

文章编号: 1006-3080(2021)02-0217-08

DOI: 10.14135/j.cnki.1006-3080.20191202008

基于 Tri-training MPLS 的半监督软测量模型

李 东, 刘乙奇, 黄道平

(华南理工大学自动化科学与工程学院, 广州 510641)

摘要:随着污水处理过程日趋复杂, 易测量变量和难测量变量的比例严重失衡, 传统的监督性软测量建模方法已经无法满足需求。针对这一问题, 提出了一种新的半监督学习的软测量模型——Tri-training MPLS 模型。首先将标记数据均分为相互独立的3个部分, 并由这3个相互独立的标记样本子集选择置信度高的未标记样本训练模型, 提高模型的预测能力。其次, 将单输出软测量模型升级为多输出模型, 对多个输出的变量直接建模预测。最后, 通过污水处理仿真模型BSM1(Benchmark Simulation Model-1)平台对本文模型进行验证。结果表明, 该软测量模型不仅具有较好的多输出预测能力, 而且对单个预测结果也有令人满意的预测表现。

关键词:软测量; 半监督学习; 多输出; Tri-training; 偏最小二乘

中图分类号:TP277

文献标志码:A

在工业生产领域, 由于生产过程的复杂性, 存在大量的难以测量或不易检测的重要参数。针对这一问题, 软测量技术提供了一个有效的解决方案^[1-3]。近年来, 软测量技术广泛应用于单输出系统的预测, 对多输出预测模型的研究极少^[4]。然而, 在复杂的生产过程中, 急需对多个难以测量的重要参数进行同时预测^[5-6]。在众多的软测量建模方法中, 目前常用的软测量建模方法有基于机理建模的方法、基于数据驱动建模的方法以及两者相结合的混合建模方法^[7]。基于机理建模的方法需要掌握生产过程的详细信息和生化反应原理, 对于复杂的工业生产过程而言获取较为困难, 使得建立相应的机理模型变得极为不易。基于数据驱动建模的方法则只需要通过一些已知的数据以及这些数据的内在联系构建软测量模型, 因此, 基于数据驱动的建模方法得到了广泛的关注^[8-9]。

在实际的工业过程中, 受到现有技术和生产环境的束缚, 很多情况下已标记样本数量有限, 如何充分使用未标记样本来提升模型质量成为当务之急。

为解决这一难题, Shahshahani 等^[10]首次提出了半监督学习的思想, 并成功应用于软测量建模^[11-13]。半监督学习方法利用少量的标记样本构建模型, 然后通过未标记样本训练模型, 提高模型质量。然而, 在训练过程中, 由于无法剔除错误的未标记数据, 导致错误累积, 影响模型的预测能力。为挑选合适的未标记数据训练模型, Blum 等^[14]提出了协同训练算法, 通过两个相互独立的回归模型, 采用交叉验证的方法挑选数据, 训练模型。

为了提高多输出模型中数据选择的正确性, 本文提出了一种新的软测量模型——Tri-training MPLS 模型。该模型通过3个具有不同特征的回归模型对未标记数据进行验证, 挑选出置信度最高的未标记数据训练模型, 建立更准确的软测量模型。此外, 该模型能够对多个难测量变量进行同步预测, 提高了预测效率。通过污水处理仿真模型BSM1平台验证, 结果表明, 当标记样本的比例较少时, Tri-training MPLS 模型具有更好的预测性能。

收稿日期: 2019-12-02

基金项目: 国家自然科学基金(61873096, 61673181); 广东省科技计划项目(201804010256)

作者简介: 李东(1994—), 男, 河南尉氏人, 博士生, 主要研究方向为软测量。E-mail: liddscut@163.com

通信联系人: 黄道平, E-mail: audhuang@scut.edu.cn

引用本文: 李 东, 刘乙奇, 黄道平. 基于 Tri-training MPLS 的半监督软测量模型[J]. 华东理工大学学报(自然科学版), 2021, 47(2): 217-224.

Citation: LI Dong, LIU Yiqi, HUANG Daoping. Semi-Supervised Soft Sensor Model Based on Tri-training MPLS[J]. Journal of East China University of Science and Technology, 2021, 47(2): 217-224.

1 预备知识

1.1 多输出偏最小二乘算法(MPLS)

