

DOI: 10.19650/j.cnki.cjsi.J2107920

基于时间近邻拉氏正则的多工况软测量回归*

徐志强,任密峰,程 兰,李 荣,阎高伟
(太原理工大学电气与动力工程学院 太原 030024)

摘 要:针对流程工业中,工况改变导致传统软测量模型预测精度下降的问题,考虑到工业数据连续性、序列性、多重共线性、数据量庞大等特殊对模型建立的影响,提出一种基于时间近邻拉普拉斯正则的多工况软测量回归模型框架。针对工业数据的多重共线性,回归框架采用非线性迭代偏最小二乘方法,同时引入域适应正则项改善工况变化对模型的影响,在此基础上,提出时间近邻拉普拉斯正则项,能够在映射过程中保持住数据的序列结构,并且大幅度减少模型训练时间以满足工业实时性要求。实验部分以三聚氰胺聚合过程多工况数据集为例,对本文模型的预测有效性以及减少训练时间的有效性进行了实验和分析。结果表明,与传统方法偏最小二乘回归相比,当目标工况为工况1到工况4时,本文方法使平均均方根误差分别降低了30.3%、31.4%、29.3%和24.1%。且相较于传统全连接法,时间近邻法构建拉普拉斯正则项能够使得四个工况上模型训练时间分别降低14.11、1.01、26.43和0.71 s,表明该模型的预测准确性和训练时间均得到有效改善。

关键词: 流程工业;过程数据;时间近邻拉普拉斯正则;多工况;软测量回归模型

中图分类号: TH7 **文献标识码:** A **国家标准学科分类代码:** 510.80

Multi-conditions soft sensor regression based on the time-nearest neighbor Laplacian regularization

Xu Zhiqiang, Ren Mifeng, Cheng Lan, Li Rong, Yan Gaowei

(College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: In the process industry, the change of working condition results in the decrease of the prediction accuracy of traditional soft-sensing models. Take the impact of industrial data continuity, sequence, multicollinearity, and huge amount of data on the model establishment into account, a multi-conditions soft sensor regression model framework based on the time-nearest neighbor Laplacian regularization is proposed. To solve the multicollinearity of industrial data, the proposed regression framework utilizes the nonlinear iterative partial least squares method. Meanwhile, the domain adaptation regular term is introduced to mitigate the influence of the change of working conditions on the model. On this basis, the time nearest neighbor Laplacian regular term is proposed, which can maintain the sequence structure of the data during the mapping process. And the model training time is greatly reduced to meet the industrial real-time requirement. In the experiment, the multi-conditions data set of the melamine polymerization process is taken as an example. The results show that compared with the traditional method of partial least squares regression, when the target conditions are conditions 1 to 4, the method in this paper reduces the average root mean square error by 30.3%, 31.4%, 29.3% and 24.1%, respectively. And compared using with the traditional function of total connecting, using the function of time-nearest neighbor connecting to construct the Laplacian regularization could deduce the training time of the four working conditions by 14.11, 1.01, 26.43, 0.71 s respectively, and indicated that the accuracy and the training time could be improved.

Keywords: process industry; industrial data; time nearest neighbor Laplace regularity; multi-conditions; soft sensor regression modeling

收稿日期:2021-05-13 Received Date: 2021-05-13

* 基金项目:国家自然科学基金(61973226, 62073232)、山西省重点研发计划项目(201903D121143)、山西省科技重大专项(20181102017)资助

0 引言

因测量成本或物理条件限制等原因,工业过程中的某些关键参量难以做到实时性地直接测量,增加了工业控制的难度,是工业智能化进程上的一个挑战。软测量技术利用易于测量的辅助变量建立数据驱动模型对关键参量进行预测,是一种有效的解决手段^[1-3]。偏最小二乘回归(partial least squares regression, PLSR)^[4]、线性判别分析(linear discriminative analysis, LDA)、支持向量机(support vector machine, SVM)^[5-7]等传统软测量模型的前提假设是训练数据和测试数据属于同一模态。但由于工作点、工作环境、设备损耗等原因,常常导致工况的改变。不同工况的数据具有模态的差异,分属于不同的域,此时源域上预先训练好的模型在目标域上的预测精度会有不同程度的下降。

域适应^[8-11](domain adaptation, DA)旨在减小源域和目标域分布差异,从而可以向目标域迁移源域上学习到的知识,将历史工况作为源域,目标工况作为目标域,构建历史工况到目标工况的域适应模型为解决多工况问题提供了不错的思路。近年来,域适应模型构建方法主要有无监督和有监督两种策略。第一种策略旨在求得一个子空间,历史工况数据和目标工况数据投影到子空间后具有相近的分布,再利用子空间中的投影向量建立回归模型^[12]。该策略能够得到求解域适应问题的最优闭式解,但是存在一个较大的缺陷,即因为没有和关键质量变量建立相关性联系,导致得到的子空间的变量缺少对关键质量变量的解释。第二种策略旨在设计一种具备域适应功能的正则项,对建立的回归模型进行约束^[13]。本质上是将数据映射到重构的标签空间中,历史工况和目标工况数据在此空间上的投影特征既具备对关键质量变量的解释性,同时具有相近的分布。在历史工况未标注的情况下,无监督提取对齐特征的方法是必要的,但在实际的工业过程中,往往存储有大量已标注的历史数据,因此,在流程工业背景下,第二种策略更为适用。当前,大量研究集中于分类任务,但由于工业过程数据存在以下3点特殊性,这些研究并不完全适用:1)过程数据无法满足紧致性准则假设,关键质量变量的连续性导致无法建立重构类别空间,数据映射到标签空间中无法满足正交性要求;2)流程工业过程采集的数据往往具有序列结构,数据点之间具有时间上的相关性,具有相对距离和方向两个几何属性;3)工业过程中,传感器采集的数据点比较密集,随着工艺过程的进行,存储的数据量往往非常庞大。

