Journal of System Simulation

Volume 36 | Issue 1 Article 19

1-20-2024

Multi-model Soft Sensor Modeling under Help-training Strategy

Luosuyang He
School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China,
hlsy1224@foxmail.com

Weili Xiong
School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi 214122, China

Multi-model Soft Sensor Modeling under Help-training Strategy

Abstract

Abstract: Due to the strong nonlinearity, multi-stage coupling, and the small number of labeled samples in complex industrial processes, it is difficult for traditional global soft sensor models to accurately describe the whole process. Therefore, a multi-model soft sensor modeling method under the helptraining strategy is proposed. This method uses a fuzzy C-means (FMC) clustering algorithm to mine similar samples in the sample set and build several sub-models. By introducing the help-training strategy, a collaborative training framework based on main and auxiliary learners is formed, and a confidence evaluation mechanism is designed to eliminate error samples and expand the modeling space of submodels. Then the fuzzy membership degree is used as the probability distribution function of D-S evidence theory to calculate the weight of the sub-model, and the output of the sub-model is fused to obtain the final model prediction result. Through the modeling and simulation of the actual data of the debutanizer industrial process, the results show that this model has good prediction performance.

Keywords

soft sensor modeling, multi-model, help-training, learner, debutanizer

Recommended Citation

He Luosuyang, Xiong Weili. Multi-model Soft Sensor Modeling under Help-training Strategy[J]. Journal of System Simulation, 2024, 36(1): 249-259.

第 36 卷第 1 期 系统仿真学报© Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

助训练策略下的多模型软测量建模

何罗苏阳[1]，熊伟丽[1,2]
(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重点实验室，江苏 无锡 214122)

摘要：由于复杂工业过程中存在强非线性、多阶段耦合以及有标签样本数量偏少的情况，传统的全局软测量模型难以精确描述整个过程。为此，提出一种助训练策略下的多模型软测量建模方法。*该方法采用模糊C均值聚类算法挖掘样本集中的相似性样本并建立若干子模型；通过引入助训练策略，形成基于主、辅学习器的协同训练框架，并设计置信度评估机制淘汰误差样本的同时扩充子模型的建模空间；进而将模糊隶属度作为D-S证据理论的概率分配函数计算出子模型权重，对子模型的输出进行融合以得到最终的模型预测结果。*通过对脱丁烷塔工业过程的实际数据进行建模仿真，结果表明此模型具有良好的预测性能。

关键词：软测量建模；多模型；助训练；学习器；脱丁烷塔
中图分类号：TP274 文献标志码：A 文章编号：1004-731X(2024)01-0249-11
DOI: 10.16182/j.issn1004731x.joss.22-0976
引用格式: 何罗苏阳, 熊伟丽. 助训练策略下的多模型软测量建模[J]. 系统仿真学报, 2024, 36(1): 249-259.
Reference format: He Luosuyang, Xiong Weili. Multi-model Soft Sensor Modeling under Help-training Strategy[J]. Journal of System Simulation, 2024, 36(1): 249-259.

Multi-model Soft Sensor Modeling under Help-training Strategy

He Luosuyang[1], *Xiong Weili*[1,2]
(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;
2. Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: Due to the strong nonlinearity, multi-stage coupling, and the small number of labeled samples in complex industrial processes, it is difficult for traditional global soft sensor models to accurately describe the whole process. Therefore, a multi-model soft sensor modeling method under the help-training strategy is proposed. *This method uses a fuzzy C-means (FMC) clustering algorithm to mine similar samples in the sample set and build several sub-models. By introducing the help-training strategy, a collaborative training framework based on main and auxiliary learners is formed, and a confidence evaluation mechanism is designed to eliminate error samples and expand the modeling space of sub-models. Then the fuzzy membership degree is used as the probability distribution function of D-S evidence theory to calculate the weight of the sub-model, and the output of the sub-model is fused to obtain the final model prediction result.* Through the modeling and simulation of the actual data of the debutanizer industrial process, the results show that this model has good prediction performance.

Keywords: soft sensor modeling; multi-model; help-training; learner; debutanizer

收稿日期：2022-08-18 修回日期：2022-09-18
基金项目：国家自然科学基金(61773182)；国家重点研发计划(2018YFC1603705-03)
第一作者：何罗苏阳(1997-)，男，硕士生，研究方向为复杂工业过程建模。E-mail：hlsy1224@foxmail.com

0 引言

复杂工业生产过程中，一些重要的质量变量往往具有强非线性和多阶段耦合等特征，并存在有标签样本少、无标签样本多的情况。伴随着工业过程信息提取能力的不断提高，软测量建模中数据驱动建模方法得到迅速发展[1]。相比于机理建模，数据驱动类方法不依赖过多的生产经验和过程知识，而是采用更加直接的“黑箱”建模方式。但由于只是根据输入–输出关系所建立，其针对复杂工业生产过程中的多工况、多阶段等特征，单一的全局模型难以对其准确描述[2]。此外，在利用过程变量中的无标签样本扩充建模空间时，容易引入误差样本导致模型退化[3]。因此，将半监督学习与集成学习的思想结合，在采用多模型建模的同时引入高质量的无标签样本不失为一种解决问题的有效方法。

半监督软测量建模方法的关键往往在于所引入无标签样本的质量。其中，如何在扩充建模空间的同时保证模型精度尤为重要。助训练[4]算法作为半监督学习的一种新型算法，已经在模式识别领域取得了较大的进展，与传统的自训练算法相比，它可以通过引入生成式模型建立有效的筛选机制，从样本中选取含有重要信息的数据，具有了更加突出的学习和修正能力[5-6]。目前，文献[7]引入Parzen窗来估计类概率密度，构成了助训练支持向量分类器，大幅提高了分类性能；文献[8]继承了助训练思想，根据样本规模选择更加适合的方法建模，提高了模型的泛化性和准确性；然而，由于训练框架中引入了仅由有标签样本训练的生成模型，文献[9]发现在此类样本无法满足一定连续性时，使用助训练进行分类所得到的决策边界会发生偏移，故其利用聚类方法揭示数据的真实空间结构后，补偿有标签样本训练的局限性以提高模型性能。因此，对多维样本空间解耦从而确保数据的流形性质，有助于样本在具有非线性特征以及空间结构不准确时建立高精度软测量模型。