MPLS算法是在PLS算法框架上的多数扩展,不同之处在于输出数据 Y 由原来的单列向量转化成了多输出矩阵。数学过程如下:

$$X = TP + E = \sum_{h=1}^a t_h p_h + E \quad (1)$$

$$Y = UQ + F = \sum_{h=1}^a u_h q_h + F \quad (2)$$

其中: $X \in \mathbf{R}^{n \times m}$ 和 $Y \in \mathbf{R}^{n \times l}$ 分别为输入和输出矩阵, n 为数据集的个数, m 为输入变量的个数, l 为输出变量的个数; $T \in \mathbf{R}^{n \times a}$ 和 $U \in \mathbf{R}^{n \times a}$ 分别是 X 和 Y 的得分矩阵, a 表示潜在的变量个数; t_h 为 T 的第 h 行; u_h 为 U 的第 h 行; $P(a \times m)$ 和 $Q(a \times l)$ 为加载矩阵; p_h 是 P 的第 h 行; q_h 是 Q 的第 h 行; E 和 F 为噪声矩阵。 u_h 和 t_h 之间的关系如下:

$$u_h = b_h t_h \quad (3)$$

其中: b_h 为 X 空间主元 t 和 Y 空间主元 u 的内部相关关系的回归系数; $U = TB$, $B \in \mathbf{R}^{a \times a}$ 表示回归矩阵。因此, X 和 Y 之间的关系可以表示为 $Y = TBQ + F$ 。

MPLS是一种常见的多元统计分析方法,不仅可以减少数据的维度,还可以建立预测模型,是解决工业过程中数据预测问题的有力工具^[15-16]。

1.2 协同训练回归

随着半监督建模方法的发展,其在建模领域得到了广泛的关注^[17-18]。然而,由于如何考察新的标记数据的置信度问题一直未能很好地解决,所以有关协同训练算法建模的方法一直未得到有效应用。为此,Zhou等^[19]提出了协同训练回归(Co-training Regression)方法。首先,由最初的标记样本集建立初始模型 h ,通过 h 计算出标记样本集的均方根误差 R_1 。然后,再由 h 计算出未标记数据 x_u 的回归值 $\hat{y}_u$,并将 $(x_u, \hat{y}_u)$ 作为新的标记样本放入最初的标记样本集中。由新的标记样本集建立模型 h' ,计算原来的标记样本集在模型 h' 下的均方根误差 R_2 。计算 R_1 和 R_2 的差值 Δ_u ,取 Δ_u 中的最大值作为新加的标记数据 $(x_u, \hat{y}_u)$ 的最高置信度。具体公式如下:

$$\Delta_u = \frac{1}{|L|} \left(\sum_{x_i \in L} (y_i - h(x_i))^2 - \sum_{x_i \in L} (y_i - h'(x_i))^2 \right) \quad (4)$$

其中: $x_i \in L$ 为标记输入数据; $y_i \in L$ 为标记输出数据; L 为标记样本集, $|L|$ 为样本集大小; h 为原始回归模型; h' 为加入新的标记数据 x_u 后产生的回归模型。

协同训练回归根据协同训练的思想建立两个相互独立的初始模型,在验证过程中,采用交叉验证的方式,减少对错误数据的选择。当达到最大迭代次数时,由最终的标记样本集建立模型 h_1 和 h_2 ,预测值由两者的均值决定:

$$h(x) = \frac{1}{2} (h_1(x) + h_2(x)) \quad (5)$$

2 Tri-training MPLS 模型

2.1 Tri-training 回归

目前,Tri-training回归的难题依旧是如何找到置信度最高的未标记数据。同样,采用协同回归方法中置信度的判别方式,通过计算 R_1 和 R_2 的差值 Δ_u 并取其最大值作为未标记数据 $(x_u, \hat{y}_u)$ 的最高置信度。然而不同的是当给一个原始的标记样本集中添加新的标记数据时,需要通过另外两个模型回归量的均值作为新的标记数据的回归量:

$$h_i(x_u) = \frac{1}{2} (h_j(x_u) + h_k(x_u)) \quad (6)$$

式中: h_i 、 h_j 、 h_k 是3个独立的回归模型。

在多输出系统下,输出数据 y 由原来的单列向量变为多列矩阵,均方根误差(RMSSD)的求解公式如下:

$$\text{RMSSD} = \sum_{x_i \in L} \text{trace} \left((y_i - h(x_i))^T (y_i - h(x_i)) \right) \quad (7)$$

式中:trace是矩阵的迹。

当达到最大迭代次数时,由最终的标记样本集建立模型 h_1 、 h_2 、 h_3 ,预测值由三者的均值决定:

$$h(x) = \frac{1}{3} (h_1(x) + h_2(x) + h_3(x)) \quad (8)$$

2.2 Tri-training MPLS 应用于软测量建模

Tri-training MPLS模型的核心是将Tri-training算法和MPLS算法相结合生成对多输出系统的软测量模型。其优势在于,首先,由上述的置信度判断依据选择正确性更高的未标记数据,模型的预测效果更好;其次,通过标记数据集建立3个相互独立的回归模型,使得预测模型具有广泛性;最后,Tri-training MPLS模型能够同时对多个输出变量预测,极大地提高了预测的效率和模型的整体性。

Tri-training PLS模型的具体过程如下:设 $L = \{X, Y\} = \{(x_1, y_1), (x_2, y_2), \dots, (x_{|L|}, y_{|L|})\}$ 表示标记数据集, $|L|$ 表示标记数据集个数, U 是未标记数据集, $|U|$ 表示未标记数据集个数。首先,将 L 平均分成3部分,记为 L_1 、 L_2 、 L_3 ,将 L_1 、 L_2 、 L_3 与传统的偏最小

