DOI：10.13382/j.jemi.B2307034

融合相似性度量加权核偏最小二乘的烷烃气体定量分析方法*

李忠兵[1]　刘雅杰[1]　梁海波[2]　倪朋勃[3]　闫　碧[1]
（1.西南石油大学电气信息学院　成都　610500；2.西南石油大学机电工程学院　成都　610500；
3.中法渤海地质服务有限公司　天津　300457）

摘　要：烃类气体含量的有效监测是油气勘探开采过程中安全保障的重要环节。红外光谱法作为一种安全高效的检测方法，受到现场工程师的关注，但主要采用离线模型进行测量，无法较好应对现场复杂的工况及变化多样的非线性影响因素，导致离线模型不更新而难以维持较高的预测精度。为此，提出了一种融合相似性度量加权核偏最小二乘的即时学习建模策略。首先设计了一种多相似性度量准则融合的样本相似性判别依据，有效筛选历史样本用于在线建模，其次在局部 PLS 模型中引入非线性核函数，实现非线性特征的有效提取，弥补线性偏最小二乘模型的非线性处理能力。在构建的多组分混合气体红外光谱数据上的实验结果验证了该方法的有效性，拟合优度 R^2 达到 0.994 1，*RMSE* 和 *MRE* 相比 PLS 模型分别提升了 43.6%和 85.8%，可有效用于烃类气体红外光谱定量分析模型的在线更新与高精度预测。

关键词：烷烃气体；红外光谱；即时学习；相似性度量；非线性核函数
中图分类号：TH741　　**文献标识码**：A　　**国家标准学科分类代码**：410.55

Weighted kernel partial least squares method based on fusion of similarity measurement criteria for quantitative analysis of alkane gases

Li Zhongbing[1]　Liu Yajie[1]　Liang Haibo[2]　Ni Pengbo[3]　Yan Bi[1]
（1. School of Electrical Engineering and Information, Southwest Petroleum University, Chengdu 610500, China;
2. School of Mechatronic Engineering, Southwest Petroleum University, Chengdu 610500, China;
3. China-France Bohai Geo services Co., Ltd, Tianjin 300457, China）

Abstract: The effective monitoring of hydrocarbon gas content is an important aspect of safety assurance in oil and gas exploration and production processes. Infrared spectroscopy, as a safe and efficient detection method, has attracted the attention of on-site engineers. However, it mainly uses offline models for measurement, which cannot cope with the complex working conditions and various nonlinear influencing factors on site, making it difficult for this non updated model to maintain high prediction accuracy. A weighted kernel partial least squares method based on fusion of similarity measurement criteria in just-in-time learning for quantitative analysis of alkane gases is proposed in this paper. Firstly, a similarity criterion based on fusion of multiple similarity measurement criteria is designed to effectively select historical samples for online modeling. Secondly, nonlinear kernel functions are introduced into local PLS models to effectively extract nonlinear features and compensate for the nonlinear processing ability of linear partial least squares models. The experimental results on the multi-component mixed gas infrared spectral data have verified the effectiveness of this method, with a goodness of R^2 of 0.994 1. Compared with that of the PLS model, the *RMSE* and *MRE* of the proposed model have improved by 43.6% and 85.8%, respectively. It can be effectively used for online updating and high-precision prediction of infrared spectral quantitative analysis models for hydrocarbon gas.

Keywords: alkane gas; infrared spectrum; just-in-time learning; similarity measurement; nonlinear kernel function

收稿日期：2023-11-09　　Received Date: 2023-11-09
*基金项目：油气藏地质及开发工程国家重点实验室开放基金项目（PLN2022-42）、国家自然科学基金项目（52074233）、四川省自然科学基金项目（2024NSFSC0202）、油气生产安全与风险控制重庆市重点实验室开放基金项目（cqsrc202101）资助

