第42卷，第11期
2022年11月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 11, pp3608-3614
November, 2022

基于PLS子空间对齐的2,6-二甲酚纯度迁移学习建模

邬云飞，栾小丽*，刘 飞

江南大学自动化研究所，轻工过程先进控制教育部重点实验室，江苏 无锡 214122

摘 要 采用近红外光谱对物质浓度进行准确的在线检测对于生产优化具有重要意义。建立检测模型需要从近红外光谱中提取相关信息，代表性样本越多，提取的信息越有效，所建模型的精度越高。随着产品纯度的提高，样本的区分度下降，样本的变异系数小，多样性不足，并且存在测量噪声以及化验室人工检测样品浓度值时的测量误差，会导致物质浓度与光谱之间缺乏相关性，传统的建模方法无法建立可靠的近红外检测模型。为了解决这个问题，提出了一种基于PLS子空间对齐的迁移学习建模方法，应用于2,6-二甲酚精馏提纯过程中产品塔高纯度产品的在线检测。在制备化工单体2,6-二甲酚过程中，存在副反应和未反应完全的杂质，生产反应后的物料要顺序经过不同的精馏塔，最后在产品塔获得纯度高于99%的产品，产品塔的质量检测尤为重要。由于产品塔检测点近红外光谱数据缺乏多样性，检测模型的泛化能力较弱。该研究采用偏最小二乘为2,6-二甲酚精馏提纯过程中不同检测点的数据集创建子空间，然后通过最小化其他检测点数据子空间与产品塔检测点数据子空间的布雷格曼(Bregman)散度，将其他检测点数据的子空间对齐到产品塔数据子空间，减小其他检测点数据子空间与产品塔检测点数据子空间的特征分布差异，既避免了投影到公共子空间产品塔检测点数据特征信息的损失，又能充分利用其他检测点数据的特征信息，然后在迁移后的子空间完成偏最小二乘回归建模，通过竞争学习加权策略确定最终的模型系数，从而提升产品塔检测模型的性能。在2,6-二甲酚纯度近红外检测数据集上进行了仿真验证，并探讨了迁移其他检测点不同数量的数据对产品塔检测模型性能的影响，产品塔检测模型的最大性能提升达到了52.19%，RMSEP值由0.059 4下降到0.028 4，与传统建模方法支持向量机回归和BP神经网络相比具有明显的优势。

关键词 近红外光谱；迁移学习；子空间对齐；2,6-二甲酚；精馏提纯

中图分类号：TP181 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)11-3608-07

引 言

物质浓度在线测量是产品质量控制的关键，主要通过测量样品体系中随物质浓度改变而变化的物理化学性质关联得到[1]。近红外光谱是一种先进的在线过程检测技术，其原理是使用13 333～4 000 cm^{-1}波长范围的电磁辐射探测样品获得光谱信息[2]，通过化学计量学方法建立光谱信息与物质浓度之间的关系，已被广泛用于化工[3-4]、制药[5]、生物[6]、食品[7]和医疗[8]等行业。近红外光谱检测结果的准确性与模型的质量密切相关，研究人员已经提出了很多方法来构建校正模型，比如偏最小二乘回归、多元线性回归、主成分回归等。随着机器学习技术的发展，Chen等[9]提出了一种用于建立光谱校正模型的贝叶斯方法，建立的贝叶斯模型具有较低的预测误差；Bian等[10]引入极限学习机算法用于复杂样品的光谱定量分析，兼顾了模型的预测精度和稳定性；Yang等[11]将深度学习用于光谱分析，提高了对数据集的特征提取能力，但是对样本量的需求较大。

在使用近红外光谱进行在线测量时，需要足够多的历史样本离线构建模型。然而在实际工业生产过程的最后阶段，样品的纯度越来越高，样本的多样性不足，物质浓度与光谱之间缺乏相关性。在这样的数据条件下，传统方法建立的模型很难达到期望的预测性能。例如，在重要的化工中间体2,6-二甲酚(2,6-dimethylphenol，2,6-DMP)的生产过程中，由于产品塔的产品纯度较高，近红外检测数据缺乏多样性，模型的泛化能力较弱。针对高质量训练数据较少的问题，迁移学习能够从相关领域迁移信息来改进目标领域的模型性能，从而减少对目标领域数据数量和质量的依赖[12]。Liu等[13]

收稿日期：2021-10-09，**修订日期**：2022-01-24

基金项目：国家自然科学基金项目(61991402,61833007)资助

作者简介：邬云飞，1998年生，江南大学自动化研究所硕士研究生 e-mail：yfwuu@vip.jiangnan.edu.cn

*通讯作者 e-mail：xlluan@jiangnan.edu.cn

采用迁移学习策略，扩展了样本的数量和多样性，提高了故障诊断结果的稳定性和准确性；Wang 等[14]运用基于局部相似特征选择的迁移学习方法，对不同原油在不同检测条件下的近红外光谱数据实现了快速建模；褚菲等[15]将迁移学习方法与多尺度核学习方法相结合，改善了间歇过程产品质量的预测精度。