另一方面，建立多模型集成的软测量模型在精度需满足实际工业要求的情况下，不但要确保足够充分的建模样本，选择可靠的子模型融合方式也至关重要。文献[10]提出一种基于模糊C均值(fuzzy C-means, FCM)聚类的多模型软测量方法，以各样本点间的相似程度作为阶段划分方法，最后利用最小二乘支持向量机(least square support vector machine, LSSVM)建模。文献[11]则在FCM划分的基础上，利用Adaboost算法优化后的高斯过程回归建模，将模糊隶属度直接作为权值因子，最后加权得到模型的预测输出。然而，传统的多模型集成方法更侧重于通过原始样本和变量维度产生更具有相似性的子样本集，未能选择可靠的多模型融合策略。基于开关切换的简单集成方法仅适用于线性系统，无法应用于复杂工业过程；而基于加权融合的集成方法尽管适用于非线性系统，但由于实际工业过程中的强非线性和多阶段耦合等特征，仍会造成多模型融合后的输出精度差，发生模型退化等现象[12-13]。

针对上述问题，本文提出一种助训练策略下的多模型软测量建模方法。首先，采用FCM方法对工业过程数据聚类，根据样本的高维空间特征对多操作工况进行剖析，以保证各阶段样本的连续性；然后，构建以孪生支持向量回归(twin support vector regression, TSVR)为主学习器，K近邻(K-nearest neighbor，KNN)为辅学习器的子模型，利用助训练策略形成一种主辅协同学习器，从而淘汰误差样本并不断扩充样本集中的有标签样本；最后，利用D-S证据理论融合不确定信息的优势，兼顾证据体的可靠性和冲突性，将模糊隶属度作为概率分配函数计算出各子模型的权值因子，提升了权值的准确度。将所建立的模型应用于脱丁烷塔浓度的预测，验证了所提方法的有效性和建模精度。

1 预备知识

1.1 助训练策略

助训练策略是一种新型半监督学习方法，其

DOI: 10.16182/j.issn1004731x.joss.22-0976

核心是引入基于生成式模型的辅学习器，通过主学习器和辅学习器之间的协同训练，依靠相似度自动地学习以达到筛选有效样本并扩充建模空间的效果[8]。其主要思想如下：

构建一个完成回归任务的学习器相当于逼近一个未知的目标函数$f:X\to Y$，或计算可能性$P(Y|X)$，其中X表示输入，Y表示输出。其策略是利用训练样本去计算出$P(X|Y)$和$P(Y)$并利用贝叶斯公式进行新样本的预测。$P(X|Y)$的分布可以描述如何以标签Y为条件生成X，所以这种模型称为生成式模型。相比之下，重点在于决策边界的判别式模型是以后验概率为基础建模，其对于每一个新样本，通过$Y=f(X)$计算其标签值，两种模型的对比如图1所示。助训练策略中，基于判别式模型的主学习器凭借其较低的学习成本和良好的预测性能，能够通过学习最优边界来建立有效的预测模型；而基于生成式模型的辅学习器则利用更快的学习收敛速度，以及对数据的信息挖掘能力，可以反映数据的特征信息并协助主学习器建模。故本文中将两种模型互补构建助训练框架，应用于半监督回归建模的过程当中。

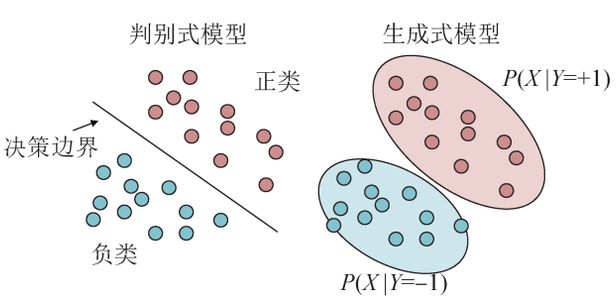


图1 生成式与判别式模型对比说明

Fig. 1 Comparison of generative model and discriminative model

助训练策略的基本框架由基于判别式方法的主学习器和基于生成式方法的辅学习器构成。辅学习器建立置信度评估机制，并通过建模挖掘样本中的有效信息。因生成模型不仅得到一种类别，还得到一个置信度$P(Y|X)$，故辅学习器可以协助主学习器选择所需样本对其标记并添加到训练集中。因此，在每一次迭代中，辅学习器用于选择其类别中高置信度的样本，所选样本构成了标记过程中的候选样本集。最后，主学习器对预筛选后的样本添加伪标签，并再次通过置信度评估选择出高置信度样本加入训练集，不断循环至所有无标签样本都添加了伪标签或达到终止条件。算法伪代码如下所示。

算法：助训练算法

输入 M=有标签样本集，N=无标签样本集，主学习器A，辅学习器B，迭代次数t

输出 新的有标签样本集W

初始化工作集，令$W=M$

while N=非空集 or 迭代次数≤t do

用W训练学习器A和B

B经置信度评估选择N中的高置信度样本

A对B所选择的样本添加伪标签

A经置信度评估选择高置信度样本加入W

从N中删除对应样本

end while

1.2 D-S 证据理论

D-S 证据理论是利用特定的合成规则将多个证据体的基本概率分配函数组合在一起计算出对每个证据的支持程度，并作为最终的决策权重[14-15]，其主要原理为：设Θ为识别框架(有限集合)，2^Θ是Θ的所有子集的集合，基本概率分配函数m是2^Θ到区间$[0,1]$的一个映射，且满足$m(\varnothing)=0$，$\sum_{A\subseteq\Theta}m(A)=1$。设在识别框架$\Theta$中有$\beta$个证据，其基本概率分配分别为$m_1,m_2,\cdots,m_\beta$，则经D-S 合成规则组合后的基本概率分配函数$m$的计算公式为

$$\begin{cases} m(A)=\dfrac{\sum\limits_{\cap A_r=A}\prod\limits_{j=1}^{\beta}m_j(A_r)}{1-\sum\limits_{\cap A_r=\varnothing}\prod\limits_{j=1}^{\beta}m_j(A_r)}, A\neq\varnothing \\ m(\varnothing)=0 \end{cases} \tag{1}$$