0　引　言

在油气钻井过程中,储层中烷烃气体压力失衡,极有可能导致井漏、井涌、井塌、溢流等安全事故。因此,储层中烷烃气体精准发现和评价,对提升油气勘探技术,实现油气的高效开发,保障油气生产安全和避免能源浪费具有重要意义,同时也是油气行业实现可持续发展的必经之路[1-2]。在气测录井中,红外光谱在线检测技术因其快速测量的优点被广泛应用[3-4],但是,红外光谱在线检测常常受生产条件、性能参数、生产过程环境等变化的影响,致使建立的离线模型不能适应工作现场的过程特性和环境变化,从而导致模型精度下降,极大地限制了红外光谱在线检测技术的发展,迫切需要红外光谱定量分析模型需要具备一定的自更新能力。

Cybenko[5] 首次提出即时学习概念(just-in-time learning,JITL),这是一种擅长应对过程特性与状态变化的局部学习策略,常被应用于自适应建模以克服非线性过程时变造成的模型失衡问题[6],Kim 等[7] 将即时学习引入了近红外光谱建模技术中,表明比传统建模方法更有效。在即时学习建模策略中,主要针对相似性度量方法和局部模型的选取开展研究[8]。常用的相似性度量方法大多基于空间距离,比如欧氏距离[9]、马氏距离[10]等。Pan 等[11] 提出组合多个加权欧氏距离的相似性度量来定义混合加权性相似性度量,提高了片剂颗粒药品样本的选择性能,表明欧氏距离可适用于不同分子结构的样品样本的筛选。Yuan 等[12] 采用欧氏距离、余弦角度等信息进行集成用于表征样本相似性,提高了局部模型的预测精度,随着样本相似性度量方法的广泛研究,不同维度的信息被用于相似性度量。Uchimaru 等[13] 采用回归系数衡量样本相似性大小,在近红外领域建立了待测样本与历史样本间的稀疏回归模型,验证了该方法在近线性关系较好的场合具有不错的效果。为了提高即时学习模型的性能,Guo 等[14] 引入互信息对输入变量进行选择,将互信息作为传统变分自编码器模型变量的权重,有效提高了数据样本较少时模型的预测精度。Xia 等[15] 使用多个余弦相似性学习器来保证多样性,在差异较大的多种数据集中展示出优越性。相似性度量方法的选取从基本的空间距离向多种相似角度扩展,但在红外光谱气测录井中,甲烷(C_1)、乙烷(C_2)、丙烷(C_3)、正丁烷(nC_4)、异丁烷(iC_4)、正戊烷(nC_5)、异戊烷(iC_5)等烷烃气体组分为重点检测对象,属于同系有机物,其分子结构高度相似,同时受其他未知气体的影响,其红外光谱呈现高相似、谱峰重叠严重及非线性等特点,常用的相似性度量方法难以应对,对烷烃气体红外光谱相似样本的筛选仍存在考虑不充分的问题。

因强大的降维能力,偏最小二乘算法(partial least squares,PLS)通常选用为即时学习中的局部建模方法[16]。然而 PLS 方法是基于因变量和自变量之间数据信息的统计相关性构建的,要求因变量与自变量之间具有较强的线性相关性。为了克服已有 PLS 建模技术非线性的不足,增强模型的可解释性,许多学者展开了研究。Kim 等[17] 提出局部加权偏最小二乘建模模型,其效果通过工业实践得到证明,可以处理过程特性的变化以及过程的非线性。潘贝等[18] 提出一种基于多样性加权相似度的集成局部加权偏最小二乘软测量建模方法,有效提升了难测变量的预测精度。Chen 等[19] 考虑原始数据信息的重要性,提出改进的递归局部加权偏最小二乘方法,并在黄酒发酵过程的近红外光谱数据集上进行相关实验,其结果表明该方法比原有局部加权偏最小二乘方法更有效。局部加权偏最小二乘方法虽然增加了模型鲁棒性,但无法解决无关变量以及噪声冗余对模型的干扰,且 PLS 算法本身的线性特征,会对异常数据产生高敏感度。Rosipal 等[20] 将核函数算子引入到 PLS 算法中,提出了核偏最小二乘(kernel partial least squares,KPLS)算法,该算法可以有效处理各种不同非线性数据。之后,KPLS 优化成为大量学者在工业生产过程参数检测领域的热点研究方向。Zhao 等[21] 将数据集划分成多个子集,采用智能优化算法选择最合适的子集建立 KPLS 模型,有效预测了样本输出。Mello-Roman 等[22] 采用基因算法确定最大化模型性能的核函数参数,从而提高 KPLS 回归预测能力。一方面,偏最小二乘法等线性局部模型无法解决红外光谱气测录井面临的高相似、谱峰重叠严重及强非线性问题,尤其随着光谱数据维数越来越高,还掺杂大量不相关变量及噪声,制约了模型的预测性能[23]。另一方面,现有核函数处理的非线性数据为离线数据,未考虑现场测试数据特点,仍存在离线数据不能表征现场待测数据与历史数据相关性的问题,难以适应复杂的过程。