针对工业过程数据的第1)点特殊性,将原始数据映射到一个隐空间中,使数据得到满足要求的特征表达是

必要的^[14]。该隐空间中的投影特征在满足正交要求的情况下,需要具备对关键质量变量的解释性,且两个域的数据的投影分布尽可能接近。同时,针对第2)点特殊性,投影时对数据点序列结构的保持也是非常重要的。文献[15-19]提出了等度量映射(isometric feature mapping, ISomap)、局部线性嵌入(locally linear embedding, LLE)等结构保持策略,在投影过程中保持住数据的潜在流形结构。但是这些流形降维方法均为无监督方法,而文献[20-22]提出的拉普拉斯正则项能够自然地嵌入到有监督或者半监督的回归框架中,达到结构保持的作用。上述流形方法均需要构建数据的邻接矩阵,当前构建邻接矩阵的方式有3种:1)采用 ϵ 阈值-近邻法;2)k-近邻法;3)全连接法。根据过程数据的第一点特殊性,由于不存在紧致性准则,使用k-NN等聚类方法无法得到一致性结果。第1)和第3)种方法均要求取所有样本之间的距离,若假设两个域中数据集的数据点个数分别为 n_s 和 n_t ,则所需的计算次数为 $n_s^2 + n_t^2$,所需计算时间随着数据点个数呈现平方趋势的增长,由于过程数据的第3点特殊性,模型训练所需时间可能超过工业过程即时性要求。针对过程数据的第2点特殊性和第3点特殊性,本文做出时间近邻假设:对于过程数据,数据点只与其时间点前后的 c 单位采样时间邻域的数据产生近邻关系。这种假设具备一定合理性,对于两个不同时段采集的数据点,其距离计算结果可能比较小,但是由于其分属于工艺过程的不同阶段,在实际意义上它们是非近邻的。根据时间近邻假设,本文提出时间近邻拉普拉斯正则项,该正则项能够在建模过程中保持数据关键结构信息,且能够大幅改善模型训练时间。

综上所述,需要将数据映射到一个隐空间中,隐空间中的变量需要满足对关键质量变量的解释性、可迁移性、保序性和正交等约束。但是,该映射的求解是一个斜投影问题,难以求解其闭式解,且工业数据普遍存在变量间多重共线性,可能求解出病态的回归系数。受PLSR^[23]和域不变迭代偏最小二乘(domain-invariant iterative partial least squares, DIPALS)^[24]的启发,采用非线性迭代偏最小二乘(nonlinear iterative partial least squares, NIPALS)方法^[25],迭代地求解隐空间中的满足上述要求的特征表达,能够有效解决上述问题,最终构建出一种适用于过程数据的基于时间拉普拉斯正则的多工况软测量回归模型(multi-conditions soft sensor regression model based on time-nearest neighbor Laplacian regularization, TNN-LR-MR)。

1 问题定义

沿用文献[14, 26]的表示,将域定义为数据分布

$P \in \mathbb{R}^k$ 和标签生成函数 $l: \mathbb{R}^k \rightarrow \mathbb{R}$ 的成对形式 $\langle P, l \rangle$ 。记源域 $D_S = \langle P, l \rangle$ 和目标域 $D_T = \langle Q, l \rangle, P \neq Q$ 。假设有源域数据集 $X_S \in \mathbb{R}^{n_S \times k}$ 采样于分布 P, n_S 为样本数, k 为样本特征维度, 其具有连续标签 Y_S 。同时, 有目标域数据集 $X_T \in \mathbb{R}^{n_T \times k}$ 采样于分布 Q, n_T 为样本数。一个经典的无监督域适应回归任务为寻找一个映射 $h: \mathbb{R}^k \rightarrow \mathbb{R}$ 使得目标域上的预测损失期望最小:

$$E_Q[|h - l|] = \int_{\mathbb{R}^k} |h - l| dQ \quad (1)$$

假设 $h = f \circ g$, 表示 h 是 f 和 g 的复合函数, 其中 $g: \mathbb{R}^k \rightarrow \mathbb{R}^L, f: \mathbb{R}^L \rightarrow \mathbb{R}, \mathbb{R}^k$ 表示 K 维源空间, \mathbb{R}^L 表示 L 维隐空间。

本文框架中, 假设投影矩阵均为线性算子, 即 $g(X) = Z = XA, A \in \mathbb{R}^{k \times L}, f(Z) = Z\beta, \beta \in \mathbb{R}^L$ 。不同工况的数据属于不同的域, 将历史工况设置为源域, 目标工况设置为目标域, 关键质量变量的标注信息作为标签。

2 相关理论

2.1 域适应正则项

针对工况差异对软测量回归模型结果的影响, 引入域适应正则项, 在历史工况和目标工况的数据投影到隐空间的过程中, 最小化两个工况数据之间的统计量差异, 从而使得投影到隐空间中的变量进行了对齐, 改善模型结果。最大均值差异度量 (maximum mean discrepancy, MMD)^[27-29], 度量在可分希尔伯特空间中源域数据与目标域数据总体均值差异, 最小化 MMD 是一种常用且有效的域适应正则策略。MMD 要求给定的函数集是充分且完备的, 仅当函数集为再生核希尔伯特空间上的单位球时可以满足要求。但是在投影矩阵均为线性算子的假设下, MMD 距离无法作用。论文[26, 30]提出了一种基于最小化二阶统计量差异的无监督降维对齐策略 (correlation alignment, CORAL), 并对其有效性进行了论证。但是降维过程中需要约束变量和标签之间的相关性, 因此需要将该策略作为域适应正则项嵌入到回归模型框架中, 对隐空间中的变量进行对齐约束。域适应正则项可表示为:

$$R_D = \|\text{Cov}_{Z_S} - \text{Cov}_{Z_T}\|_F^2 = \left\| \frac{1}{n_S - 1} A^T X_S^T X_S A - \frac{1}{n_T - 1} A^T X_T^T X_T A \right\|_F^2 = \left\| A^T \left(\frac{1}{n_S - 1} X_S^T X_S - \frac{1}{n_T - 1} X_T^T X_T \right) A \right\|_F^2$$