二乘算法结合生成相互独立的3个回归模型 P_1 、 P_2 、 P_3 。然后,取未标记数据集 U 中的数据 x_u , 分别代入到 P_1 、 P_2 、 P_3 中计算回归量。在学习过程中,用其中两个回归量的均值来更新另一个标记样本集。随着迭代的进行,该过程不仅不断地建立3个相互独立的回归模型,而且在差异越来越大的回归量上更新标记数据集。为了选取合适的未标记数据,利用上述方法估计未标记数据的置信度,然后对满足置信条件的未标记数据计算回归量,将两个回归量的均值作为新的标记数据放入另一个标记样本集中。最后,由新的标记样本集 L_1 、 L_2 、 L_3 建立新的模型 h_1 、 h_2 、 h_3 , 通过求取3者均值作为最终的预测值。Tri-training MPLS 的详细流程如下:

输入: 标记样本集 L (包含输入变量 x_l 和输出变量 y), 未标记样本集 U (仅包含输入 x_u), 测试样本集 P , 最大迭代次数 T

过程:

分别将 x_l 和 x_u 均分为3部分: x_l^1 、 x_l^2 、 x_l^3 和 x_u^1 、 x_u^2 、 x_u^3

$$L_1 = \{(x_l^1, y_1), (x_l^1, y_2), \dots, (x_{|L|}^1, y_{|L|})\}$$

$$L_2 = \{(x_l^2, y_1), (x_l^2, y_2), \dots, (x_{|L|}^2, y_{|L|})\}$$

$$L_3 = \{(x_l^3, y_1), (x_l^3, y_2), \dots, (x_{|L|}^3, y_{|L|})\}$$

进行 T 次迭代:

For $j \in \{1, 2, 3\}$ do

$$h_j = \text{pls}(L_j)$$

for each $x_u \in U$ do

$$y_u = h_j(x_u^j)$$

$$h'_j = \text{pls}(L_j \cup \{x_u^j, y_u\})$$

$$\nabla_{x_u} = \frac{1}{|L|} \sum_{x_j \in L} \left(\text{trace} \left((y_i - h_j(x_i^j))^T (y_i - h_j(x_i^j)) \right) - \right.$$

$$\left. \text{trace} \left((y_i - h_j(x_i^j))^T (y_i - h_j(x_i^j)) \right) \right)$$

end of for

如果存在 $\nabla_{x_u} > 0$

$$x_n = \arg\max_{x_u} \nabla_{x_u}; y_n^j = h_j(x_n)$$

$$\hat{y}_n^1 = (y_n^2 + y_n^3) / 2; \hat{y}_n^2 = (y_n^1 + y_n^3) / 2; \hat{y}_n^3 = (y_n^1 + y_n^2) / 2$$

然后 $U = U - x_n$

$$\pi_j = \{(x_n, \hat{y}_n^j)\}$$

$$U = U$$

否则

$$\pi_j = \emptyset$$

End of for

$$L_j = L_j \cup \pi_j$$

达到最高迭代次数, 结束迭代

输出新的标记样本集 L_1 、 L_2 and L_3

for $j \in \{1, 2, 3\}$ do

$$h_j = \text{pls}(L_j)$$

end of for

验证模型预测能力:

$$x \in P$$

$$y_p = \frac{1}{3} (h_1(x) + h_2(x) + h_3(x))$$

为比较 Co-training MPLS 模型和 Tri-training MPLS 模型算法的复杂性, 图1示出了两种模型的流程图。由图1可知, 两种模型的算法都被分为训练环节和测试环节两部分, 不同之处在于训练环节中, Co-training MPLS 模型将标记数据均分为两部分, 由两组标记数据子集同时评价未标记数据的置信度。而 Tri-training MPLS 算法是将标记数据三等分, 有3组标记数据子集同时评价未标记数据的置信度, 并选择置信度较高的未标记数据加入到标记数据集中。3个相互独立的标记样本子集同时训练, 避免了因单个数据级判断错误而影响选择的质量, 致使错误累积, 最终导致预测结果较差。从算法复杂度的角度分析, 无论是2组标记子集还是3组标记子集, 训练过程是一致的, 只是 Tri-training MPLS 算法的训练时间会相对较长。

为了评价 Tri-training MPLS 模型的预测性能, 采用传统的均方根误差(RMSE)和 D 值指标评价, 定义如下:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}, \quad i = 1, 2, \dots, n \quad (9)$$

$$D = |\hat{y}_i - y_i| \quad (10)$$

其中: n 表示样本数量; $\hat{y}_i$ 为预测值; y_i 为真实值。

3 仿真实例

污水处理仿真模型 BSM1 平台的设备布局如图2所示, 由一个生物反应器 (5999 m³) 和一个二次沉淀池 (深4 m, 10层, 6000 m³) 组成。生物反应器包含5个混合小单元, 前2个单元 (每个1000 m³) 为非曝气的, 后3个单元 (每个1333 m³) 为曝气的。处理污水的平均流量为20000 m³/d, 可生物降解的化学需氧量(COD)的平均质量浓度为300 mg/L。为了去除有机物, 需要进行硝化和反硝化反应^[20-21]。