针对上述问题,本文从欧氏距离、马氏距离、余弦距离、光谱梯度角、皮尔逊相关系数、光谱信息散度 6 个角度考虑烷烃气体红外光谱样本在距离、角度、形状及信息熵等层面的相似性,提出了一种相似度融合策略,并在即时学习建模中引入非线性核函数,构建了加权核偏最小二乘法局部建模方法,使待测样本和历史样本通过相似度融合准则在迭代过程中更新构建新的权重矩阵,并与输出变量关联,改善线性 PLS 模型的非线性解释能力,有效降低强干扰与冗余数据的影响。本文的主要贡献如下:1)提出了一种相似度融合准则,充分利用距离、角度、形状、信息熵等光谱信息衡量待测样本和历史数据样本之间的相似性,构建具有多样性的相似度指标。2)提出了一种加权核偏最小二乘测量模型,相似性权重矩阵加权到历史输入、输出样本,

充分融合现有的数据信息,优选核函数增加模型的非线性解释能力。3)在烷烃类混合气体的红外光谱数据集中开展大量实验,验证所提模型对烷烃气体红外光谱定量分析精度的影响。

1　方法原理

即时学习算法是一种基于数据驱动的建模策略,根据当前状态信息在历史数据样本集中搜索相似的样本集,以此建立预测模型[24]。在即时学习中,相似性度量方法的准确建立是局部测量模型拥有良好预测性能的前提,局部测量模型能够准确表征数据特征是即时学习克服过程时变和解决非线性问题的保障。因此,本文提出一种融合相似性准则的加权核偏最小二乘(SF-KPLS)即时学习框架,融合多种相似性度量准则加权于历史输入、输出样本信息数据中,从多角度充分提取相似样本用于在线建模,并在PLS模型中优选非线性核函数来增强线性回归模型的非线性解释能力,提高烷烃气体红外光谱分析模型的预测精度。

1.1　融合相似性度量准则

基于即时学习模型的红外光谱分析能够从历史数据中找出与当前待测样本模态相匹配的数据样本,依据新数据样本在线滚动建立局部模型,使得模型不断的更新和优化,能够较为充分的考虑到过程阶段与状态变化。针对烷烃气体的红外光谱过程数据分布失衡问题,经过相似度度量方法计算,可以根据与当前待测样本相似度的大小,为每个历史数据设置权重。

为了更有效筛选相似样本,本文从距离、角度、形状、信息熵等不同角度,选取欧氏距离、马氏距离、余弦距离、光谱梯度角、皮尔逊相关系数、光谱信息散度6个相似性度量准则进行加权融合,其融合公式为:

$$w_f = \sum_{i=1}^{6} a_i w_i \tag{1}$$

式中:w_f为融合相似度值,w_i(i=1,2,…,6)依次表示欧氏距离、马氏距离、余弦距离、光谱梯度角、皮尔逊相关系数、光谱信息散度相似度量值,a_i为各相似度量值的融合加权系数,满足$\sum_{i=1}^{6} a_i = 1$。为了兼顾即时学习模型的精度及建模速度,加权系数a_i由各单一相似性度量准则建立定量分析模型时的预测精度$RMSE$值更新确定,其公式为:

$$a_i = \frac{\dfrac{1}{RMSE_i}}{\sum_{i=1}^{6} \dfrac{1}{RMSE_i}} \tag{2}$$

将距离、角度、形状、信息熵等不同角度的相似性度量方法进行融合加权,更加全面地衡量待测数据样本和历史数据样本特征之间的关联性,筛选出更加贴近真实的模型训练样本,通过为较小$RMSE$相似性度量方法分配更大的权重,使得局部测量模型精度得到提升。