记 $A = \frac{1}{n_S - 1} X_S^T X_S - \frac{1}{n_T - 1} X_T^T X_T$, 则有 $R_D = \|A^T A A\|_F^2$ 。

2.2 时间近邻拉普拉斯正则项

只考虑域适应正则项的映射过程, 可能会导致训练模型畸变数据的序列结构。针对这一情况, 在映射过程中应当尽可能保持两个工况数据的序列结构。文献[17]里指出, 在平滑性假设下, 拉普拉斯正则化能够提取出边缘分布的几何结构, 拉普拉斯正则项的一般表达式为:

$$R_L = \sum_{v=S, T} \sum_{i, j \in D_v} w_{ij}^v (g(x_i^v) - g(x_j^v)) \quad (3)$$

其中, $v = S$ 表示历史工况, $v = T$ 表示目标工况。 w_{ij}^v 为 $\{v | v = S, T\}$ 工况的近邻矩阵 W 的第 i 行第 j 列的元素。由于过程数据无紧致性准则, 所以一般情况下 k -NN 近邻法无法得到一致的聚类结果, 而使用 ε 阈值-近邻法和全连接法均需要 $n_S^2 + n_T^2$ 次运算。根据引言中所述的时间近邻假设, 本文构建一种适用于过程数据的时间近邻矩阵 W , 其元素为:

$$w_{ij} = \begin{cases} e^{-\text{dis}(x_i, x_j)}, & t_{x_j} \subset U_c(t_{x_i}) \\ 0, & \text{其他} \end{cases} \quad (4)$$

其中, $t_{x_j} \subset U_c(t_{x_i})$ 表示 t_{x_j} 包含于 t_{x_i} 的 c 单位采样时间邻域内。若数据集中每个样本均为单位采样时间得到, 则当 $x_j = \{x_k | i - c \leq k \leq i + c\}$ 时, 认为 x_j 为 x_i 的邻域。 $\text{dis}(x_i, x_j) = \|x_i - x_j\|^2$ 。通过推导, 可得:

$$R_L = \sum_{v=S, T} \sum_{i, j \in D_v} g(x_i^v) L_{ij}^v g(x_j^v) = A^T \left(\sum_{v=S, T} X_v^T L_v X_v \right) A \quad (5)$$

其中, $L_v = D_v - W_v$ 为拉普拉斯矩阵; D_v 为度矩阵, 其对角元素为 $D_{ii}^v = \sum_i w_{ij}$, 其他元素为 0。若记 $M = \sum_{v=S, T} X_v^T L_v X_v$, 则 $R_L = A^T M A$ 。

$$K(n_S, n_T, c) = \begin{cases} 2(n_S + n_T)c - 2c^2, & c \leq n_S \\ n_S^2 + 2n_T c - c^2, & n_S < c \leq n_T \\ n_S^2 + n_T^2, & n_T < c \end{cases} \quad (6)$$

假设历史工况数据集样本数为 n_S , 目标工况数据集样本数为 n_T , 并后续表示方便, 假设 $n_S \leq n_T$ 。则在采样时间邻域设置为 c 单位的情况下, 需要的计算量 K 与样本数以及时间变量 c 的数值之间的关系如式(6)所示。

当 $c \geq \max\{n_S, n_T\} = n_T$ 时, 在两个工况上, 时间变量 c 对应的近邻矩阵等同于全连接阵。由式(6)可以看出, 计算全连接阵的时间复杂度是 $O(n^2)$, 而当 $c \leq \min\{n_S, n_T\} = n_S$ 时, 其计算时间复杂度仅为 $O(n)$, 由此可以得出, 在有庞大训练数据量的情况下, 该策略能够有效改善模型的训练时间。比较而言, 全连接法能够得到数据集全局的结构, 如果采用欧几里得距离这类线性距离, 则全局结构保持为刚性, 而时间近邻法仅能保持每个样本的近邻局部结构, 未考虑全局结构上的畸变, 会带来精确度一定程度的降低, 精度和实时性的平衡需要进行衡量。但是由于过程数据的序列性, 近邻样本的选择是否对称会影响结构保持效果, 因此全连接法并不是最佳选择。以上结论在实验部分将得到探讨。

3 基于时间近邻拉普拉斯正则的多工况软测量回归模型

为了在映射过程中尽可能减小历史工况与目标工况之间的差异以及保持数据的序列结构,并保持对标签的解释性,在 NIPALS 中引入域适应正则项和时间近邻拉普拉斯正则项。标准 NIPALS 采用迭代的方法求解隐空间中满足要求的各个投影向量,包含了 4 个基本步骤,即:初始化、映射、回归和直交补。为了适应 NIPALS 的迭代形式的求解方法,需要将域适应正则项和时间近邻拉普拉斯正则项改写成式(7)所示的向量形式并引入映射步骤,以达到对映射过程的约束。

$$\alpha^* = \arg \min_{\alpha, \alpha^T \alpha = I} \|X_S - Y_S \alpha^T\|_F^2 + \gamma \operatorname{tr}(\alpha^T \Lambda \alpha) + \eta \operatorname{tr}(\alpha^T M \alpha) \quad (7)$$

其中, α 为某次迭代过程中的投影向量, γ, η 分别为域适应对齐正则项与时间近邻拉普拉斯正则项的正则系数,为了平衡各项量级上的误差, γ, η 分别设计为:

$$\gamma = \frac{\|X_S - Y_S \alpha_0^T\|_F^2}{\operatorname{tr}(\alpha_0^T \Lambda \alpha_0)}$$

$$\eta = \frac{\|X_S - Y_S \alpha_0^T\|_F^2}{\operatorname{tr}(\alpha_0^T M \alpha_0)} \quad (8)$$

$$\alpha_0 = \arg \min_{\alpha} \|X_S - Y_S \alpha^T\|_F^2 = \frac{X_S^T Y_S}{Y_S^T Y_S}$$