借鉴迁移学习中不同数据域的知识传递思想，本研究针对高纯度的产品塔 2,6-DMP 在线检测问题，利用其他塔的光谱数据和产品塔光谱数据的相似性，提出了一种基于偏最小二乘(PLS)子空间对齐的迁移学习建模方法。首先借助偏最小二乘为产品塔和其他塔的数据集创建子空间，然后学习其他塔到产品塔数据集子空间的映射函数，并将其他塔的样本数据投影到对齐之后的子空间以生成新的特征表示，最后采用迁移之后的新特征建立模型。基于子空间的迁移学习中，通常采取寻找公共子空间或者构建一组中间表示的策略，不仅可能代价高昂，还会造成源域和目标域信息的损失。通过 PLS 子空间对齐可以直接比较各数据域的样本特征，无需进行其他的投影。该方法可以充分利用其他塔多样性较好的光谱数据，为产品塔建立具有可靠性和高预测精度的模型，从而实现 2,6-DMP 产品质量的实时调控。

1　实验部分

1.1　偏最小二乘回归算法

在产品分离提纯过程中，混合体系由多组分构成，组分的种类和含量未知，不同种类物质的近红外光谱特征吸收峰存在重叠，不能对一系列标准溶液做出校正曲线。偏最小二乘(partial least squares，PLS)方法通过在特征空间内提取主成分来描述光谱数据与纯度值之间的关系，适用于样本数较少而变量数较多的过程建模，能够用于产品纯度的在线测量[1]。

光谱数据 $\boldsymbol{X}=\{\boldsymbol{x}_i;\ i=1,\cdots,n\}$，与 2,6-DMP 纯度 $\boldsymbol{Y}=\{y_i;\ i=1,\cdots,n\}$，其中 n 是样本数目。PLS 分别从 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中提取主成分 $\boldsymbol{t}_1$ 和 $\boldsymbol{u}_1$，它们必须满足：

(1) $\boldsymbol{t}_1$ 和 $\boldsymbol{u}_1$ 应该分别携带尽可能多的 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的变异信息，即 $\mathrm{var}(\boldsymbol{t}_1)\rightarrow\max$，$\mathrm{var}(\boldsymbol{u}_1)\rightarrow\max$；

(2) $\boldsymbol{t}_1$ 和 $\boldsymbol{u}_1$ 之间的相关性最大，即 $r(\boldsymbol{t}_1,\boldsymbol{u}_1)\rightarrow\max$。

上述两个条件综合起来，即要求 $\boldsymbol{t}_1$ 和 $\boldsymbol{u}_1$ 的协方差达到最大

$$\mathrm{cov}(\boldsymbol{t}_1,\boldsymbol{u}_1)=\sqrt{var(\boldsymbol{t}_1)var(\boldsymbol{u}_1)}\,\mathrm{r}(\boldsymbol{t}_1,\boldsymbol{u}_1)\rightarrow\max \quad (1)$$

式(1)中，$\mathrm{cov}(\cdot,\cdot)$ 表示协方差，$\mathrm{var}(\cdot)$ 表示方差，$r(\cdot,\cdot)$表示相关系数。式(1)可转化成下列优化问题

$$\max_{\boldsymbol{w}_1,\boldsymbol{c}_1}\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{c}_1 \quad \text{s.t.}\begin{cases}\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{w}_1=1\\ \boldsymbol{c}_1^{\mathrm{T}}\boldsymbol{c}_1=1\end{cases} \quad (2)$$

式(2)中，$\boldsymbol{E}_0$ 和 $\boldsymbol{F}_0$ 分别是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 标准化后的矩阵，$\boldsymbol{w}_1$ 和 $\boldsymbol{c}_1$ 分别是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第一个轴。$\boldsymbol{w}_1$ 和 $\boldsymbol{c}_1$ 可以通过特征值分解法获得，$\boldsymbol{w}_1$ 是矩阵 $\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0$ 最大特征值的单位特征向量，$\boldsymbol{c}_1$ 是矩阵 $\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0$ 最大特征值的单位特征向量。

提取第一个主成分后，建立 $\boldsymbol{E}_0$，$\boldsymbol{F}_0$ 对 $\boldsymbol{t}_1$ 的回归模型，然后运用 $\boldsymbol{E}_0$，$\boldsymbol{F}_0$ 被 $\boldsymbol{t}_1$ 解释后的残差信息提取第二个主成分 $\boldsymbol{t}_2$，重复该过程，直到提取 A 个主成分满足精度要求。PLS 算法提取主成分过程如下($i=1:A$)：