1.2　相似性加权核偏最小二乘在线建模

当工业过程数据呈现线性特性的时候,PLS算法建模可以取得较理想的质量预测结果,但目前工业过程极其复杂,数据呈现非线性特点,将数据样本集采用核函数投影到高维空间,能够给出工业过程中的过程变量矩阵和质量变量矩阵之间的某些特殊信息,以解决原始空间中数据的非线性问题。

基于非线性核函数分析技术,首先需要构造一个合适的投影函数f,将原始光谱数据输入变量x投影到高维特征空间:

$$f(x) \in F \subseteq \Re^n \tag{3}$$

其中,$\Re^n$是n维希尔伯特空间,F表示高维特征空间,需要注意的是,特征空间的维度是任意的,可以为无限大。通过使用合适的非线性映射函数,可以把原始低维空间中线性不可分数据集映射到高维空间中变得线性可分,核映射理论基本原理如图1所示。

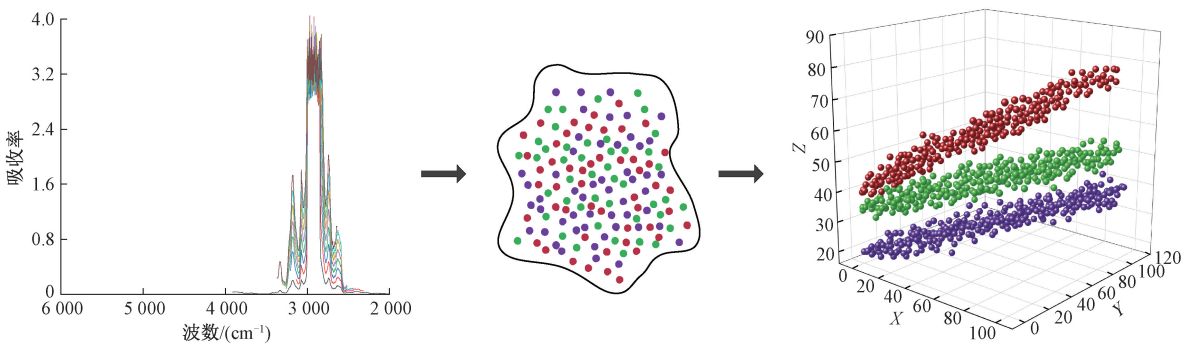


图1　核映射理论基本原理

Fig. 1　Basic principles of kernel mapping theory

融合相似性加权核偏最小二乘(SF-KPLS)首先通过核函数将原始数据集的输入数据和待测样本的输入数据 x_q 映射到高维特征空间 $\boldsymbol{\Phi}_{train}$ 和 φ_q,进而得到核特征矩阵 $\boldsymbol{K}_{train}=\boldsymbol{\Phi}_{train}\boldsymbol{\Phi}_{train}^{\mathrm{T}}$ 与核向量 $\boldsymbol{K}_q=\varphi_q\boldsymbol{\Phi}_{train}^{\mathrm{T}}$。然后根据融合相似性度量高维特征空间中待测样本与各个训练样本的对应样本权值 $\boldsymbol{\omega}$,把各个训练样本对应的权值构成权重矩阵 $\boldsymbol{\Omega}=\mathrm{diag}\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_N\}$。通过权重矩阵对高维特征空间中的样本加权后,运用 KPLS 回归算法建立局部模型来估计当前待测样本的输出 y_q,SF-KPLS 算法结构流程如图 2 所示。