式中：A，A_r为识别框架Θ中的焦元；$r=1,2,\cdots,\lambda$，λ为Θ中焦元的个数；$j=1,2,\cdots,\beta$，β为证据体个数。

2 基于助训练策略的多模型软测量方法

传统基于半监督策略的软测量建模方法往往只建立单一的全局质量预测模型，但复杂工业过程中强非线性和多操作阶段耦合等特性广泛存在，所以全局模型无法兼顾各阶段特征[16]。因此，采用FCM聚类划分工业过程中各操作阶段，针对每个操作阶段建立子模型；在助训练策略下，主、辅学习器不断进行半监督学习，通过每一次训练来扩充原有标签样本集；最后利用D-S证据理论进行决策融合，融合各阶段局部模型的输出，得到最终的全局软测量模型。所提算法的建模包括3部分内容：①无监督聚类；②半监督训练；③决策融合。算法流程如图2所示，具体的过程分析如下。

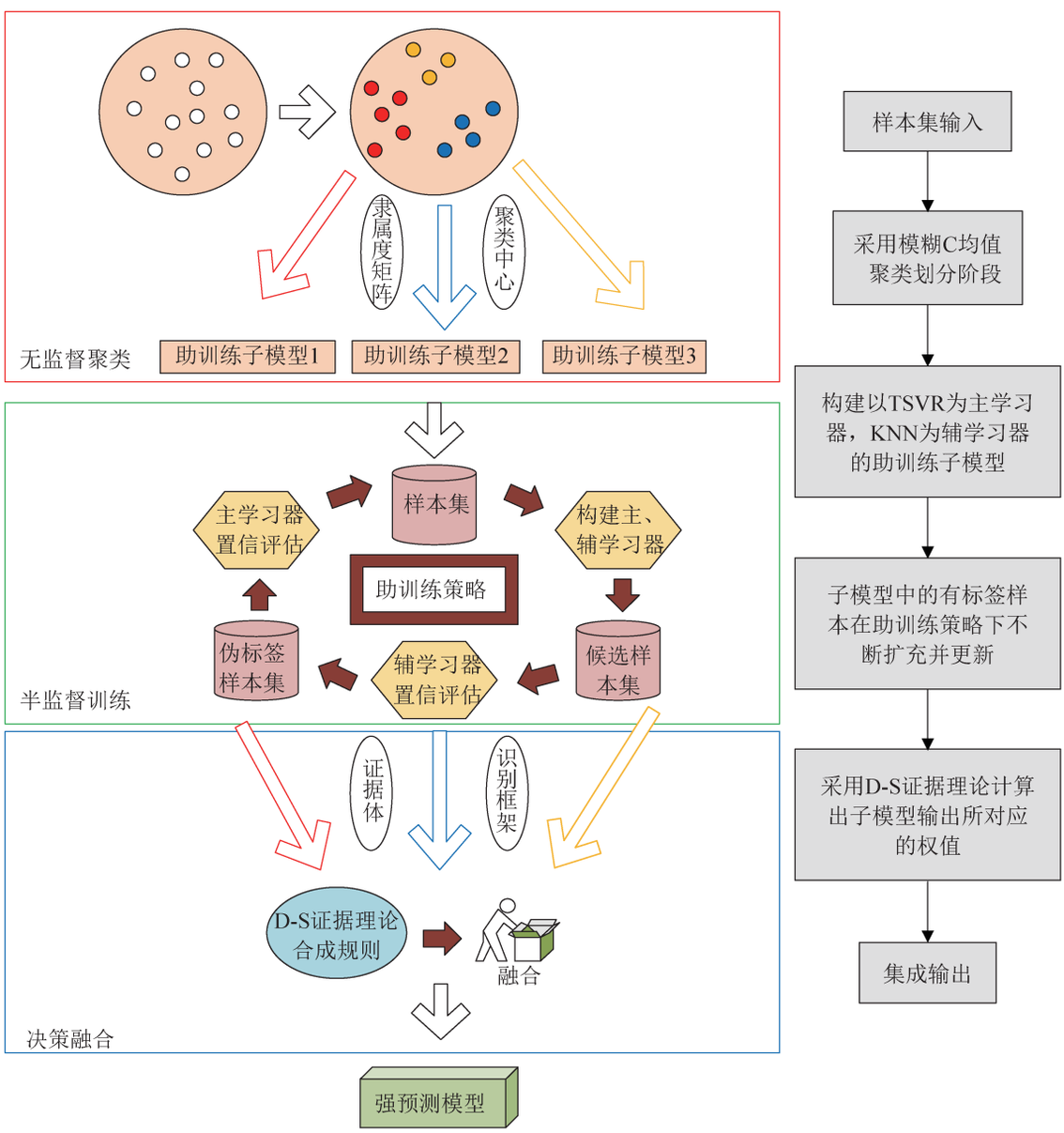


图2　所提算法的具体流程
Fig. 2　Specific flow of proposed algorithm

DOI: 10.16182/j.issn1004731x.joss.22-0976

2.1 基于FCM聚类的子模型划分

由于化工过程具有强非线性和多操作阶段耦合性的特征，因此在建立全局软测量模型之前可以利用样本在多维空间内的分散性来划分阶段。先将预处理后的样本集作为FCM的输入样本集$X=\{x_1,x_2,\cdots,x_n\}_{n\times d}$，其中$n$为样本数，$d$为样本空间维度；而后用[0,1)间的随机数初始化隶属度矩阵$\boldsymbol{U}=[u_{ik}]_{c\times n}$，$u_{ik}$表示第$i$个样本$x_i$属于第$k$个类别的隶属度，$1\leqslant i\leqslant n$，$1\leqslant k\leqslant c$，并且满足式(2)中条件[17]：

$$\begin{cases}\sum_{i=1}^{c}u_{ik}=1,\\ u_{ik}\geqslant 0,\end{cases}\quad u_{ik}\in[0,1) \tag{2}$$