(1) 通过特征值分解法获得 $\boldsymbol{w}_i$ 和 $\boldsymbol{c}_i$；

(2) 计算主成分 $\boldsymbol{t}_i$，载荷向量 $\boldsymbol{p}_i$，系数向量 $\boldsymbol{r}_i$ 以及残差信息 $\boldsymbol{E}_i$ 和 $\boldsymbol{F}_i$：

$$\begin{aligned}\boldsymbol{t}_i&=\boldsymbol{E}_{i-1}\boldsymbol{w}_i\\ \boldsymbol{p}_i&=\boldsymbol{E}_{i-1}^{\mathrm{T}}\boldsymbol{t}_i/\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_i\\ \boldsymbol{r}_i&=\boldsymbol{F}_{i-1}^{\mathrm{T}}\boldsymbol{t}_i/\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_i\\ \boldsymbol{E}_i&=\boldsymbol{E}_{i-1}-\boldsymbol{t}_i\boldsymbol{p}_i^{\mathrm{T}}\\ \boldsymbol{F}_i&=\boldsymbol{F}_{i-1}-\boldsymbol{t}_i\boldsymbol{r}_i^{\mathrm{T}}\end{aligned}$$

通过以上步骤，提取到主成分 $\boldsymbol{T}=(\boldsymbol{t}_1,\cdots,\boldsymbol{t}_A)$，获得载荷矩阵 $\boldsymbol{P}=(\boldsymbol{p}_1,\cdots,\boldsymbol{p}_A)$，系数矩阵 $\boldsymbol{R}=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_A)$和 $\boldsymbol{W}=(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_A)$。

主成分与矩阵 $\boldsymbol{E}_0$ 的关系为

$$\boldsymbol{T}=\boldsymbol{E}_0\boldsymbol{V} \quad (3)$$

式(3)中，$\boldsymbol{V}=\boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}$表示投影矩阵。

最终可以得到模型回归系数，见式(4)

$$\boldsymbol{\beta}=\boldsymbol{V}\boldsymbol{R}^{\mathrm{T}} \quad (4)$$

1.2　基于 PLS 的子空间对齐方法

子空间对齐(subspace alignment，SA)方法使用主成分分析(principal component analysis，PCA)为源域数据和目标域数据提取 d 个特征向量，作为源域和目标域子空间的基，用 $\boldsymbol{Z}_S$ 和 $\boldsymbol{Z}_T$ 表示。然后使用映射矩阵 $\boldsymbol{M}\in R^{d\times d}$对齐两个域的基向量，将源域子空间坐标系转换为目标域子空间坐标系，矩阵 $\boldsymbol{M}$ 通过最小化布雷格曼矩阵散度(Bregman matrix divergence)获得，见式(5)

$$F(\boldsymbol{M})=\|\boldsymbol{Z}_S\boldsymbol{M}-\boldsymbol{Z}_T\|_F^2 \quad (5)$$

因为 $\boldsymbol{Z}_S$ 和 $\boldsymbol{Z}_T$ 正交，所以满足 $\boldsymbol{Z}_S^{\mathrm{T}}\boldsymbol{Z}_S=\boldsymbol{I}_d$，$\boldsymbol{Z}_T^{\mathrm{T}}\boldsymbol{Z}_T=\boldsymbol{I}_d$，其中 $\boldsymbol{I}_d\in R^{d\times d}$为单位阵，可以得到最优的矩阵 $\boldsymbol{M}$[见式(6)]

$$\boldsymbol{M}^*=\boldsymbol{Z}_S^{\mathrm{T}}\boldsymbol{Z}_T \quad (6)$$

主成分回归进行降维时仅考虑光谱数据 $\boldsymbol{X}$，而在偏最小二乘回归中考虑了光谱数据 $\boldsymbol{X}$ 与纯度值 $\boldsymbol{Y}$ 之间的关系，不仅能概括光谱数据中所包含的信息，也能更好地解释纯度值。将子空间对齐方法拓展到偏最小二乘回归中，具体描述如下：

(1) 输入：其他塔的样本集($\boldsymbol{X}_S$，$\boldsymbol{Y}_S$)，产品塔样本集($\boldsymbol{X}_T$，$\boldsymbol{Y}_T$)，标准化处理后分别为($\boldsymbol{E}_S$，$\boldsymbol{F}_S$)和($\boldsymbol{E}_T$，$\boldsymbol{F}_T$)。

(2) 首先采用 PLS 算法分别获得其他塔和产品塔各自子空间的投影矩阵 $\boldsymbol{V}_S$ 和 $\boldsymbol{V}_T$；

(3) 借助子空间对齐方法获得其他塔的映射矩阵 $\boldsymbol{M}_S^*$[见式(7)]

$$\boldsymbol{M}_S^*=\arg\min F(\boldsymbol{M}_S)=\|\boldsymbol{V}_S\boldsymbol{M}_S-\boldsymbol{V}_T\|_F^2 \quad (7)$$

$$\boldsymbol{M}_S^*=\boldsymbol{V}_S^{+}\boldsymbol{V}_T$$

式(7)中，“+”表示广义逆；

(4) 计算迁移后其他塔的投影矩阵 $\boldsymbol{V}_{\mathrm{trans}}$和主成分 $\boldsymbol{T}_{\mathrm{trans}}$，见式(8)

$$\begin{aligned}\boldsymbol{V}_{\mathrm{trans}}&=\boldsymbol{V}_S\boldsymbol{M}_S^*\\ \boldsymbol{T}_{\mathrm{trans}}&=\boldsymbol{X}_S\boldsymbol{V}_{\mathrm{trans}}\end{aligned} \quad (8)$$