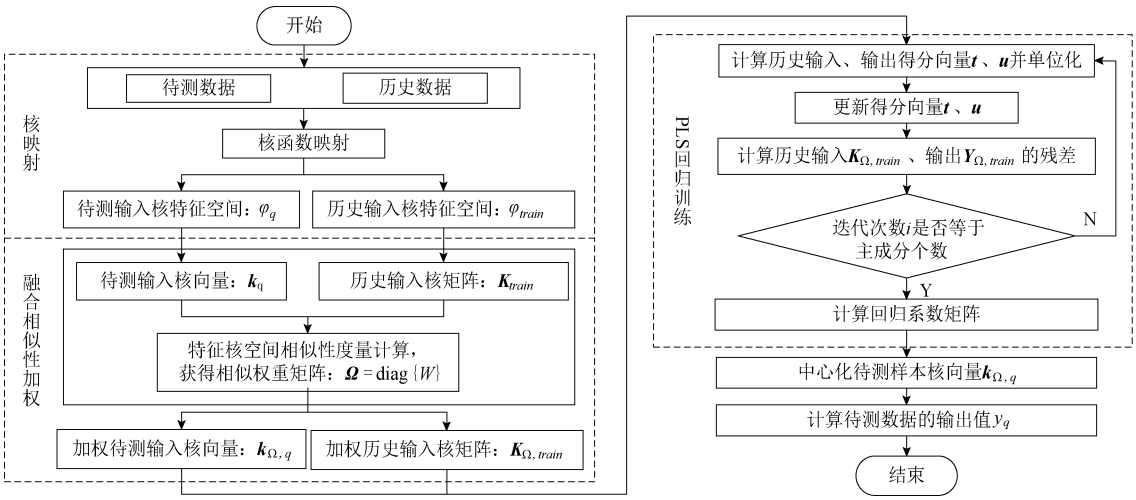


图 2　相似性度量加权核偏最小二乘在线建模流程

Fig. 2　Online modeling process of similarity measurement weighted kernel partial least squares

其中,在高维特征空间中采用式(4)对 $\boldsymbol{\Phi}_{train}$ 和对应输出 $\boldsymbol{Y}$ 进行加权,在一定程度上解决最小二乘类算法对异常数据敏感的问题。

$$\begin{cases}\boldsymbol{\Phi}_{\Omega,train}=\boldsymbol{\Omega}\boldsymbol{\Phi}_{train}\\ \boldsymbol{Y}_{\Omega,train}=\boldsymbol{\Omega}\boldsymbol{Y}\end{cases}\tag{4}$$

由此可得核矩阵 $\boldsymbol{K}_{train}$ 与核向量 $\boldsymbol{K}_q$ 的加权形式如式(5)所示。

$$\begin{cases}\boldsymbol{K}_{\Omega,train}=\boldsymbol{\Phi}_{\Omega,train}\boldsymbol{\Phi}_{\Omega,train}^{\mathrm{T}}=\boldsymbol{\Omega}\boldsymbol{\Phi}_{train}\boldsymbol{\Phi}_{train}^{\mathrm{T}}\boldsymbol{\Omega}^{\mathrm{T}}=\boldsymbol{\Omega}\boldsymbol{K}_{train}\boldsymbol{\Omega}^{\mathrm{T}}\\ k_{\Omega,q}=\varphi_q\boldsymbol{\Phi}_{\Omega,train}^{\mathrm{T}}=\varphi_q\boldsymbol{\Phi}_{train}^{\mathrm{T}}\boldsymbol{\Omega}^{\mathrm{T}}=k_q\boldsymbol{\Omega}^{\mathrm{T}}\end{cases}\tag{5}$$

根据非线性核函数的 PLS 算法对待测样本按如下步骤建立局部加权回归模型:

步骤 1)令 $i=1$,$\boldsymbol{K}_1=\boldsymbol{K}_{\Omega,train}$,$\boldsymbol{Y}_1=\boldsymbol{Y}_{\Omega,train}$,初始化 u_i、主成分个数 H、收敛条件 δ 和最大迭代次数 γ(可以设置 u 等于输出变量 $\boldsymbol{Y}$ 中的任何一列,本文将浓度值作为二者的初始值);