再采用式(4)计算聚类中心$V=\{v_1,v_2,\cdots,v_c\}_{c\times d}$将样本数据分为$c$个子类，$c$为聚类个数；最后利用式(3)和式(4)对模糊隶属度矩阵和聚类中心迭代更新，直到式(5)目标函数J与上一次迭代之差达到迭代终止条件η或迭代次数$iter$大于最大迭代次数后停止，$\eta=1\times10^{-5}$。经此过程后，原样本集被分为3个相似性更高的子集合。

$$u_{ik}=\sum_{i=1}^{c}\left(\frac{\left\|x_k-v_i\right\|}{\left\|x_k-v_t\right\|}\right)^{-\frac{2}{\alpha-1}} \tag{3}$$

$$v_i=\sum_{k=1}^{n}\left(u_{ik}^{\alpha}x_k\right)/\sum_{k=1}^{n}u_{ik}^{\alpha} \tag{4}$$

$$J=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{\alpha}\left\|x_k-v_i\right\| \tag{5}$$

式中：$\alpha=2$为模糊系数，最大迭代次数max_iter=2 000。$\left\|x_k-v_i\right\|$为第k个样本点与第i个聚类中心的欧氏距离，主要流程如图3所示。

2.2 助训练策略下的回归模型构建

样本聚类完成后，以每一子类的样本数据作为输入，采用TSVR和KNN分别构造主、辅学习器，建立基于助训练策略的半监督学习模型。在助训练策略下，首先利用子类中的有标签样本集M来训练基于TSVR的主学习器和基于KNN的辅学习器；同时，从子类无标签样本集N中随机选择出部分无标签样本组成新的无标签样本集N'，辅学习器通过找出对应的近邻样本并对其进行置信度评估，筛选出N'中置信度最高的无标签样本组成候选样本集R；经筛选后的候选样本集R由主学习器对其中样本进行置信度评估，以筛选出置信度最高的样本添加伪标签，将其加入并更新原子类样本集，以增加每个子类中有标签样本的数量，扩充建模空间以提高子模型的学习精度和泛化性能，其主要流程如图4所示。

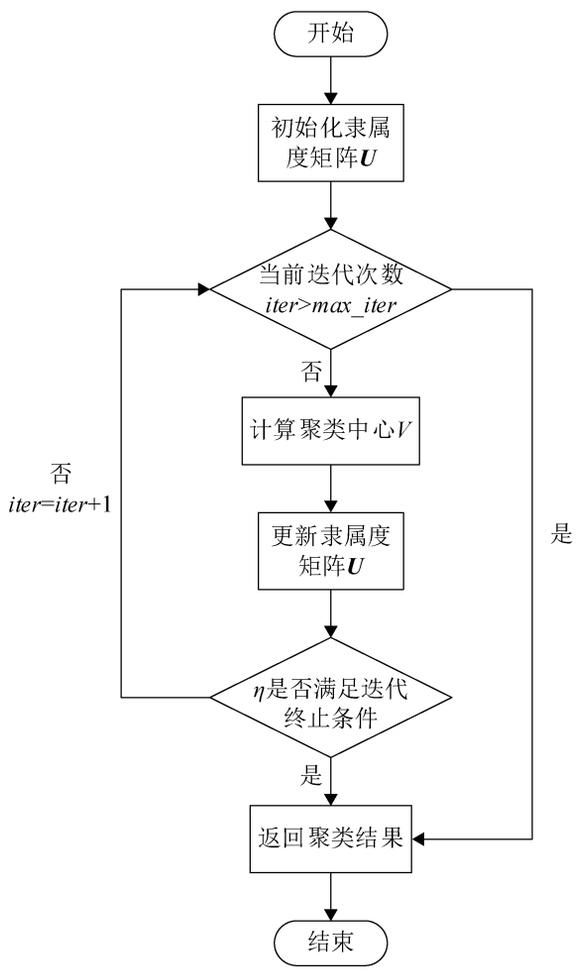


图3 FCM聚类的主要流程
Fig. 3 Main process of FMC clustering

2.3 基于证据理论的子模型融合

本文以D-S证据理论作为融合规则，需将聚类所得的c个子模型作为证据理论中的识别框架Θ，将其中某个子模型作为焦元$C_i(i=1,2,\cdots,c)$；

根据式(3)计算出子模型中样本x_1对应其类别的模糊隶属度，将其作为一个证据体，则该证据体的基本概率分配函数为$m\left(C_1x_1\right)=u_{11}$；同理对于所有样本数据$X=\left\{x_1,x_2,\cdots,x_n\right\}_{n\times d}$，可以得到$n$个证据体，并把基本概率分配函数记作$m\left(C_1x_q\right)=u_{q1}\left(q=1,2,\cdots,n\right)$；最后，利用式(1)对这些基本概率分配函数进行融合，融合后的基本概率分配函数作为子模型的基本概率分配函数输出。