利用拉格朗日乘子法计算式(7)的最优解,结果如下:

$$\alpha^* = \left(\frac{\gamma \Lambda + \eta M}{Y_S^T Y_S} + I \right)^{-1} \frac{X_S^T Y_S}{Y_S^T Y_S} \quad (9)$$

基于时间拉普拉斯正则的多工况软测量回归模型的算法伪代码如算法 1 所示。

4 实验结果与分析

实验部分采用三聚氰胺间歇多工况过程数据集,在 4 种不同配方的三聚氰胺树脂的聚合过程中,光纤探头持续检测近红外吸光度光谱数据,预测聚合物开始变得浑浊时的浊点温度(以 $^{\circ}\text{C}$ 表示)。单域实验中,训练集和测试集的分割采用随机抽样方法。跨域实验中,目标工况中训练集和测试集的分割同样采用随机抽样方法。为消除随机性,实验均重复 10 次取均值作为最终结果。4 个过程的数据集样本数如表 1 所示。

为了评价方法的预测性能,采用均方根误差(root mean square error, RMSE)作为预测精度的评价指标,其计算公式如式(10)所示。

算法 1: 基于时间拉普拉斯正则的多工况软测量回归算法

Input: 历史工况数据 X_S , 目标工况数据 X_T , 历史工况标签 Y_S , 隐空间维度 L , 时间变量 c

Output: 回归系数 $\beta \in \mathbf{R}^k$, 目标工况数据预测标签 \hat{Y}

初始化:

$$X_S = X_S - E[X_S], X_T = X_T - E[X_T],$$

$$Y_S = Y_S - E[Y_S]; P = [], A = [], r = []$$

for $i = 1; L$

映射:

按照式(8)计算 α_0 以及正则化系数 γ, η

按照式(9)计算 α^* , 并进行范数标准化 $\alpha^* = \alpha^* / \|\alpha^*\|$

回归:

$$t_s = X_S \alpha^*, t_r = X_T \alpha^*$$

$$\tilde{r} = (t_s^T t_s)^{-1} t_s^T Y_S$$

$$p_s^T = (t_s^T t_s)^{-1} t_s^T X_S, p_r^T = (t_r^T t_r)^{-1} t_r^T X_T$$

直交补:

$$X_S = X_S - t_s p_s^T, X_T = X_T - t_r p_r^T, Y_S = Y_S - \tilde{r} t_s$$

$$P = [P, p_s], A = [A, \alpha^*], r = [r, \tilde{r}]$$

end

$$\beta = A(P^T A)^{-1} r$$

$$\hat{Y}_T = X_T \beta + E[Y_S]$$

表 1 三聚氰胺数据集不同工况数据样本数

Table 1 The number of samples of melamine data set under different working conditions

工况编号	工况 1	工况 2	工况 3	工况 4
样本数	3 032	733	3 890	472

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (10)$$

其中, N 为测试样本的数量; y_i 为第 i 个样本的真实值; \hat{y}_i 为第 i 个样本的估计值。

4.1 验证时间近邻拉普拉斯正则项有效性

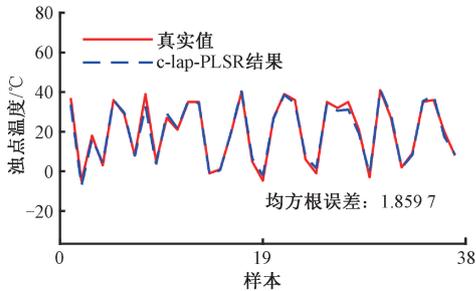
为了验证时间近邻拉普拉斯正则项对回归模型的提升效果,第一个实验分别对三聚氰胺数据集的 4 个工况数据进行单域实验。将传统 PLSR 模型与在 PLSR 模型基础上嵌入时间近邻拉普拉斯正则项的单域回归模型(简称 $c\text{-Lap-PLSR}$)进行实验比较,两种模型均取相同的隐变量个数,本实验设定数据集的内在维度为模型的隐变量个数。不同工况实验中, c 均取 100。10 次实验的均方根误差结果平均值如表 2 所示。

表 2 不同工况单域实验均方根误差

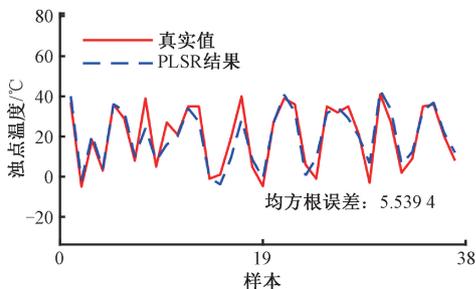
Table 2 Root mean square error of single domain experiment under different working conditions

工况编号	c-Lap-PLSR	PLSR
工况 1	2.174	3.289
工况 2	1.859 7	5.539 4
工况 3	2.194	3.149
工况 4	3.282	4.360

从表 2 中可以看出,在隐变量相同的情况下,c-Lap-PLSR 模型在所有工况上的表现均好于 PLSR 模型。选取工况 2 的其中一次结果进行展示,为了更清晰地表示结果,随机选取工况 2 中 5%的数据作为验证集进行结果验证及图绘制,结果如图 1 所示。



(a) c-lap-PLSR结果
(a) The result of c-lap-PLSR



(b) PLSR结果
(b) The result of PLSR

图 1 工况 2 上两种方法单域回归结果对比

Fig. 1 Comparison of single-domain regression results of the two methods in working condition 2