步骤 2)计算 $\boldsymbol{K}_i$ 的得分向量 $\boldsymbol{t}_i$,并单位化:

$$\boldsymbol{t}_i=\frac{\boldsymbol{K}_i u_i}{\|\boldsymbol{K}_i u_i\|}\tag{6}$$

步骤 3)计算 $\boldsymbol{Y}_i$ 的得分向量 $\boldsymbol{u}_i$,并单位化:

$$\boldsymbol{u}_i=\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{t}_i}{\|\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{t}_i\|}\tag{7}$$

步骤 4)如果满足收敛条件 $\frac{\|\boldsymbol{t}_i-\boldsymbol{t}_{i-1}\|}{\|\boldsymbol{t}_{i-1}\|}\leqslant\delta$ 或者达到最大迭代次数 γ,转至步骤 5);否则转至步骤 2);

步骤 5)令 $\boldsymbol{t}_{i+1}=\boldsymbol{t}_i$,$\boldsymbol{u}_{i+1}=\boldsymbol{u}_i$ 并计算残差:

$$\boldsymbol{K}_{i+1}=(\boldsymbol{I}-\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}})\boldsymbol{K}_i(\boldsymbol{I}-\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}})=$$
$$\boldsymbol{K}_i-\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{K}_i-\boldsymbol{K}_i\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}}+\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{K}_i\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}}\tag{8}$$

$$\boldsymbol{Y}_{i+1}=(\boldsymbol{I}-\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}})\boldsymbol{Y}_i=\boldsymbol{Y}_i-\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{Y}_i\tag{9}$$

步骤 6)判断 i 是否等于 H,若是,则停止迭代,算法结束,转至步骤 8),若否,则转至步骤 2);

步骤 7)根据以上步骤。分别得到关于输入 $\boldsymbol{K}_{\Omega,train}$、输出 $\boldsymbol{Y}_{\Omega,train}$ 的主成分得分矩阵 $\boldsymbol{T}=[t_1,t_2,\cdots,t_N]$、$\boldsymbol{U}=[u_1,u_2,\cdots,u_N]$,则回归系数矩阵为:

$$\boldsymbol{B}=\boldsymbol{\Phi}_{\Omega,train}^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{T}^T\boldsymbol{K}_{\Omega,train}\boldsymbol{U})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y}_{\Omega,train}\tag{10}$$

$k_{\Omega,q}$ 经过中心化后,待测样本的预测输出值如式(11)所示进行计算:

$$y_q=\varphi_q\boldsymbol{B}=k_{\Omega,train}\boldsymbol{U}((\boldsymbol{T}^{\mathrm{T}}\boldsymbol{K}_{\Omega,train}\boldsymbol{U})^{-1})\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y}_{\Omega,train}\tag{11}$$

2　数据来源及评价指标

2.1　数据来源

本文在自主搭建的录井气体红外光谱采集系统(如图 3 所示)采用研发的红外光谱气测录井仪在常温常压下采集了以氮气为背景载气的甲烷、乙烷、丙烷、正丁烷、异丁烷、正戊烷、异戊烷、二氧化碳(CO_2)及其混合物为目标气体的红外光谱数据。采集时,混合配气系统控制各通道的流量为 1 000 mL/min,输出的样品气体经可反

向吹扫气体的干燥管消除水分子的影响，再经过滤器过滤掉粉尘等颗粒物后进入有效光程长度为 4.8 m、体积为 400 mL 的光程池，光程池外部恒温装置控制内部温度恒定在 27.5 ℃，主要仪器及参数设置说明如表 1 所示。