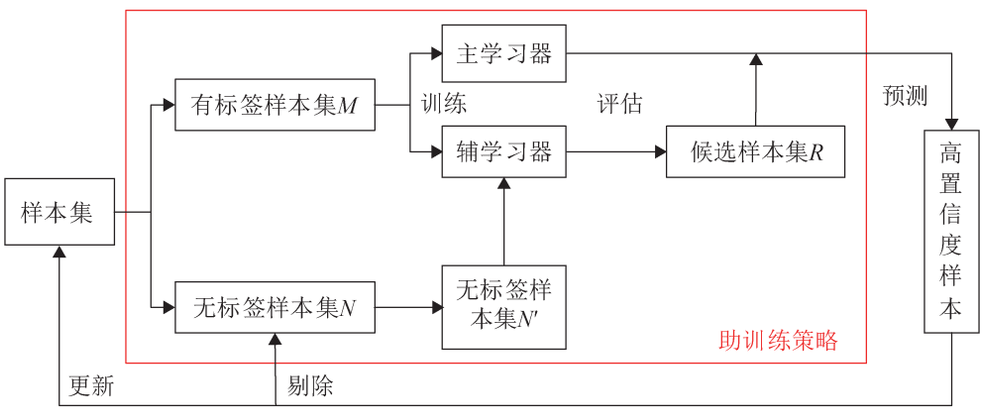


图4 助训练策略的主要流程

Fig. 4 Main process of help-training strategy

对于2.1节中的所有子模型，按照上述方法可得到c个子模型的基本概率分配函数，将其分别记作$m\left(C_1X\right),m\left(C_2X\right),\cdots,m\left(C_cX\right)$，子模型的预测输出记作$\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_c$，并将子模型的基本概率分配函数作为其加权值，融合后得到各个子模型的集成输出结果$\hat{y}$，则全局模型的输出为

$$\hat{y}=m\left(C_1X\right)\hat{y}_1+m\left(C_2X\right)\hat{y}_2+\cdots+m\left(C_cX\right)\hat{y}_c \tag{6}$$

综上分析，所提助训练策略下的多模型软测量建模方法的具体步骤如下：

step 1：对样本集中的数据进行预处理，预处理后的数据作为FCM的输入；

step 2：初始化各项参数后对样本进行无监督聚类，得到模糊隶属度矩阵；

step 3：根据子类的样本数据构建子模型，在助训练策略下不断更新子样本集；

step 4：将模糊隶属度作为D-S合成规则中的概率分配函数计算出各子模型的权值因子；

step 5：对所有子模型输出进行集成，得到最终的全局软测量模型预测结果。

3 仿真实验

3.1 数值仿真

为了分析所提算法的有效性，对式(7)所表示的非线性函数系统进行仿真研究：

$$y=\frac{\sin\sqrt{x_1^2+x_2^2}}{\sqrt{x_1^2+x_2^2}}+\varepsilon,\ x_1,x_2\in[-4\pi,4\pi] \tag{7}$$

式中：x_1、x_2为系统的输入；y为系统的输出；ε为均值为0、方差为0.01的高斯白噪声。为了便于建模和测试，随机选取1 000个样本点，并对其中30%的样本添加标签。采用FCM将样本集划分为$c=3$个操作阶段，建立以TSVR为主学习器，KNN为辅学习器的助训练子模型，其中主要参数配置为迭代次数$t=5$，每一代的无标签样本数$s=20$。为突出本文算法的性能，对3种软测量模型进行比较：

(1) 有监督TSVR全局软测量模型(TSVR)；

(2) 基于助训练策略的半监督TSVR全局软测量方法(H-TSVR)；

(3) 本文算法——助训练策略下的多模型软测

DOI: 10.16182/j.issn1004731x.joss.22-0976

量方法(HFD-TSVR)。

仿真结果采用均方根误差(root mean square error, RMSE)、相关性系数(correlation coefficient, COR)作为评估回归模型精度的指标，定义为

$$RMSE=\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i-y_i\right)^2/N} \tag{8}$$

$$COR=\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)\cdot\left(\hat{y}_i-\bar{y}_{\mathrm{p}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2\cdot\sum_{i=1}^{N}\left(\hat{y}_i-\bar{y}_{\mathrm{p}}\right)^2}} \tag{9}$$

式中：$\hat{y}_i$为实际样本的预测值；y_i为实际样本的真实值；$\bar{y}$为真实值的均值；$\bar{y}_{\mathrm{p}}$为预测值的均值；N为实际样本个数。

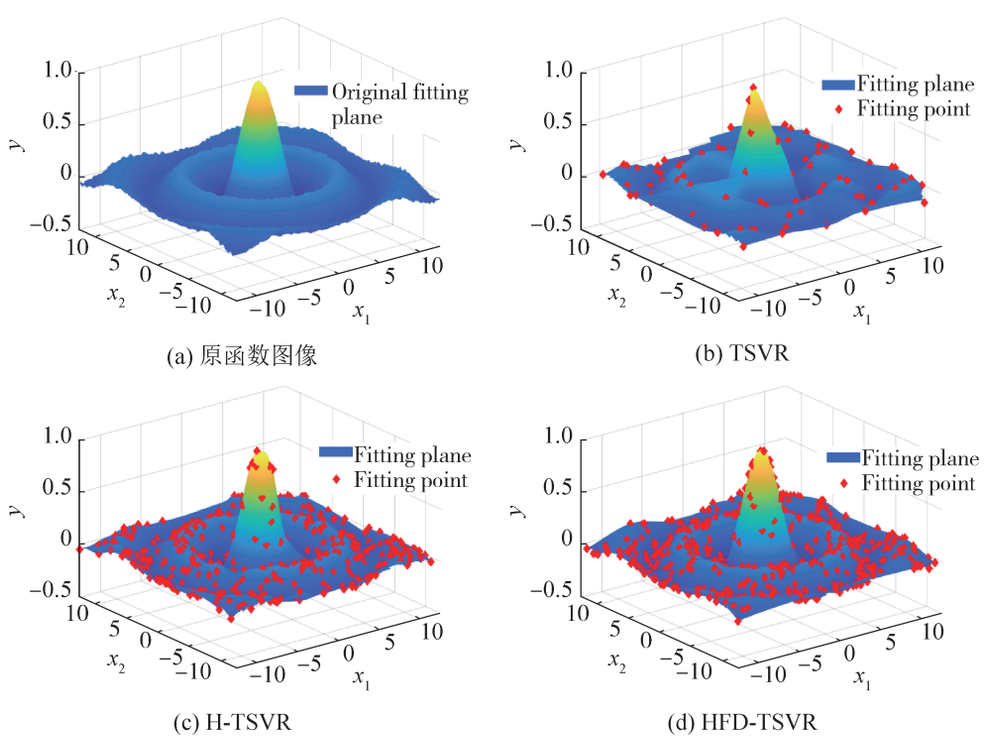
表 1 给出了 3 种算法建模后预测的 RMSE 值和 COR 值，由此看出 HFD-TSVR 的预测效果最好。因为 TSVR 算法仅仅利用了有标签样本，无法利用无标签样本中蕴含的重要信息，所以预测效果最差；相比于 TSVR，H-TSVR 算法不仅实现了对无标签样本的利用，而且淘汰了大误差样本，但由于单模型建模无法准确地描述各阶段特征，其预测精度还有待提升；而本文算法采用聚类方法对多种操作工况进行拆解，在充分利用无标签样本信息的情况下扩充建模空间，经融合策略优化子模型权值后集成输出，因此，其表现出更优的预测精度。