从图 1 可知,增加了时间近邻拉普拉斯正则项,有益于提升模型拟合效果。PLSR 模型在隐空间中增强自变量和因变量对应的隐变量之间的相关性,并最大程度地保持了数据变量的方差,保留了最主要信息。但是 PLSR 只对数据的映射进行了约束,忽略了过程数据的序列结构。对于过程数据,隐变量应当具有和源数据近似的序列结构,因此,时间近邻拉普拉斯正则项能够有效改善工业过程数据的软测量回归模型的效果。

4.2 验证改进算法的有效性

为了验证 TNN-LR-MR 算法在多工况软测量回归任务中具有比较好的性能改进,跨域实验任取两两工况数据分别作为历史工况和当前工况进行交叉验证。同时,将域适应模型 CORAL、子空间对齐(subspace alignment, SA)^[31]以及 DIPALS 和本文所提模型 TNN-LR-MR 进行比较,并设置 PLSR 作为对照组。TNN-LR-MR 中, c 均取 100。10 次实验的均方根误差结果平均值如表 3 所示。

表 3 5 种方法工况间交叉预测均方根误差

Table 3 Cross prediction of root mean square error among the five methods

	10 次实验取平均	工况 1	工况 2	工况 3	工况 4
工况 1	PLSR	3.377	4.839	5.070	3.494
	CORAL	3.377	4.836	5.070	3.479
	SA	5.335	31.959	9.881	5.485
	DIPALS	2.432	4.091	3.722	3.242
	TNN-LR-MR	1.947	3.291	2.927	2.891
工况 2	PLSR	4.112	4.632	5.618	3.702
	CORAL	4.111	4.632	5.618	3.688
	SA	5.713	6.474	10.618	4.993
	DIPALS	4.077	4.051	4.156	3.529
	TNN-LR-MR	3.024	2.730	3.278	3.165
工况 3	PLSR	3.792	4.659	2.937	4.210
	CORAL	3.793	4.660	2.937	4.183
	SA	13.256	42.671	6.241	9.595
	DIPALS	2.862	2.988	2.527	3.796
	TNN-LR-MR	2.252	2.391	2.043	2.664
工况 4	PLSR	4.860	5.105	6.164	3.570
	CORAL	4.845	5.101	6.163	3.569
	SA	5.985	6.541	12.596	5.251
	DIPALS	4.929	5.080	6.200	3.057
	TNN-LR-MR	4.282	4.908	5.977	2.577

选取工况 3 为历史工况,工况 2 为当前工况,选择 10 次实验的其中一次结果进行展示,同样随机选取工况 2 中 5%的数据作为验证集进行结果验证及图绘制,结果如图 2 所示。

从表 3 可以看出,相对于 PLSR 对照组,DIPALS 和 TNN-LR-MR 都能够多工况回归上面发挥作用,验证了域适应策略的有效性。但是,TNN-LR-MR 预测结果均好于 DIPALS,这表明时间近邻拉普拉斯正则项在映射过程中保持过程数据的序列结构,对于多工况软测量回归是有利的。同时,TNN-LR-MR 结果大多好于 CORAL 和

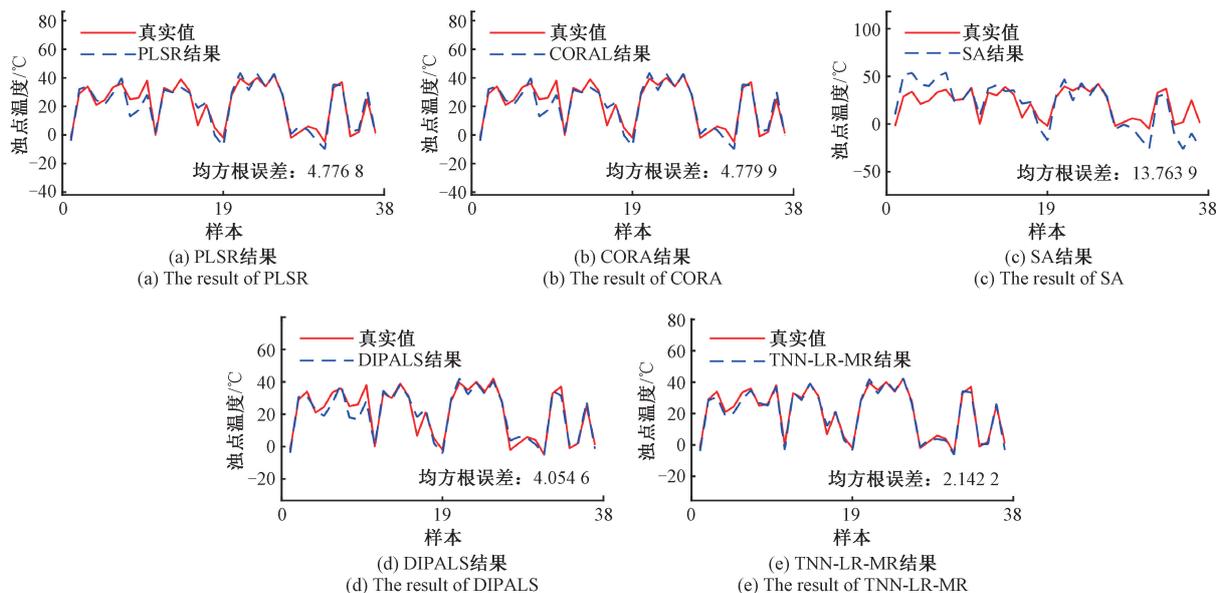


图2 工况2上5种方法跨域回归结果对比

Fig. 2 Comparison of cross-domain regression results of the five methods in working condition 2

SA,表明在映射过程中不约束数据和标签的相关性,得到的隐变量对标签不具备好的解释性,且由于过程数据变量的多重共线性,回归系数矩阵有病态的风险,削弱了回归模型的泛化性和鲁棒性。训练时间消耗改善有效性。