(5) 计算迁移后其他塔的载荷矩阵 $\boldsymbol{P}_{\text{trans}}$ 和系数矩阵 $\boldsymbol{R}_{\text{trans}}(i=1:A)$

$$\boldsymbol{t}_i = \boldsymbol{T}_{\text{trans}}(:, i)$$

$$\boldsymbol{p}_i = \frac{\boldsymbol{E}_{Si-1}^{\text{T}}\boldsymbol{t}}{\boldsymbol{t}_i^{\text{T}}\boldsymbol{t}_i}$$

$$\boldsymbol{r}_i = \frac{\boldsymbol{F}_{Si-1}^{\text{T}}\boldsymbol{t}_i}{\boldsymbol{t}_i^{\text{T}}\boldsymbol{t}_i}$$

$$\boldsymbol{E}_{Si} = \boldsymbol{E}_{Si-1} - \boldsymbol{t}_i\boldsymbol{p}_i^{\text{T}}$$

$$\boldsymbol{F}_{Si} = \boldsymbol{F}_{Si-1} - \boldsymbol{t}_i\boldsymbol{r}_i^{\text{T}}$$

$$\boldsymbol{P}_{\text{trans}}(:, i) = \boldsymbol{p}_i$$

$$\boldsymbol{R}_{\text{trans}}(:, i) = \boldsymbol{r}_i$$

(6) 计算迁移后的回归系数，见式(9)

$$\boldsymbol{\beta}_{\text{trans}} = \boldsymbol{V}_{\text{trans}}\boldsymbol{R}_{\text{trans}}^{\text{T}} \tag{9}$$

(7) 输出：采用竞争学习加权策略(winner-takes-all based weighting method)[16]，即计算产品塔回归系数 $\boldsymbol{\beta}_t$ 和迁移后其他塔回归系数 $\boldsymbol{\beta}_{\text{trans}}$ 对应的交叉验证均方根误差，选择误差较小的作为最终的模型回归系数 $\boldsymbol{\beta}_f$，算法流程如图 1 所示。

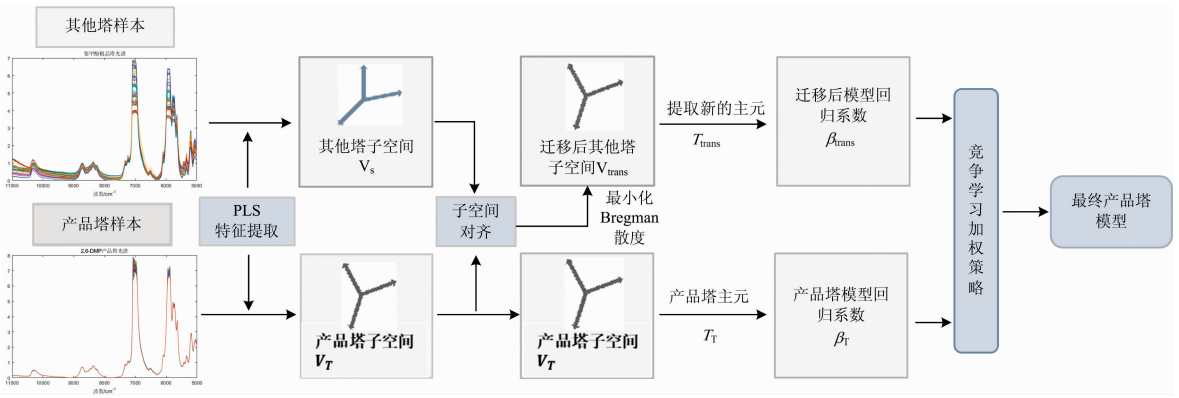

图 1 迁移学习算法建模流程图

Fig. 1 Modeling flow diagram of transfer learning algorithm

2 结果与讨论

2.1 过程描述

目前国内外合成 2,6-DMP 的主要方法有天然分离法、苯胺重氮化水解法、甲苯氯化水解法及苯酚烷基化法。工业上常用苯酚烷基化法，该方法选择性较高且成本较低，适宜连续生产，本文研究的 2,6-DMP 制备过程如图 2(a)所示。