表 1　3 种模型的预测评价指标

Table 1　Prediction evaluation indexes of three models

模型	RMSE	COR
TSVR	0.467 1	0.797 4
H-TSVR	0.233 5	0.903 1
HFD-TSVR	**0.217 8**	**0.915 6**

为了更加直观地体现出 3 种方法的回归效果，图 5 分别给出了每个方法的实验结果。其中图 5(a)为原函数图形，图 5(b)(c)(d)中的红色点表示建模样本点。由于 TSVR 算法仅利用有标签样本，故图 5(b)中的建模样本点最少；图 5(c)由于引入助训练策略，使得建模样本点增多；而图 5(d)引入了多模型建模，以每一操作工况为中心再次扩展，更加准确地描述出不同阶段之间的差异和阶段之内的联系，提高了预测精度，其样本点所拟合的平面最接近原函数，表现出最优的预测性能。



(a) 原函数图像　(b) TSVR

(c) H-TSVR　(d) HFD-TSVR

图 5　3 种方法的预测结果对比

Fig. 5　Comparison of prediction results of three methods

3.2 脱丁烷塔仿真

脱丁烷塔是石油炼制过程的重要组成部分，用于脱硫和石脑油裂解[18]。图6为脱丁烷塔的工艺流程示意图。在工业过程中，需要从石脑油中去除丁烷，也就是使塔底的丁烷含量最小，因此需要对丁烷浓度实时测量。但丁烷浓度通常很难直接检测并会产生一定的测量延迟，需建立软测量模型预测丁烷浓度[19-20]。所选取的输入辅助变量如表2所示。

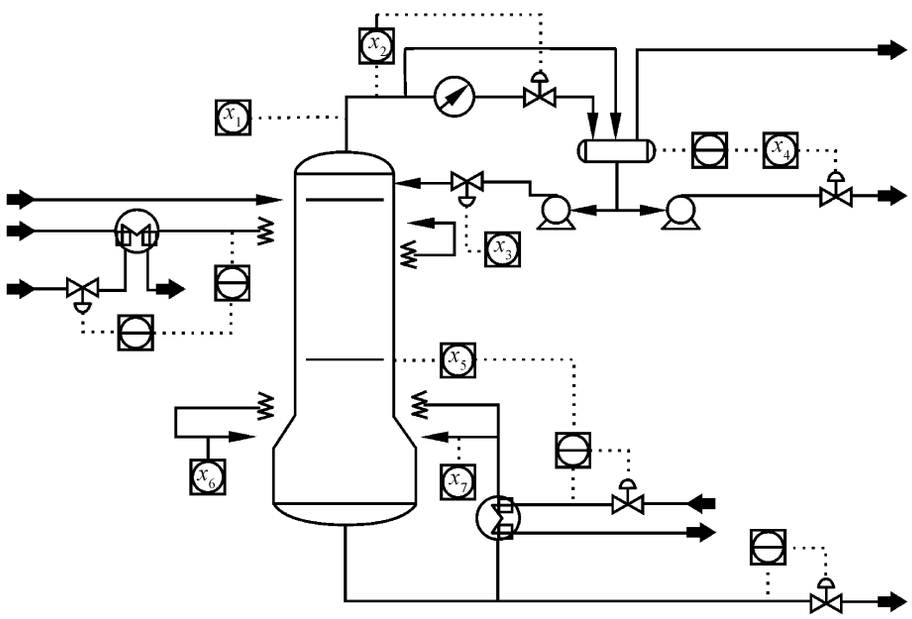


图6 脱丁烷塔工艺流程
Fig. 6 Process flow of debutanizer

表2 辅助变量选择
Table 2 Selection of auxiliary variables

主要变量	变量名称
x_1	顶层温度
x_2	顶层压力
x_3	回流流量
x_4	流向下个过程的流量
x_5	第六塔板温度
x_6	塔底温度1
x_7	塔底温度2

工业过程数据来源于石脑油裂解过程中的实时采样，共2 394组样本。为了模拟复杂工业过程中有标签样本少，无标签样本多的情况，使有标签样本在全部样本中的比例为30%；然后使用所提算法建模预测，其中主要参数配置为迭代次数 $t=10$，每一代的无标签样本数 $s=100$，模型内其他参数选择同数值仿真实验。为体现本文算法的性能，对5种软测量模型进行比较：

(1) 有监督TSVR全局软测量模型(TSVR)；

(2) 基于自训练策略的半监督TSVR全局软测量方法(S-TSVR)；

(3) 基于助训练策略的半监督TSVR全局软测量方法(H-TSVR)；

(4) 基于助训练策略的多模型(无D-S融合)软测量方法(HF-TSVR)；

(5) 本文算法——基于助训练策略的多模型软测量方法(HFD-TSVR)。

由表3可知不同建模方法的预测性能。将有监督TSVR方法和S-TSVR、H-TSVR方法对比可以发现，后两种建模方法的RMSE值更小，COR值更大，说明通过引入半监督学习的思想，充分利用了无标签样本，扩充了模型空间，改良了预测性能。而助训练相比与自训练的优点则在于对无标签样本的筛选，从而淘汰误差样本。相比于S-TSVR和H-TSVR方法，同样引入半监督学习思

