodules by semi-supervised FCM based on prior distribution[J]. CAAI transactions on intelligent systems, 2017, 12(5): 729–734.

[4] 刘杨磊, 梁吉业, 高嘉伟, 等. 基于 Tri-training 的半监督多标记学习算法 [J]. 智能系统学报, 2013, 8(5): 439–445.
LIU Yanglei, LIANG Jiye, GAO Jiawei, et al. Semi-supervised multi-label learning algorithm based on Tri-training[J]. CAAI transactions on intelligent systems, 2013, 8(5): 439–445.

[5] 徐蓉, 姜峰, 姚鸿勋. 流形学习概述 [J]. 智能系统学报, 2006, 1(1): 44–51.
XU Rong, JIANG Feng, YAO Hongxun. Overview of manifold learning[J]. CAAI transactions on intelligent systems, 2006, 1(1): 44–51.

[6] LIN Tong, ZHA Hongbin. Riemannian manifold learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(5): 796–809.

[7] 赵立杰, 王海龙, 陈斌. 基于流形正则化半监督学习的污水处理操作工况识别方法 [J]. 化工学报, 2016, 67(6): 2462–2468.
ZHAO Lijie, WANG Hailong, CHEN Bin. Identification of wastewater operational conditions based on manifold regularization semi-supervised learning[J]. CIESC journal, 2016, 67(6): 2462–2468.

[8] 杜永贵, 李思思, 阎高伟, 等. 基于流形正则化域适应湿式球磨机负荷参数软测量 [J]. 化工学报, 2018, 69(3): 1244–1251.
DU Yonggui, LI Sisi, YAN Gaowei, et al. Soft sensor of wet ball mill load parameter based on domain adaptation with manifold regularization[J]. CIESC journal, 2018, 69(3): 1244–1251.

[9] 陈定三, 杨慧中. 基于局部重构融合流形聚类的多模型软测量建模 [J]. 化工学报, 2011, 62(8): 2281–2286.
CHEN Dingsan, YANG Huizhong. Multiple model soft sensor based on local reconstruction and fusion manifold clustering[J]. CIESC journal, 2011, 62(8): 2281–2286.

[10] ZHOU Zhihua, LI Ming. Semi-supervised regression with co-training[C]//Proceedings of the 19th International Joint Conference on Artificial Intelligence. Scotland, UK, 2005: 908−913.

[11] CALDAS W L, GOMES J P P, MESQUITA D P P. Fast Co-MLM: an efficient semi-supervised co-training method based on the minimal learning machine[J]. New generation computing, 2018, 36(6): 41–58.

[12] CHEN Minmin, WEINBERGER K Q, BLITZER J C. Co-training for domain adaptation[C]//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain, 2011: 2456−2464.

[13] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. Wisconsin, USA, 1998: 92−100.

[14] 程玉虎, 冀杰, 王雪松. 基于 Help-Training 的半监督支持向量回归 [J]. 控制与决策, 2012, 27(2): 205–210.
CHENG Yuhu, JI Jie, WANG Xuesong, et al. Semi-supervised support vector regression based on help-training[J]. Control and decision, 2012, 27(2): 205–210.

[15] 盛高斌, 姚明海. 基于半监督回归的选择性集成算法 [J]. 计算机仿真, 2009, 26(10): 198–201.
SHENG Gaobin, YAO Minghai. An ensemble selection

algorithm based on semi-supervised regression[J]. Computer simulation, 2009, 26(10): 198–201.

[16] BESEMER J, LOMSADZE A, BORODOVSKY M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions[J]. Nucleic acids research, 2001, 29(12): 2607–2618.

[17] LI Fan, CLAUSI D A, XU Linlin, et al. ST-IRGS: a region-based self-training algorithm applied to hyperspectral image classification and segmentation[J]. IEEE transactions on geoscience and remote sensing, 2018, 56(1): 3–16.

[18] SALI L, DELSANTO S, SACCHETTO D, et al. Computer-based self-training for CT colonography with and without CAD[J]. European radiology, 2018, 28(11): 4783–4791.

[19] 张博锋, 白冰, 苏金树. 基于自训练 EM 算法的半监督文本分类 [J]. 国防科技大学学报, 2007, 29(6): 65–69.
ZHANG Bofeng, BAI Bing, SU Jinshu. Semi-supervised text classification based on self-training EM algorithm[J]. Journal of national university of defense technology, 2007, 29(6): 65–69.

[20] LI Yuanqing, GUAN Cuntai, LI Huiqi, et al. A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system[J]. Pattern recognition letters, 2008, 29(9): 1285–1294.

[21] 仝小敏, 吉祥. 基于自训练的回归算法 [J]. 中国电子科学研究院学报, 2017, 12(5): 498–502.
TONG Xiaomin, JI Xiang. Regression algorithm based on self training[J]. Journal of China academy of electronics and information technology, 2017, 12(5): 498–502.

[22] KUMAR S, HEGDE R M, TRIGONI N. Gaussian process regression for fingerprinting based localization[J]. Ad hoc networks, 2016, 51: 1–10.

[23] 熊伟丽, 李妍君, 姚乐, 等. 一种动态校正的 AGMM-GPR 多模型软测量建模方法 [J]. 大连理工大学学报, 2016, 56(1): 77–85.
XIONG Weili, LI Yanjun, YAO Le, et al. A dynamically corrected AGMM-GPR multi-model soft sensor modeling method[J]. Journal of Dalian University of Technology, 2016, 56(1): 77–85.

[24] ZHOU Zhihua, LI Ming. Semisupervised regression with cotraining-style algorithms[J]. IEEE transactions on knowledge and data engineering, 2007, 19(11): 1479–1493.

[25] TANHA J, VAN SOMEREN M, AFSARMANESH H. Semi-supervised self-training for decision tree classifiers[J]. International journal of machine learning and cybernetics, 2017, 8(1): 355–370.

[26] LAW M T, YU Yaoliang, CORD M, et al. Closed-form training of mahalanobis distance for supervised clustering[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3909−3917.

[27] KNORR E M, NG R T. A unified notion of outliers: properties and computation[C]//Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. Newport Beach, USA, 1997: 219−222.

[28] 阮宏镁, 田学民, 王平. 基于联合互信息的动态软测量方法 [J]. 化工学报, 2014, 65(11): 4497–4502.
RUAN Hongmei, TIAN Xuemin, WANG Ping. Dynamic soft sensor method based on joint mutual information[J]. CIESC journal, 2014, 65(11): 4497–4502.

[29] FORTUNA L, GRAZIANI S, RIZZO A, et al. Soft sensors for monitoring and control of industrial processes[M]. London: Springer, 2007: 229−231.

[30] 程康明, 熊伟丽. 一种双优选的半监督回归算法 [J]. 智能系统学报, 2019, 14(4): 689–696.
CHENG Kangming, XIONG Weili. A dual-optimal semi-supervised regression algorithm[J]. CAAI transactions on intelligent systems, 2019, 14(4): 689–696.

作者简介:

程康明, 硕士研究生, 主要研究方向为工业过程建模、机器学习和大数据分析。

熊伟丽, 教授, 博士生导师, 主要研究方向为复杂工业过程建模及优化、智能优化算法及应用。主持国家自然科学基金面上项目、江苏省产学研等纵向项目 8 项; 参与国家 863 计划、重点研发计划等多项。发 表学术论文 60 余篇。