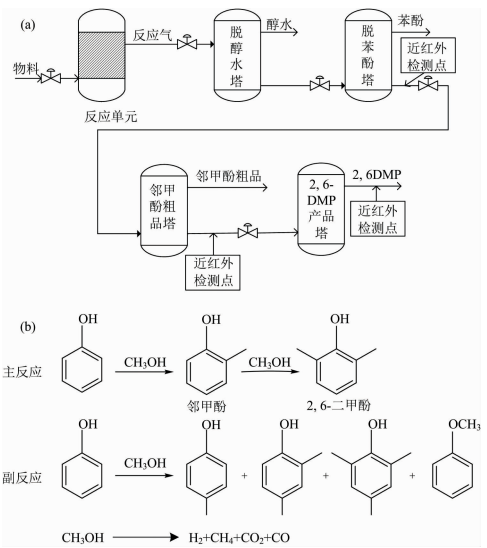

图 2 工艺流程与反应原理

(a)：工艺流程图；(b)：反应原理图

Fig. 2 Process flow diagrams and Reactive principle sketch

(a): Process flow diagram; (b): Reactive principle sketch

以苯酚和甲醇为原料，选择合适的催化剂后在固定床管式反应器进行烷基化反应，反应原理如图 2(b)所示。反应气依次经过脱醇水塔、脱苯酚塔和邻甲酚粗品塔，到达 2,6-DMP 产品塔，在产品塔顶部获得产品 2,6-DMP。由图 2(b)可知，产物含有邻甲酚及其他杂质，为了在线检测生产流程中各组分的含量，分别在脱苯酚塔的底部、邻甲酚粗品塔的底部和 2,6-DMP 产品塔的顶部安装了近红外光谱仪和检测探头，在线收集管道中物料的近红外光谱数据，通过已建立的模型得到产品纯度的预测值。

2,6-DMP 产品塔的产品纯度高，化验室通过气相色谱法标注的纯度值分布在一个较小的区间内，且存在很高的重复性，传统的建模方法无法精确建模。在产品 2,6-DMP 的精馏提纯过程中，不同检测点处样品有机物的含量和种类不同，脱苯酚塔和邻甲酚粗品塔检测点采集的光谱数据与产品

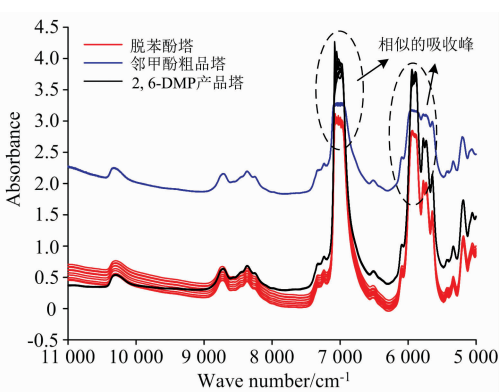

图 3 不同检测点处的光谱比较

Fig. 3 Spectral comparison at different detecting points

塔存在差异，但是样品中有机成分种类有重合，近红外光谱在相同波数处存在相似的的吸收峰，如图 3 所示。本工作提出的一种基于 PLS 子空间对齐的迁移学习建模方法能够利用这种相似性，借助其他检测点的光谱数据，有效提升产品塔检测点模型的性能。

2.2　数据

建模所用的原始光谱来自某合成材料公司的 2,6-DMP 制备过程。使用布鲁克在线近红外光谱仪采集样本光谱，扫描光谱范围为 12 500～4 000 cm^{-1}，实际使用范围 11 000～5 000 cm^{-1}。近红外光谱数据集对应的 2,6-DMP 纯度值由化验室通过气相色谱法分析离线获得，各个检测点的纯度值数据特征如表 1 所示。采用变异系数来衡量样本离散程度，因为标准差是一个绝对指标，当用其来对同一总体的不同时期进行对比时，由于平均值不同，缺乏可比性。变异系数是标准差与平均值的比值，可以消除平均值不同对样本集离散程度对比的影响。通过比较变异系数发现，产品塔的变异系数明显小于脱苯酚塔和邻甲酚粗品塔，说明产品塔的样本区分度低。

表 1　不同检测点的 2,6-DMP 纯度值特征

Table 1　2,6-DMP purity distribution at different detecting points

样本集	样本数	均值/%	变异系数/%	最小值/%	最大值/%
脱苯酚塔检测点	300	81.32	2.270	73.90	83.97
邻甲酚粗品塔检测点	300	97.49	0.370	96.47	98.67
2,6-DMP 产品塔检测点	50	99.86	0.031	99.81	99.95

2.3　实验结果

(1)将 2,6-DMP 产品塔检测点获得的数据集中 30 组数据作为训练集，20 组数据作为测试集。模型性能的评价指标采用近红外光谱分析方法中最常用的预测均方根误差(root mean square error of prediction，RMSEP)[2]，计算公式如式(10)

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{k=1}^{r}(\hat{y}_k-y_k)}{r-1}} \tag{10}$$

式(10)中，r 为测试集的样本个数，$\hat{y}_k$ 为第 k 个测试样本的纯度预测值，y_k 为第 k 个测试样本的实际纯度值。

为了更直观地观察 PLS 子空间对齐方法的效果，引入指标性能提升百分比 IP，计算公式如式(11)所示。

$$\mathrm{IP}=\frac{\mathrm{RMSEP}_{\mathrm{PLS}}-\mathrm{RMSEP}_{\mathrm{PLS\text{-}SA}}}{\mathrm{RMSEP}_{\mathrm{PLS}}} \tag{11}$$

式(11)中，$\mathrm{RMSEP}_{\mathrm{PLS}}$ 为仅使用产品塔训练集训练 PLS 模型的预测均方根误差，$\mathrm{RMSEP}_{\mathrm{PLS\text{-}SA}}$ 为使用 PLS 子空间对齐方法训练模型的预测均方根误差。