本实验旨在验证 TNN-LR-MR 算法中,时间近邻拉普拉斯正则项提升整体模型训练效率。本实验第一部分验证在分别使用全连接法与时间近邻法建立近邻矩阵时,样本数量与训练消耗时间上的关系。使用时间近邻

法建立近邻矩阵时,设定 $c = 100$,使用全连接法等同于设置 $c \geq \max\{n_s, n_r\}$ 。

设置工况4为历史工况,分别建立与工况1、2、3的 TNN-LR-MR 模型。当前工况数据随机抽取百分之六十作为训练数据,剩下百分之四十作为测试数据,下文所述样本数 n_r 均指当前工况上训练集的样本数。运行相关结果如表4所示。以样本为 X 轴,运行时间为 Y 轴,按照样本数增序排列,样本数-时间关系如图3所示。

表4 两种方法训练消耗时间对比

Table 4 The training time consumed comparison of two methods

实验设定	工况1	工况2	工况3	工况4
样本数 nT ($nS=472$)	1 819	439	2 334	283
全连接法运行时间: t_1/s	16.432 465	1.895 659	29.351 291	1.404 143
时间近邻法运行时间: t_2/s ($c=100$)	2.318 591	0.882 918	2.924 344	0.694 898

误差允许范围内,可以认为图3中全连接法运行时间与样本数量大致呈现一种二次平方的关系,而时间近邻法运行时间与样本数量大致呈现一种线性关系,这与2.2小节中关于样本数的时间复杂度分析结果相同。

随后探讨时间变量 c 的选取对运行时间和运行结果的影响,历史工况和当前工况分别设置为工况1和工况3, c 从1~3 500 递增,重复运行模型,记录实际运行时间以及测试结果。同时,在上述设置下将式(6)绘制成理论运行次数曲线,并与实际运行时间绘制在同一图中进行比较,结果如图4所示。

从图4可以看出,在机器运行速度变化误差允许范围内,可以认为实际运行时间满足式(6)理论分析结果。为了表示模型测试结果与时间变量 c 的关系,以及更直观地展示测试结果与运行时间的平衡,将实际运行时间与均方根误差绘制在同一张图上,结果如图5所示。

图5中一个有趣的现象是,误差允许范围内,均方根误差呈现先减小后增加的趋势,且当 $c \approx 1 500$ 时,存在一个最低点。这是近邻样本选择的对称性和数量之间的冲突和平衡造成的,接下来一小节对该现象进行分析。

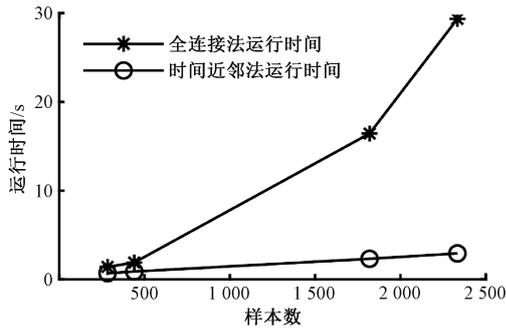


图 3 两种方法消耗时间与样本数的关系
Fig. 3 The relationship between the time consumed by the two methods and the sample size

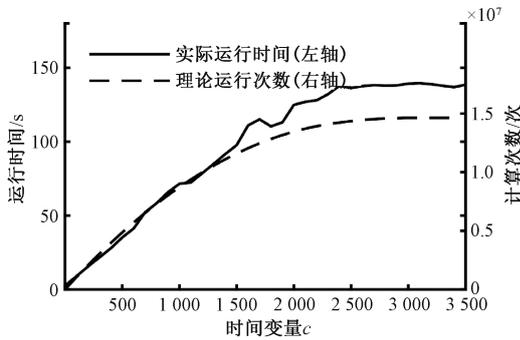


图 4 时间变量 c 的影响实际与理论结果对比图
Fig. 4 Comparison of actual and theoretical results of the influence of time variable c

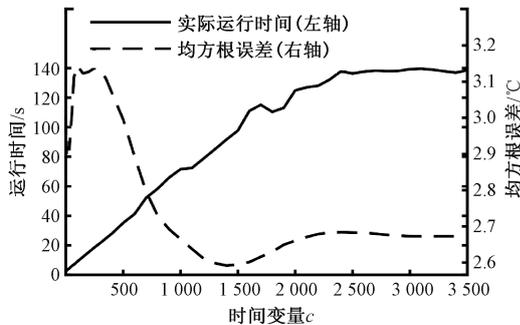


图 5 时间变量 c 对运行时间和运行结果的影响
Fig. 5 The influence of time variable c on running time and running results

4.3 c 的选择与均方根误差的关系分析

以源域数据集工况 1 为例,其近邻矩阵按照式(4)构建为 $W_s \in \mathbb{R}^{n_s \times n_s}$,时间变量设置为 c 。近邻矩阵可以使用图 6 表示。