DOI: 10.16182/j.issn1004731x.joss.22-0976

想的基础上又结合了多模型集成的HF-TSVR和HFD-TSVR方法具有更优的预测性能。这是因为FCM算法可以最小化类间的相似性，最大化类内的相似性，更准确地描述不同操作阶段的差异性，体现了模型对多操作阶段的解耦。相比于其他4种方法，本文所提的HFD-TSVR结合了多模型集成和半监督学习的思想，并利用D-S证据理论决策优化融合权值，进一步提高模型的预测精度，其预测结果的RMSE值最小，COR值最大，拥有最优的预测性能。

表3　5种模型的预测评价指标
Table 3　Prediction evaluation indexes of five models

模型	RMSE	COR
TSVR	0.141 2	0.510 4
S-TSVR	0.137 4	0.589 7
H-TSVR	0.137 1	0.594 4
HF-TSVR	0.128 9	0.701 2
HFD-TSVR	**0.112 8**	**0.720 5**

为了更直观地对比各类方法的预测效果，选取3个最具代表性的方法TSVR、H-TSVR和HFD-TSVR分别绘制其前1 197个样本在30%标签率下对丁烷浓度真实值的跟踪效果对比图和预测误差对比图。图7为3种方法对真实值的跟踪效果对比图，其中黑色45°对角线为基准线，样本点越靠近基准线，其预测精度越高。图8为3种方法对真实值的预测误差对比图，误差值越靠近基准线，其预测越准确。

由图7和图8可知，本文方法HFD-TSVR的预测效果最好。HFD-TSVR方法的优越性在于考虑利用无标签样本的同时，兼顾准确性与冲突性，凭借助训练策略中的两次置信度评估，很大程度地提高了无标签样本的利用率，扩充了样本空间，提升了模型的预测精度和泛化性，并有效防止模型退化；此外，多模型集成的软测量建模方法对于复杂工业过程中的强非线性、多操作阶段耦合性特征尤为有效，同时也保证了子样本集内一定的连续性，为建立准确的助训练子模型打下重要基础。图9是3个子模型以及融合后全局模型的RMSE值和COR值。融合后得到最终模型的预测效果优于单一的子模型，更加体现出多模型建模的有效性。

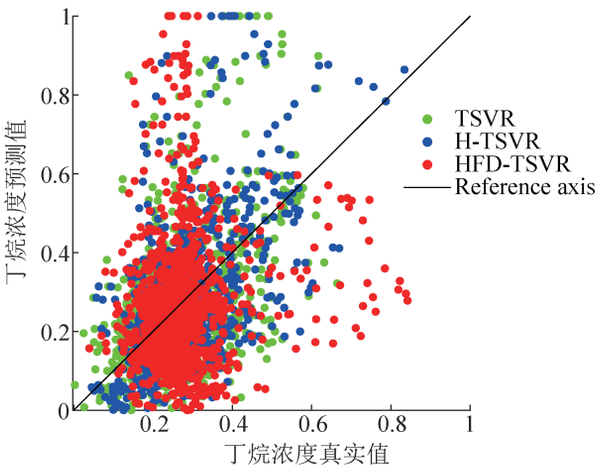


图7　丁烷浓度预测散点图
Fig. 7　Scatter plot of butane concentration prediction

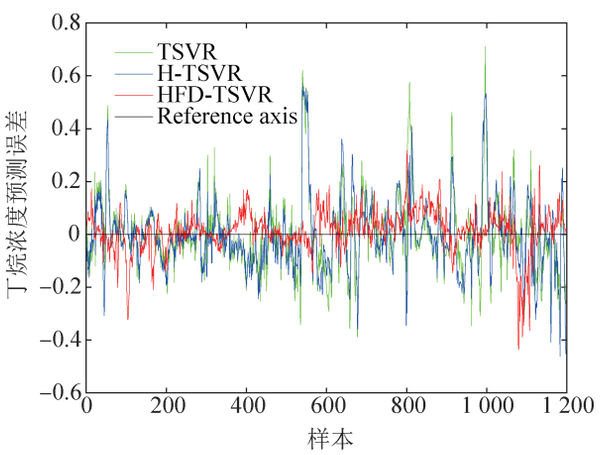


图8　丁烷浓度预测误差图
Fig. 8　Predicted errors of butane concentration

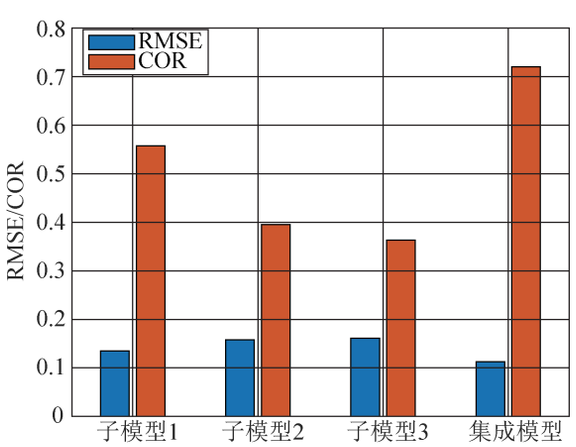


图9　各模型预测指标
Fig. 9　Prediction indexes of each model

4 结论

针对工业过程中的强非线性和多操作阶段耦合性特征以及有标签样本数量少的问题，本文提出了助训练策略下的多模型软测量建模方法。所提方法采用FCM对多操作工况进行剖析，定量描述样本之间的模糊关系，充分挖掘了样本间的相似性信息，确保每一阶段样本更具连续性以便建立助训练策略下的子模型。基于主辅协同学习器形成助训练策略，尽可能地剔除误差较大样本并扩充建模空间，避免了数据信息缺失导致的模型退化问题。而后又凭借D-S证据理论的聚焦特性，将基于子模型所构造的证据体进行融合，可以准确地优化权值，提高了软测量模型的预测性能。通过在一个数值例子和一个脱丁烷塔实际工业过程的仿真实验表明，与其他4种方法相比，助训练策略下的多模型软测量建模方法预测精度较高，验证了所提方法的优越性和有效性。由于本文结合了半监督学习与集成学习来构建软测量模型，后续还可以考虑如何更深层地挖掘数据特征信息以及拓展子模型的多样性等方面进一步研究。

参考文献：

[1] Yuan Xiaofeng, Gu Yongjie, Wang Yalin, et al. A Deep Supervised Learning Framework for Data-driven Soft Sensor Modeling of Industrial Processes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(11): 4737-4746.