(2)分析不同数量的辅助光谱对 2,6-DMP 产品塔模型性能的影响，将脱苯酚塔检测点和邻甲酚粗品塔检测点采集的光谱数据，按照不同数量(30～300)依次加入到产品塔训练集中。为了证明 PLS 子空间对齐方法的有效性，与传统机器学习方法支持向量机回归和 BP 神经网络进行了比较，支持向量机回归选择高斯核函数，脱苯酚塔样本作为辅助数据时核参数选择 4.5，邻甲酚粗品塔样本作为辅助数据时核参数选择 5.5；BP 神经网络隐藏层神经元个数选择 5，最大迭代次数为 100。

PLS 子空间对齐方法中唯一的参数是主成分数，选取与产品塔训练集同样数目的脱苯酚塔样本和邻甲酚粗品塔样本，比较不同主成分数下的模型性能，结果如图 4 所示。在主成分数较少时，模型的预测均方根误差较小，性能较高，最后选择因子数为 7，此时仅使用产品塔训练集构建模型的预测均方根误差值为 0.059 4。

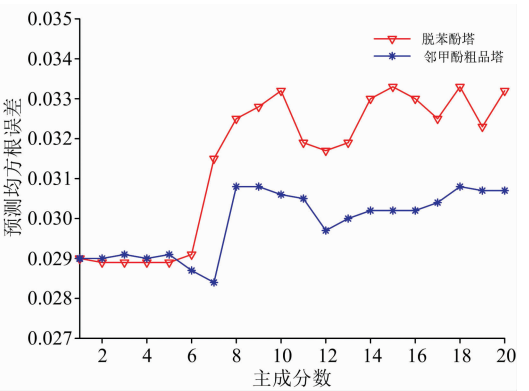

图 4　不同主成分数对模型性能的影响

Fig. 4　Different principal component numbers impact on model performance

图 5(a)和图 6(a)分别是迁移脱苯酚塔不同数量样本辅助建模所得的建模误差和迁移邻甲酚粗品塔不同数量样本辅助建模所得的建模误差，红色曲线表示运用 PLS 子空间对齐方法的建模误差，蓝色曲线表示其他塔样本与产品塔样本合并后运用 PLS 算法的建模误差，黑色曲线表示运用支持向量机回归算法的建模误差，绿色曲线表示使用 BP 神经网络算法的建模误差。图 5(b)和图 6(b)分别是迁移脱苯酚塔样本数据后的性能提升百分比和迁移邻甲酚粗品塔样本数据后的性能提升百分比。

从图 5 和图 6 中可以看出，相较于传统方法，PLS 子空间对齐方法对产品塔的模型性能有明显的提升。随着样本数的增加，模型的性能提升呈下降趋势，且脱苯酚塔的样本作为辅助数据，在样本量超过 240 时，对模型性能已没有提升，表明随着辅助数据数量的增加，引入了对产品塔模型有害的样本，导致了负迁移。

观察图 5 和图 6 可知，借助邻甲酚粗品塔 30 个样本时，模型性能的提升最大，此时预测均方根误差为 0.028 4，性能提升百分比为 52.19%。图 7(a)是此时的模型曲线，图 7(b)是预测值与实际值的散点图。从图 7 可以看出，PLS 子空间对齐方法建立的模型预测效果更好。

3　结　论

针对产品生产最后阶段物质浓度提高，样本区分度低，多样性不足，无法精确建模的问题，提出了一种基于 PLS 子

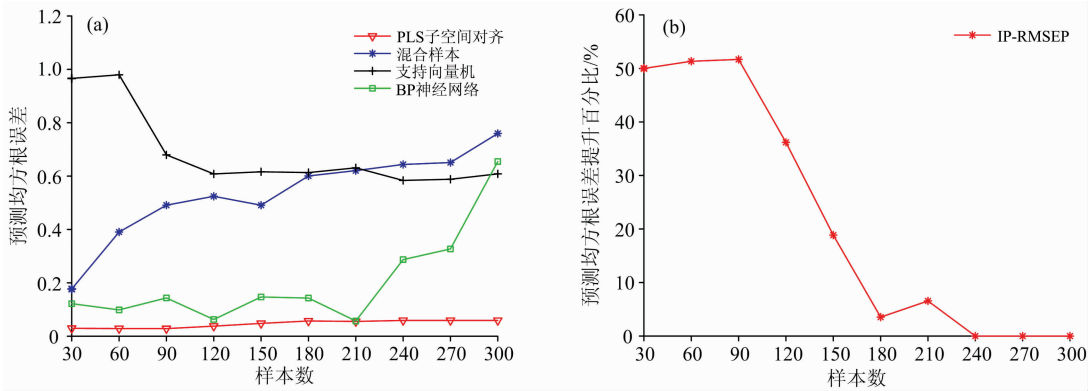

图 5　迁移脱苯酚塔不同样本数对模型性能的影响

(a)：迁移脱苯酚塔不同样本数对模型性能的影响；(b)：迁移脱苯酚塔不同样本数的模型性能提升百分比

Fig. 5　Different sample numbers of dephenolization tower impact on model performance