仿真过程中, 共纳入了15个输入变量和5个输出变量, 详细的变量信息如表1所示。本案例的研究目的是验证 Tri-training MPLS 模型对 SS-E、SNH-E、SNO-E、COD-E 和 BOD₅-E 等不易测量变量的预测

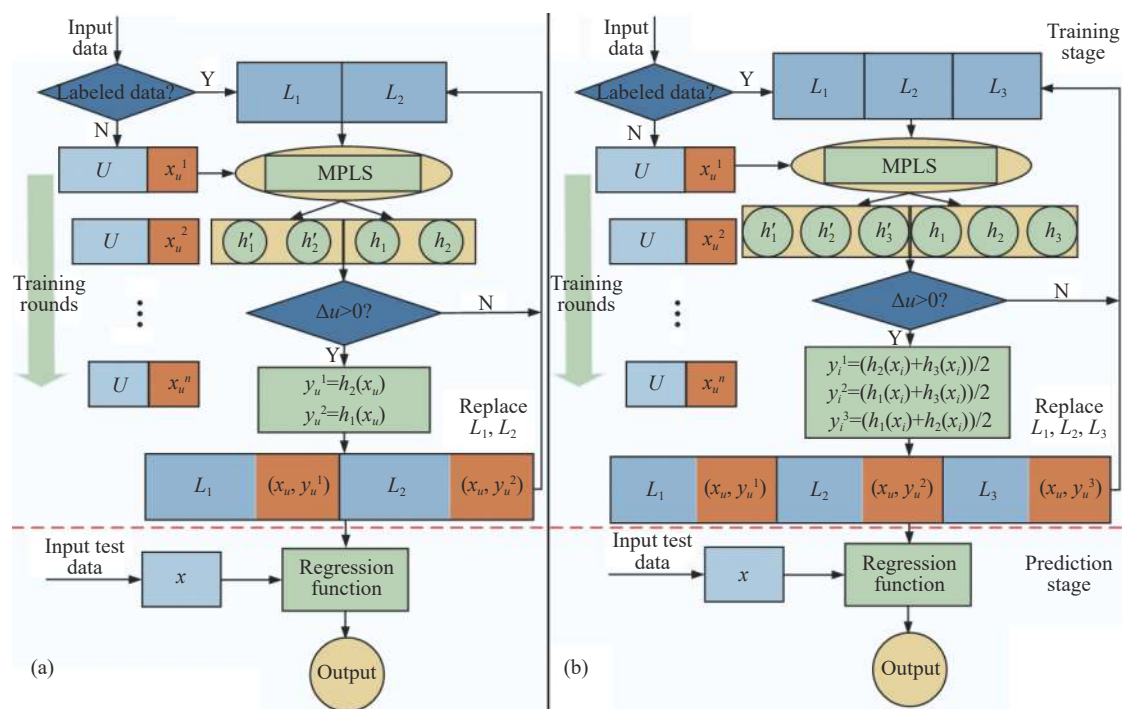


图 1 Co-training MPLS(a)和 Tri-training MPLS(b)模型的流程图

Fig. 1 Flow chart of Co-training MPLS (a) and Tri-training MPLS (b) models

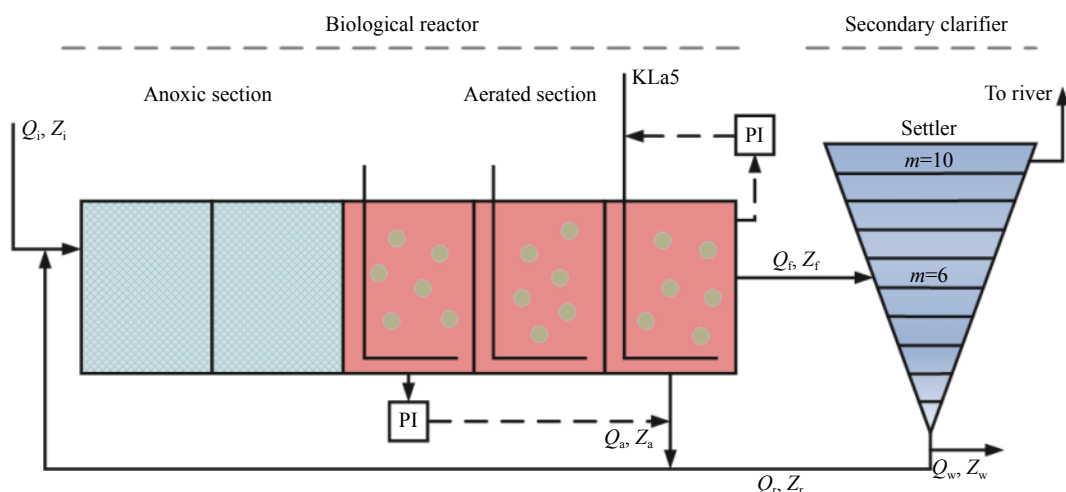


图 2 BSM1 平台设备布局示意图

Fig. 2 Schematic diagram of BSM1 platform equipment layout

性能。每隔 15 min 对各个变量采样一次, 模拟 14 d, 共 1 344 组数据。分别将第 1 天、第 2 天、…、第 7 天的样本作为已标记数据集, 剩余部分作为未标记数据集, 最大学习次数设置为 5 次。最后, 用剩余 7 d 的数据作为测试样本集, 对模型进行测试。