图 6 中,阴影部分即为近邻矩阵元素非 0 点,阴影面积越大表示计算量越大,近邻关系考虑的越多。 A 与 B 两条点线之间的阴影区域为对称区域,对称区域表示若

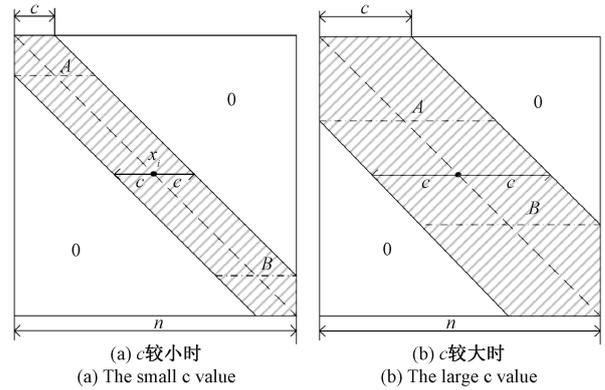


图 6 时间近邻矩阵示意图
Fig. 6 Diagram of time nearest neighbor matrix

对角线某个位置对应的数据点 x_i , 则其前后各 c 个连续样本均能被选择到,因此对称区域上下两边 A, B 之间的距离表示有多少个这样的数据点,而 A, B 的边长表示这个数据点考虑到了前后多少个样本的近邻关系,为表述方便,下文均以对称区域的“边距”和“底长”分别表示上述两种概念。图 6(a) 和 (b) 对比可以得到,随着 c 的增加,总计算量增大,对称区域的边距减小,底长增大。容易得到, $c \geq n_s/2$ 时,对称区域面积为 0。近邻关系考虑得越多,结构保持的效果越好,但是不对称的近邻关系选取,会制约结构保持的效果,最终反映在均方根误差结果上,即:对称区域的边距和底长对均方根误差均为正相关,对称区域面积与均方根误差正相关,不对称区域面积与均方根误差负相关。

由图 6 容易计算得到,总阴影部分面积为 $K_1(n_s, c) = 2n_s c - c^2$,对称区域阴影部分面积为 $K_2(n_s, c) = 2n_s c - 4c^2$,不对称区域阴影部分面积为 $K_3(n_s, c) = 3c^2$ 。本节实验中, $n_s = 3\ 032, n_r = 2\ 334$,将 3 个部分阴影面积与 c 的关系作图表示为图 7 所示。

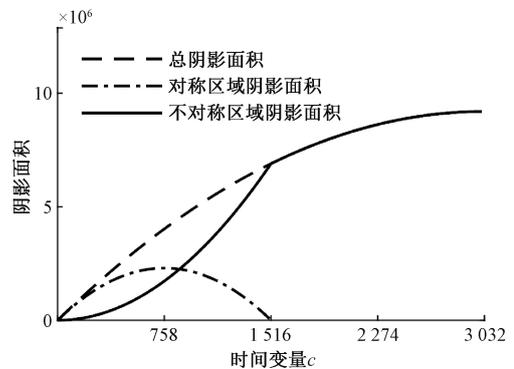


图 7 两个阴影部分面积与时间变量 c 的关系
Fig. 7 The relationship between the area of the two shaded parts and the time variable c

按照对称区域阴影面积是否大于0,将整个过程划分为两个阶段, $0 < c \leq n_s/2 = 1516$ 为第1阶段, $1516 = n_s/2 \leq c \leq n_s = 3032$ 为第2阶段。第1阶段中,当 $0 < c \leq n_s/4 = 758$ 时,总阴影面积与对称区域阴影面积均在增大,此时均方根误差会呈现较快减小趋势,但是当 $n_s/4 < c \leq n_s/2 = 1516$ 时,对称阴影面积在减小,但是总阴影面积仍呈现较快增长,此时均方根误差减小趋势变缓。第2阶段中,对称区域面积已经减小到0,不对称阴影面积与总阴影面积相等,表现制约作用,此时均方根误差会呈现轻微增大的趋势。图5结果符合上述规律。

需要注意的是,虽然在 c 较小的情况下,均方根误差处于最大值附近,但是仍然远小于其他对比方法。上述对时间变量 c 的选取进行了定性分析,但是对于变量的定量分析及优化仍存在较大困难,是下一阶段对该算法进行改善的重要考虑方向。实际应用过程中可以根据实际工艺需要选择合适的 c 值,以满足实时性要求。

5 结 论

本文通过引入域适应正则项改善工况差异对模型的影响,同时提出时间近邻拉普拉斯正则项,在模型训练过程中保持数据结构信息的同时能够大幅度降低训练时间,满足工业软测量的实时性要求。仿真分析结果表明,本文提出TNN-LR-MR方法建立的多工况软测量模型在三聚氰胺数据集上的预测精度均高于其他对比方法。最后,对时间近邻正则项中时间变量 c 的选择与训练时间和均方根误差的关系的分析结果表明,所提算法TNN-LR-MR在保证较好训练精度的同时,能够满足工业的实时性要求。应当说明的是,本文所提模型并不适用于非线性工业过程的软测量建模,后续可引入核方法对非线性情况进行研究和探讨。

参考文献

- [1] 汤健,赵立杰,岳恒,等. 湿式球磨机筒体振动信号分析及负荷软测量[J]. 东北大学学报(自然科学版), 2010,31(11):1521-1524.
TANG J, ZHAO L J, YUE H, et al. Analysis of vibration signal of wet ball mill shell and soft sensing for mill load[J]. Journal of Northeastern University (Natural Science), 2010, 31(11): 1521-1524.
- [2] 汤健,柴天佑,余文,等. 在线KPLS建模方法及在磨机负荷参数集成建模中的应用[J]. 自动化学报, 2013,39(5): 471-486.
TANG J, CHAI T Y, YU W, et al. On-line KPLS algorithm with application to ensemble modeling parameters of mill load[J]. Acta Automatica Sinica, 2013, 39(5): 471-486.
- [3] 汤健,郑秀萍,赵立杰,等. 基于频域特征提取与信息融合的磨机负荷软测量[J]. 仪器仪表学报, 2010, 31(10):2161-2167.
TANG J, ZHENG X P, ZHAO L J, et al. Soft sensing of mill load based on frequency domain feature extraction and information fusion[J]. Chinese Journal of Scientific Instrument, 2010, 31(10): 2161-2167.
- [4] GELADI P, KOWALSKI B R. Partial least-squares regression: A tutorial [J]. Analytica Chimica Acta, 1986, 185(1):1-17.
- [5] SHAWE-TAYLOR J, SUN S. A review of optimization methodologies in support vector machines [J]. Neurocomputing, 2011, 74(17):3609-3618.
- [6] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features [C]. Proc. Conference on Machine Learning. Springer, Berlin, Heidelberg, 1998, 1398:137-142.
- [7] 赵彦涛,单泽宇,杨黎明,等. 基于互信息和约减LSSVM的水泥生料细度软测量[J]. 电子测量与仪器学报, 2019,33(12): 173-182.
ZHAO Y T, SHAN Z Y, YANG L M, et al. Soft sensor modeling of cement raw material fineness based on mutual information and Reduced LSSVM [J]. Journal of Electronic Measurement and Instrumentation, 2019, 33(12): 173-182.
- [8] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10):1345-1359.
- [9] BAKTASHMOTLAGH M, HARANDI M T, LOVELL B C, et al. Domain adaptation on the statistical manifold[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014: 2481-2488.
- [10] CHEN S, HARANDI M, JIN X, et al. Domain adaptation by joint distribution invariant projections[J]. IEEE Transactions on Image Processing, 2020, 29: 8264-8277.
- [11] ADEL T, HAN Z, WONG A. Unsupervised domain adaptation with a relaxed covariate shift assumption[C]. 31st AAAI Conference on Artificial Intelligence, 2017: 1691-1697.
- [12] GHOLENJI E, TAHMORESNEZHAD J. Joint discriminative subspace and distribution adaptation for unsupervised domain adaptation [J]. Applied Intelligence, 2020, 50(7):2050-2066.
- [13] WANG J, FENG W, CHEN Y, et al. Visual domain adaptation with manifold embedded distribution alignment[C]. ACM International Conference on Multimedia. ACM, 2018:402-410.