(a)：Impact on model performance for transferring different sample numbers of dephenolization tower;

(b)：Model performance improvement percentage for transferring different sample numbers of dephenolization tower

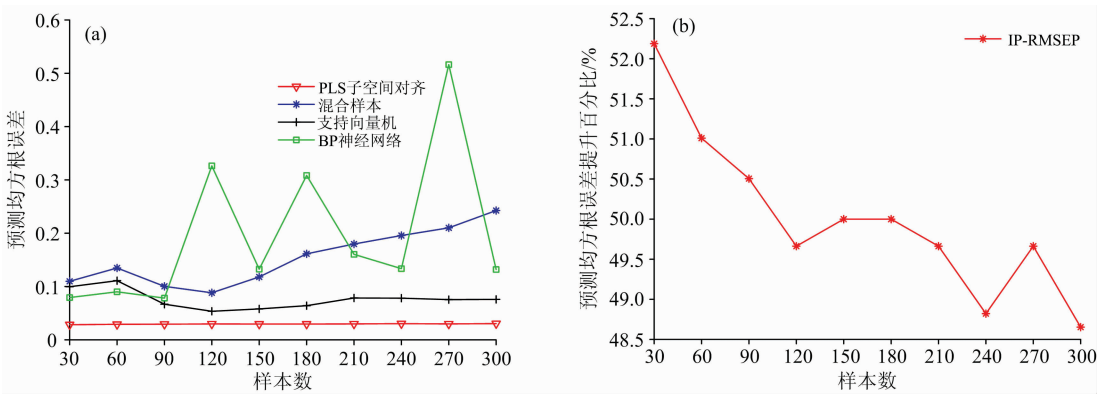

图 6　迁移邻甲酚粗品塔不同样本数对模型性能的影响

(a)：迁移邻甲酚粗品塔不同样本数对模型性能的影响；(b)：迁移邻甲酚粗品塔不同样本数的模型性能提升百分比

Fig. 6　Different sample numbers of o-cresol tower impact on model performance

(a)：Impact on model performance for transferring different sample numbers of o-cresol tower;

(b)：Model performance improvement percentage for transferring different sample numbers of o-cresol tower

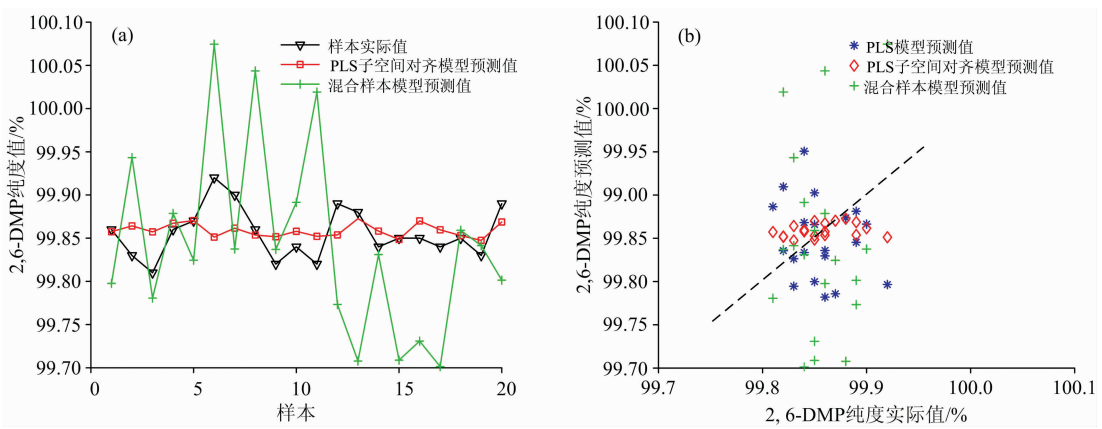

图 7　模型曲线与散点图

(a)模型曲线；(b)：预测值与实际值散点图

Fig. 7　Model curve and Scatter plot

(a)：Model curve；(b)：Scatter plot of prediceted and actual values

空间对齐的迁移学习建模方法。在偏最小二乘回归算法的基础上，为两个域创建子空间，然后最小化布雷格曼矩阵散度从而获得源子空间到目标子空间的映射，完成源域特征到目标域特征的迁移。在某公司制备 2,6-二甲酚过程的近红外检测数据集上进行了仿真验证，结果表明，所提方法能够借助其他塔的数据来提升产品塔高浓度产品检测模型的稳定性和准确性，具有一定的实用价值。在后续的工作中将进一步研究迁移模型性能与样本数量之间的定量关系以及如何避免负迁移。