图 3 示出了标记样本为 4 d 时两种模型的预测曲线和 D 值。可以看出, Co-training MPLS 模型和 Tri-training MPLS 模型均能够同时对多个输出变量进行预测, 且预测效果良好, 尤其是对污水排放指标中的重要参数 BOD_5-E 的预测, RMSE 值分别达到了 0.08 和 0.07, 预测曲线与真实值基本一致, 其他具体

参数见表 2。MPLS 算法作为一种线性建模方法, 显然对于波动较大的输出数据的预测性能较差, 特别是预测曲线中的峰值点和谷值点, 都没有达到良好的预测结果。

图 4 示出了输出变量 RMSE 值的变化曲线图。通过表 2 和图 4 分析, 当标记数据所占天数仅为 1 d 时, Tri-training MPLS 模型中的 SS-E、SNO-E、COD-E 和 BOD_5-E 的 RSME 值均小于 Co-training MPLS 模型的 RSME 值。因此, 在标记数据极少的情况下, Tri-training MPLS 模型对于污水处理过程中的重要指标参数的预测能力优于 Co-training MPLS 模型。

表1 BSM1平台中的变量
Table 1 Variables of the BSM1 platform

No.	Variables	Comments	No.	Variables	Comments
1	SS-1	Readily biodegradable substrate-1	11	SO-2	Oxygen-2
2	SS-in	Readily biodegradable substrate influent	12	SO-5	Oxygen-5
3	SNH-1	$\text{NH}_4^+ + \text{NH}_3$ nitrogen-1	13	$Q\text{-tr}$	Flow rate internal recycling
4	SNH-2	$\text{NH}_4^+ + \text{NH}_3$ nitrogen-2	14	$Q\text{-in}$	Flow rate influent
5	SNH-3	$\text{NH}_4^+ + \text{NH}_3$ nitrogen-3	15	COD-in	Chemical oxygen demand for influent
6	SNH-in	$\text{NH}_4^+ + \text{NH}_3$ nitrogen influent	16	SS-E	Readily biodegradable substrate effluent
7	SNO-1	Nitrate and nitrite nitrogen-1	17	SNH-E	$\text{NH}_4^+ + \text{NH}_3$ nitrogen effluent
8	SNO-2	Nitrate and nitrite nitrogen-2	18	SNO-E	Nitrate and nitrite nitrogen effluent
9	SNO-3	Nitrate and nitrite nitrogen-3	19	COD-E	Chemical oxygen demand for effluent
10	SO-1	Oxygen-1	20	$\text{BOD}_5\text{-E}$	Five-day biological oxygen demand effluent