[2] Wang Bo, Wang Xingyu, He Mengyi, et al. Study on Multi-model Soft Sensor Modeling Method and Its Model Optimization for the Fermentation Process of Pichia pastoris[J]. Sensors, 2021, 21(22): 7635.

[3] Shao Weiming, Ge Zhiqiang, Song Zhihuan, et al. Nonlinear Industrial soft Sensor Development Based on Semi-supervised Probabilistic Mixture of Extreme Learning Machines[J]. Control Engineering Practice, 2019, 91: 104098.

[4] Mathias M Adankon, Mohamed Cheriet. Help-training for Semi-supervised Support Vector Machines[J]. Pattern Recognition, 2011, 44(9): 2220-2230.

[5] Zhang Lei, Yang Lin, Ma Tianwu, et al. A Self-training Semi-supervised Machine Learning Method for Predictive Mapping of Soil Classes with Limited Sample Data[J]. Geoderma, 2021, 384: 114809.

[6] 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.

[7] Mathias M Adankon, Mohamed Cheriet. Help-training for Semi-supervised Discriminative Classifiers. Application to SVM[C]//2008 19th International Conference on Pattern Recognition. Piscataway, NJ, USA: IEEE, 2008: 1-4.

[8] Mathias M Adankon, Mohamed Cheriet. Help-training Semi-supervised LS-SVM[C]//2009 International Joint Conference on Neural Networks. Piscataway, NJ, USA: IEEE, 2009: 49-56.

[9] Gan Haitao, Sang Nong, Huang Rui, et al. Using Clustering Analysis to Improve Semi-supervised Classification[J]. Neurocomputing, 2013, 101: 290-298.

[10] 钟伟民, 李杰, 程辉, 等. 基于FCM聚类的气化炉温度多模型软测量建模[J]. 化工学报, 2012, 63(12): 3951-3955.
Zhong Weimin, Li Jie, Cheng Hui, et al. A Soft Sensor Multi-modeling for Furnace Temperature of Gasifier Based FCM Clustering[J]. CIESC Journal, 2012, 63(12): 3951-3955.

[11] 嵇小辅, 张翔. 基于FCM与集成高斯过程回归的赖氨酸发酵软测量[J]. 智能系统学报, 2015, 10(1): 156-162.
Ji Xiaofu, Zhang Xiang. Soft Measurement of Lysine Fermentation Based on FCM and Integrated Gaussian Process Regression[J]. CAAI Transactions on Intelligent Systems, 2015, 10(1): 156-162.

[12] 李卫, 杨煜普, 王娜. 基于核模糊聚类的多模型LSSVM回归建模[J]. 控制与决策, 2008, 23(5): 560-562, 566.
Li Wei, Yang Yupu, Wang Na. Multi-model LSSVM Regression Modeling Based on Kernel Fuzzy Clustering [J]. Control and Decision, 2008, 23(5): 560-562, 566.

[13] Yuan Xiaofeng, Qi Shuaibin, Wang Yalin, et al. A Dynamic CNN for Nonlinear Dynamic Feature Learning in Soft Sensor Modeling of Industrial Process Data[J]. Control Engineering Practice, 2020, 104: 104614.

[14] Wang Hui, Guo Lili, Dou Zheng, et al. A New Method of Cognitive Signal Recognition Based on Hybrid Information Entropy and D-S Evidence Theory[J]. Mobile Networks and Applications, 2018, 23(4): 677-685.

[15] 刘准钆, 程咏梅, 潘泉, 等. 基于证据距离和矛盾因子的加权证据合成法[J]. 控制理论与应用, 2009, 26(12):

DOI: 10.16182/j.issn1004731x.joss.22-0976

1439-1442.
Liu Zhunga, Cheng Yongmei, Pan Quan, et al. Combination of Weighted Belief Functions Based on Evidence Distance and Conflicting Belief[J]. Control Theory & Applications, 2009, 26(12): 1439-1442.

[16] Xiong Weili, Zhang Wei, Xu Baoguo, et al. JITL Based MWGPR Soft Sensor for Multi-mode Process with Dual-updating Strategy[J]. Computers & Chemical Engineering, 2016, 90: 260-267.

[17] Wang Liangjie, Kazuhide Sawada, Shuji Moriguchi. Landslide Susceptibility Analysis with Logistic Regression Model Based on FCM Sampling Strategy[J]. Computers & Geosciences, 2013, 57: 81-92.

[18] Meng Yanmei, Lan Qiliang, Johnny Qin, et al. Data-driven Soft Sensor Modeling Based on Twin Support Vector Regression for Cane Sugar Crystallization[J]. Journal of Food Engineering, 2019, 241: 159-165.

[19] Iratxe Niño-Adan, Itziar Landa-Torres, Diana Manjarres, et al. Soft-sensor for Class Prediction of the Percentage of Pentanes in Butane at a Debutanizer Column[J]. Sensors, 2021, 21(12): 3991.

[20] Hiromasa Kaneko, Kimito Funatsu. Adaptive Soft Sensor Based on Online Support Vector Regression and Bayesian Ensemble Learning for Various States in Chemical Plants[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 137: 57-66