References

[1] WANG Jin-jin, LI Xiu-xi(王津津, 李秀喜). Chemical Industry and Engineering Progress(化工进展), 2011, 30(10): 2163.
[2] Pasquini C. Analytica Chimica Acta, 2018, 1026: 8.
[3] Wang K, He K, Du W, et al. Chemical Engineering Science, 2021, 242: 116672.
[4] Lian X, Zhang M, Sun X, et al. Polymer Testing, 2021, 93: 106584.
[5] Cogoni G, Liu Y A, Husain A, et al. International Journal of Pharmaceutics, 2021, 602: 120620.
[6] Hebbi V, Thakur G, Rathore A S. Journal of Biotechnology, 2021, 325: 303.
[7] Loudiyi M, Temiz H T, Sahar A, et al. Critical Reviews in Food Science and Nutrition, 2022, 62(11): 3063.
[8] Yu Y, Huang J, Zhu J, et al. IEEE Sensors Journal, 2020, 21(3): 3506.
[9] Chen T, Martin E. Analytica Chimica Acta, 2009, 631(1): 13.
[10] Bian X H, Li S J, Fan M R, et al. Analytical Methods, 2016, 8(23): 4674.
[11] Yang J, Xu J, Zhang X, et al. Analytica Chimica Acta, 2019, 1081: 6.
[12] Pan S J, Yang Q. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345.
[13] Liu J, Ren Y. IEEE Transactions on Industrial Informatics, 2021, 17(9): 6073.
[14] Wang Y, Wang K, Zhou Z, et al. Frontiers of Chemical Science and Engineering, 2019, 13(3): 599.
[15] CHU Fei, PENG Chuang, JIA Run-da, et al(褚　菲, 彭　闯, 贾润达, 等). CIESC Journal(化工学报), 2021, 72(4): 2178.
[16] Shan P, Zhao Y, Wang Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 215: 97.

Transfer Learning Modeling of 2,6-Dimethylphenol Purity Based on PLS Subspace Alignment

WU Yun-fei, LUAN Xiao-li*, LIU Fei

Key Laboratory of Advanced Process Control for Light Industry of the Ministry of Education, Institute of Automation, Jiangnan University, Wuxi　214122, China

Abstract　The highly accurate on-line detection of solute concentration by using near-infrared spectroscopy analysis technology is of great significance for production optimization. Establishing a detection model needs to extract relevant information from the near-infrared spectrum, more representative samples will extract more effective information, and the model will also be more accurate. However, the purity of products has been continuously improved, and sample discrimination is reduced. The coefficient of variation of the sample is small, which leads to the diversity of samples being insufficient. Moreover, there are measurement errors when measuring the sample concentration in the laboratory, which will lead to the lack of correlation between the solute concentration and the spectrum. It is hard to establish a reliable and highly accurate near-infrared detection model. A new transfer learning modeling method based on PLS subspace alignment is proposed and applied in the near-infrared on-line detection of 2,6-dimethylphenol high purity in the product tower of the distillation purification process. In preparing the chemical monomer 2,6-dimethylphenol, there are side reactions and unreacted impurities, the materials after reaction must flow through different rectifying towers in sequence, and finally, the product with purity higher than 99% is obtained in the product tower. The product quality inspection of the product tower is particularly important. Due to the lack of diversity of the detection data of product tower detection point, the generalization ability of the detection model is weak. We create subspaces for the data sets of different detection points in the separation and purification of 2,6-dimethylphenol using the partial least squares method. Then, a mapping that aligns the other tower subspace into the product tower is learned by minimizing a Bregman matrix divergence function, reducing the feature distribution discrepancy between other towers and product towers. It avoids information loss in product tower data when projecting to a common subspace and makes full use of the feature information in other towers. The partial least squares regression modeling is completed on the transferred subspace and the final model

coefficients are determined by the winner-takes-all-based weighting method. After the above method, the product tower detection model's performance has improved. Finally, the effectiveness of this method is verified with the near-infrared detection data set of 2,6-dimethylphenol. We discuss the influence of transferring different amounts of data from other detection points on the performance of the product tower detection model. The performance improvement of the product tower detection model is improved by 52.19% in the best case, and the root means square error of prediction (RMSEP) dipped from 0.059 4 to 0.028 4. Compared with the traditional modeling methods such as support vector machine regression (SVR) and Back-Propagation (BP) neural network, it has better performance.

Keywords Near infrared spectroscopy; Transfer learning; Subspace alignment; 2,6-dimethylphenol; Distillation purification

(Received Oct. 9, 2021; accepted Jan. 24, 2022)

* Corresponding author

本 刊 声 明

近期以来，一些不法分子假冒《光谱学与光谱分析》期刊社名义，以虚假网站等形式欺骗广大作者、读者。这些虚假网站公然假冒《光谱学与光谱分析》期刊名义进行大肆的征稿并骗取作者的审稿费和版面费。经部分作者及读者举报，现有关部门已就此介入调查。我刊将通过法律途径向假冒者追究相应的责任，维护本刊权益。

本刊官方网站已正式开通，网址为

http://www.gpxygpfx.com/

在此郑重声明，本网址为《光谱学与光谱分析》期刊唯一开通运行的官方网站。本刊从未授权任何单位或个人以任何形式（包括网上网下）代理本刊征稿、审稿等项业务。

希望广大读者和作者切实维护好自身的合法权益，防止受骗上当。

《光谱学与光谱分析》期刊社

2019 年 3 月 15 日