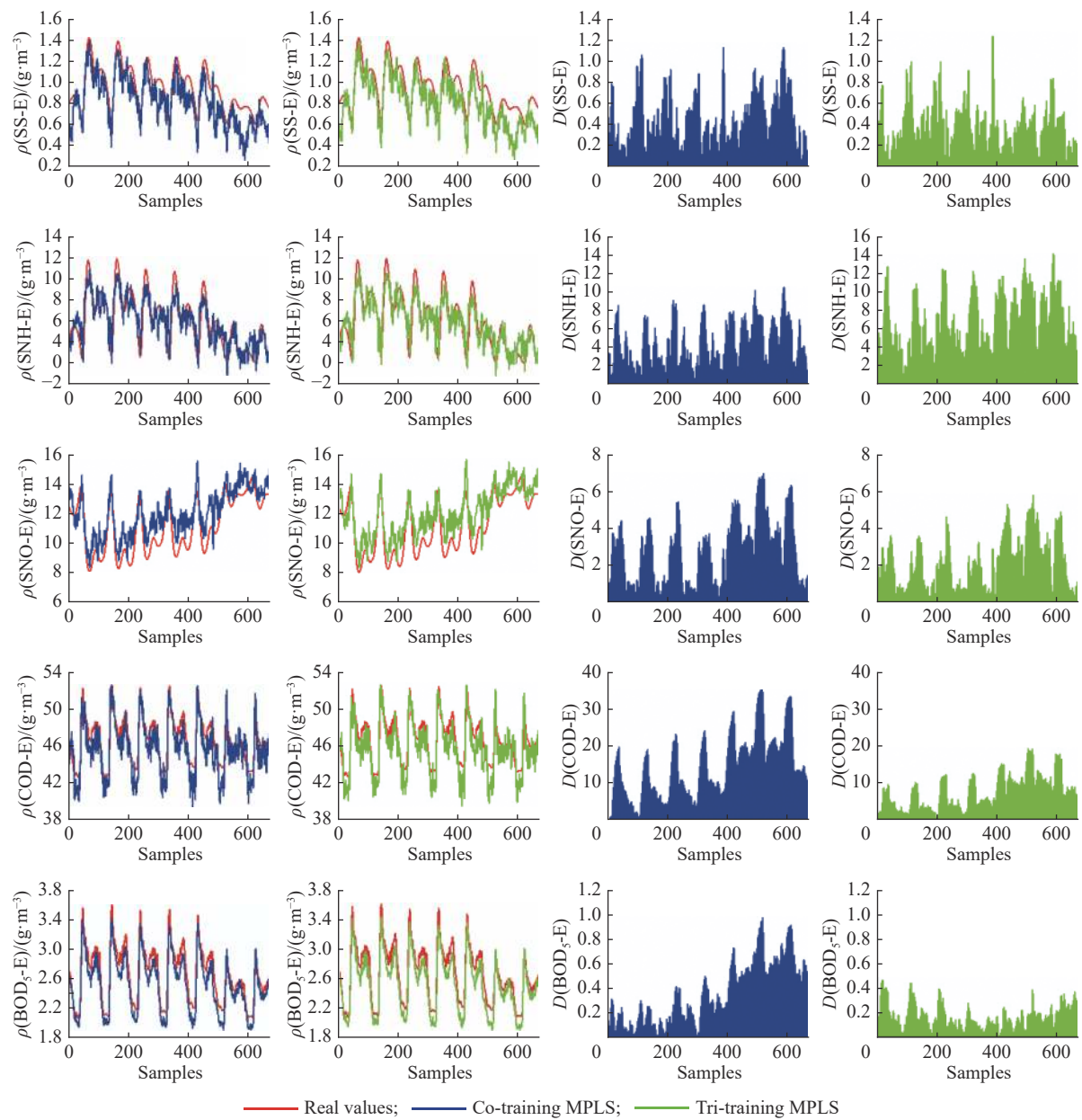


图3 两种模型的预测曲线和D
Fig. 3 Prediction curves and D of two models

表 2 输出变量的 RSME 值
Table 2 RMSE values of output variables

Labeled data/d	RSME by Co-training MPLS					RSME by Tri-training MPLS				
	SS-E	SNH-E	SNO-E	COD-E	BOD ₅ -E	SS-E	SNH-E	SNO-E	COD-E	BOD ₅ -E
1	0.20	20.46	9.61	257.12	0.18	0.14	48.55	5.34	64.37	0.03
2	0.06	5.40	1.76	16.42	0.07	0.06	7.15	1.74	15.56	0.07
3	0.06	5.96	1.72	15.23	0.11	0.07	4.39	1.95	22.40	0.12
4	0.10	5.54	1.70	10.50	0.08	0.08	5.69	1.72	11.02	0.07
5	0.10	5.17	1.79	6.10	0.05	0.08	4.96	1.64	6.79	0.05
6	0.06	3.19	1.18	3.37	0.03	0.06	3.02	1.19	3.40	0.03
7	0.03	2.02	1.41	2.49	0.03	0.03	2.01	1.43	2.50	0.03