- [14] BEN-DAVID S, BLITZER J, CRAMMER K, et al. Analysis of representations for domain adaptation [C]. International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2006: 137-144.
- [15] 刘丽娜, 马世伟, 温加睿. 基于局部约束字典学习的数据降维和重构方法[J]. 仪器仪表学报, 2016, 37(1): 99-108.
LIU L N, MA SH W, WEN J R. Dimensionality reduction and reconstruction method based on locality constrained dictionary learning [J]. Chinese Journal of Scientific Instrument, 2016, 37(1): 99-108.
- [16] TENENBAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
- [17] ROWEIS S, SAUL L. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.
- [18] HE X, DENG C, YAN S, et al. Neighborhood preserving embedding [C]. 10th IEEE International Conference on Computer Vision. IEEE, 2005: 1208-1213.
- [19] 潘国锋. 基于紫外吸收光谱的水质硝酸盐含量 LLE-SVR 建模 [J]. 仪器仪表学报, 2011, 32(12): 2869-2873.
PAN G F. LLE-SVR modeling for nitrate measurement in water based on UV absorption spectra [J]. Chinese Journal of Scientific Instrument, 2011, 32(12): 2869-2873.
- [20] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from examples [J]. The Journal of Machine Learning Research, 2006, 7: 2399-2434.
- [21] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2003, 15(6): 1373-1396.
- [22] 张旭, 黄亦翔, 张旭东, 等. 基于 Kullback-Leibler 距离的起重机回转系统健康评估[J]. 电子测量与仪器学报, 2021, 35(2): 25-32.
ZHANG X, HANG Y X, ZHANG X D, et al. Kullback-Leibler distance based health performance evaluation for rotary system of crane truck [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(2): 25-32.
- [23] JONG S D. SIMPLS: An alternative approach to partial least squares regression [J]. Chemometrics & Intelligent Laboratory Systems, 1993, 18(3): 251-263.
- [24] NIKZAD-LANGERODI R, ZELINGER W, SAMINGER-PLATZ S, et al. Domain adaptation for regression under Beer-Lambert's law [J]. Knowledge-Based Systems, 2020, 210: 106447.
- [25] WOLD H. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach [J]. Journal of Applied Probability, 1975, 12(S1): 117-142.
- [26] BEN-DAVID S, BLITZER J, CRAMMER K, et al. A theory of learning from different domains [J]. Machine Learning, 2010, 79(1-2): 151-175.
- [27] BORGWARDT K M, GRETTON A, RASCH MJ, et al. Integrating structured biological data by kernel maximum mean discrepancy [C]. ISMB, 2006: 49-57.
- [28] HUANG J Y, SMOLA A J, GRETTON A, et al. Correcting sample selection bias by unlabeled data [C]. Proceedings of NIPS, 2006: 601-608.
- [29] QUADRANTO N, PETERSON J, SMOLA A J. Distribution matching for transduction [C]. Proceedings of NIPS, 2009: 1500-1508.
- [30] SUN B, FENG J, SAENKO K. Return of frustratingly easy domain adaptation [C]. 30th AAAI Conference on Artificial Intelligence, 2016: 2058-2065.
- [31] FERNANDO B, HABRARD A, SEBBAN M, et al. Unsupervised visual domain adaptation using subspace alignment [C]. IEEE International Conference on Computer Vision (ICCV), 2013: 2960-2967.

作者简介



徐志强, 2018 年于太原理工大学获得学士学位, 现于太原理工大学攻读硕士学位, 主要研究方向为软测量、机器学习、迁移学习。
E-mail: 15735181522@163.com

Xu Zhiqiang received his B.Sc. degree from Taiyuan University of Technology in 2018. He is currently a master student at Taiyuan University of Technology. His main research interests include soft sensors, machine learning and transfer learning.



阎高伟 (通信作者), 1992 年于太原工业大学获得学士学位, 2003 年于太原理工大学获得硕士学位, 2007 年于太原理工大学获得博士学位, 现为太原理工大学教授, 主要研究方向为机器学习与人工智能软测量系统。

E-mail: yangaowei@tyut.edu.cn

Yan Gaowei (Corresponding author) received his B.Sc. degree, M.Sc. degree, and Ph.D. degree all from Taiyuan University of Technology in 1992, 2003, and 2007, respectively. He is currently a professor at Taiyuan University of Technology. His main research interests include machine learning, artificial intelligence, and soft sensor system.