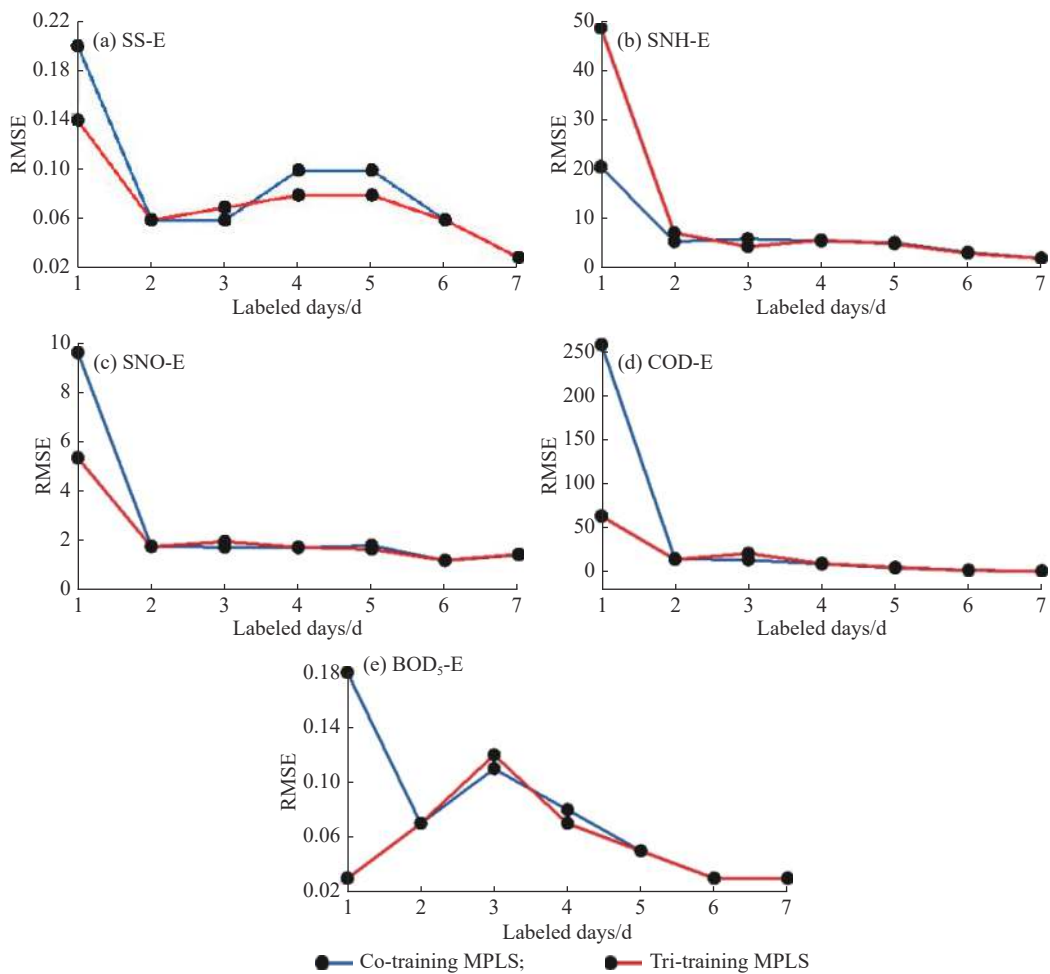


图 4 不同标记样本率下的 RMSE 变化曲线

Fig. 4 RMSE curves under different labeled data rates

但观察发现,在标记数据为 1 d 时 COD-E 的 Co-training MPLS 模型和 Tri-training MPLS 模型的 RSME 值分别为 257.12 和 64.37,明显偏大。随着标记数据所占天数的增加,各个输出变量的 RMSE 值均在减小,尤其是 COD-E 的 RSME 值由 257.12 和 64.37 迅速减小到 2.49 和 2.50。这说明当标记样本

较少时,对于波动较大的输出变量,Co-training MPLS 模型和 Tri-training MPLS 模型需要通过少量的标记数据将建立 2 个或者 3 个回归模型,导致模型不准确,预测表现不佳。当标记数据充足时,两个模型均能通过标记数据建立更准确的预测模型,并且通过未标记数据训练模型,对多个输出变量进行精

准的预测。综上所述,在仅仅具有少量的标记数据情况下, Tri-training MPLS模型对多个输出变量的预测效果优于 Co-training MPLS 模型;当标记数据增多时,两种模型的预测表现差别不大。

表 3 列出了两种模型在不同标记样本率下的时间消耗对比结果。无论是 Co-training MPLS 模型还是

Tri-training MPLS 模型,随着标记样本率的增加,时间消耗也在增加。在相同的标记样本率下, Tri-training MPLS 模型的时间消耗相对于 Co-training MPLS 模型有所增长,这主要是因为 3 组标记数据子集的训练要比 2 组费时,但是 Tri-training MPLS 模型的预测结果较好。

表 3 不同的标记样本率下的时间消耗
Table 3 Time consumption under different labeled data rate

Model	Time consumption/s						
	1 d	2 d	3 d	4 d	5 d	6 d	7 d
Co-training MPLS	22.124	34.186	45.782	60.015	71.689	81.473	93.159
Tri-training MPLS	25.657	39.249	49.561	61.954	72.019	84.554	96.233

4 结 论

针对多输出系统中的软测量建模问题,当输入和输出数据严重不平衡时,本文提出了一种新的软测量模型——Tri-training MPLS 模型。通过污水处理仿真模型 BSM1 平台的仿真研究,验证了模型的有效性。仿真结果表明,虽然 Tri-training MPLS 模型在预测时间上相较于 Co-training MPLS 模型有一定的滞后,但在预测效果上却有明显的提升,尤其是对重要的出水指标 COD-E,其 RSME 值比 Co-training MPLS 模型平均减小 26.46。该模型为多输出系统中难以测量的变量预测问题提供了一个有效的解决方法。然而,不难发现,模型对于峰值点和谷值点的预测效果不佳。在实际的工业生产过程中,也会存在着正常波动和离群点,影响模型的预测表现。针对这一问题,下一步的研究会将 Tri-training 算法与非线性的多输出回归算法结合,并且,在建模之前对数据进行标准化处理,以降低因数据波动对预测表现造成的影响。

参考文献:

[1] 黄道平,刘乙奇,李艳. 软测量在污水处理过程中的研究与应用[J]. 化工学报, 2011, 62(1): 1-9.
[2] 杨逸俊. 基于 WOA-GRU 的精馏塔丙烷浓度软测量建模 [C]//第 30 届中国过程控制会议(CPCC 2019). 云南昆明: 中国自动化学会过程控制专业委员会, 2019: 2-16.
[3] 蒋昕祎,李绍军,金宇辉. 基于慢特征重构与改进 DPLS 的软测量建模[J]. 华东理工大学学报(自然科学版), 2018, 44(4): 535-542.
[4] LIU Y Q, LIU B, ZHAO X J, *et al.* Development of RVM-

based multiple-output soft sensors with serial and parallel stacking strategies[J]. *IEEE Transactions on Control Systems Technology*, 2018, 27(6): 2727-2734.
[5] LIU Y, LIU B, Zhao X, *et al.* Development of RVM-based multiple-output soft sensors with serial and parallel stacking strategies[J]. *IEEE Transactions on Control Systems Technology*, 2019, 27(6): 2727-2734.
[6] XIAO H, BAI B, LI X, *et al.* Interval multiple-output soft sensors development with capacity control for wastewater treatment applications: A comparative study[J]. *Chemometrics and Intelligent Laboratory Systems*, 2019, 184: 82-93.
[7] 王忠建. 基于软测量法的带钢酸洗液组分浓度在线检测研究[D]. 上海: 华东理工大学, 2016.
[8] WANG X, LIU H. A knowledge-and data-driven soft sensor based on deep learning for predicting the deformation of an air preheater rotor[J]. *IEEE Access*, 2019, 7: 159651-159660.
[9] GOPAKUMAR V, TIWARI S, RAHMAN I. A deep learning based data driven soft sensor for bioprocesses[J]. *Biochemical Engineering Journal*, 2018, 136: 28-39.
[10] SHAHSHAHANI B M, LANDGREBE D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. *IEEE Transactions on Geoscience & Remote Sensing*, 1994, 32(5): 1087-1095.
[11] CHI S, LI X, TIAN Y, *et al.* Semi-supervised learning to improve generalizability of risk prediction models[J]. *Journal of Biomedical Informatics*, 2019, 92: 103-117.
[12] ZHU J, GE Z, SONG Z. Quantum statistic based semi-supervised learning approach for industrial soft sensor development[J]. *Control Engineering Practice*, 2018, 74: 144-152.
[13] SHAO W, GE Z, SONG Z, *et al.* Nonlinear industrial soft sensor development based on semi-supervised probabilistic

- mixture of extreme learning machines[J]. *Control Engineering Practice*, 2019, 91: 98-104.
- [14] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//*Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. USA: ACM, 1998: 92-100.
- [15] 程剑, 宋淑群, 张凌波, 等. 基于PLS-混合Pi-Sigma模糊神经网络模型的甲醇合成装置变换工序CO变换率软测量建模[J]. *华东理工大学学报(自然科学版)*, 2015, 41(1): 66-71.
- [16] LIU Y, PAN Y, HUANG D. Development of a novel adaptive soft-sensor using variational Bayesian PLS with accounting for online identification of key variables[J]. *Industrial & Engineering Chemistry Research*, 2015, 54(1): 338-350.
- [17] BAO L, YUAN X, GE Z. Co-training partial least squares model for semi-supervised soft sensor development[J]. *Chemometrics and Intelligent Laboratory Systems*, 2015, 147: 75-85.
- [18] FERREIRA V, SOUZA F A A, ARAÚJO R. Semi-supervised soft sensor and feature ranking based on co-regularised least squares regression applied to a polymerization batch process[C]//*2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*. USA: IEEE, 2017: 257-262.
- [19] ZHOU Z H, LI M. Semisupervised regression with cotrain-ing-style algorithms[J]. *IEEE Transactions on Knowledge & Data Engineering*, 2007, 19(11): 1479-1493.
- [20] 吴菁, 刘乙奇, 刘坚, 等. 基于动态多核相关向量机的软测量建模研究[J]. *化工学报*, 2019, 70(4): 1472-1484.
- [21] 刘大伟, 沈文浩. 废水处理仿真基准模型BSM1简介[J]. *广州环境科学*, 2007, 22(1): 11-15.

Semi-Supervised Soft Sensor Model Based on Tri-training MPLS

LI Dong, LIU Yiqi, HUANG Daoping

(School of Automation Science & Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: With the sewage treatment process becoming more and more complex, the proportion of easy-to-measure and hard-to-measure variables is seriously out of balance such that the traditional supervised soft-sensor modeling cannot meet the actual requirements. Aiming at this problem, this paper proposes a new soft-sensor model, semi-supervised Tri-training MPLS. The labeled data are divided into three independent parts, from which the unlabeled data with high confidence will be selected to improve the prediction ability of the model. In addition, the single-output model is upgraded to the multi-output model to directly predict multiple output variables. Finally, it is shown from the simulation results via the BSM1 platform (Benchmark Simulation Model-1) that the proposed soft-sensor model has good prediction ability on multiple output prediction and satisfactory prediction on single target.

Key words: soft sensor; semi-supervised; multi-output; Tri-training; partial least squares