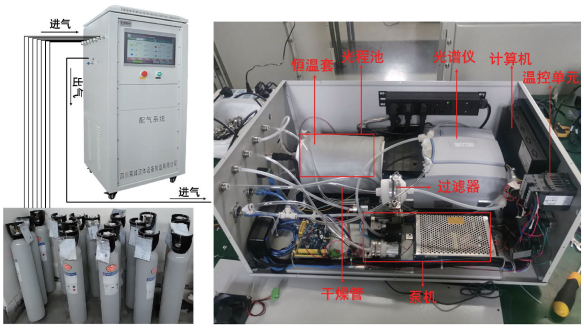


图 3　录井气体红外光谱采集系统硬件结构

Fig. 3　Hardware architecture diagram of logging gas infrared spectroscopy acquisition system

表 1　系统中各单元的型号及参数设置说明

Table 1　The parameters for each unit in the system

单元	型号	设置及说明
气源	标准气瓶（大连计量检验检测研究院）	C_1、C_2、C_3、nC_4、iC_4、nC_5、iC_5、CO_2、N_2（载气）
混合配气系统	LFIX-7000（莱峰，成都，中国）	6 通道，输出气体浓度误差为气源浓度的±1%
干燥管	MD-070-24F-4091119-02（Perma Pure，US）	长度 1 m 反向吹扫
过滤器	GWT40（纪维，沈阳，中国）	过滤颗粒尺寸≥10 μm
光程池	PMG10030（荧飒，上海，中国）	有效光程长度为 4.8 m 恒温 27.5℃
光谱仪	ALPHA Ⅱ（Bruker，Germany）	采集范围：2 000~6 500 cm^{-1} 采集间隔：1 cm^{-1}

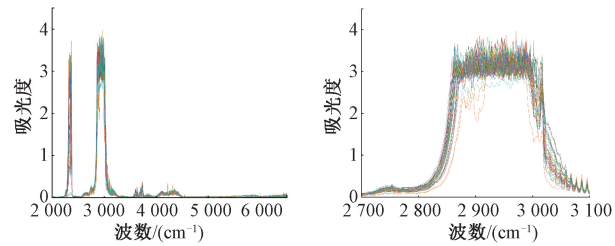
在制备混合样本过程中，第 i 个组分的目标浓度 C_i（随机设置）和其气源气瓶的标气浓度 C_i^* 应当满足式（12）的约束条件：

$$0 \leqslant \frac{C_{N_2}}{C_{N_2}^*} + \sum_{i=1}^{n} \frac{C_i}{C_i^*} = 1 \tag{12}$$

其中，$C_{N_2}/C_{N_2}^*$ 表示背景氮气填充，$C_{N_2}^*$ 为纯氮气浓度。

为了保证数据集中各样本浓度的准确性，根据气源的浓度及数目划分为 6 组分数据集和 7 组分数据集。其中，6 组分数据集由 40 组标气浓度为 1% 的甲烷、乙烷、丙烷、二氧化碳、正丁烷、正戊烷共 6 种气体混合配比组成，7 组分数据集由 359 组标气浓度为 100% 的甲烷、乙烷、丙烷、二氧化碳及标气浓度为 5% 的正丁烷、标气浓度为 10% 的异丁烷、标气浓度为 4% 的异戊烷共 7 种气体混合配比组成。

图 4 所示为 6 组分数据集原始光谱曲线，其中横轴表示波数，纵轴表示吸光度，图 4（a）和（b）分别是 6 组分中浓度数据集的真实红外光谱曲线和光谱局部放大图（2 700~3 100 cm^{-1}），可以看出不同组分的吸收峰彼此相互重叠，相互间形成较强干扰。



(a) 6组分数据集光谱
(a) Spectrum of the six-component dataset

(b) 图(a)局部放大图
(b) Partial enlarged detail of Fig.(a)

图 4　原始光谱

Fig. 4　Original spectrum

2.2　评价指标

为了评价模型的性能，本文定量分析模型采用决定系数（R^2）、均方根误差（*RMSE*）和平均相对误差（*MRE*）为评价指标。其中决定系数表示的是实际值与预测值之间的相关性，该值参考范围为（0，1），越接近于 1 意味着该模型的拟合优度越高。误差评价指标反映了实际值与预测值的差异，均方根误差数值越小代表模型预测精度越高，平均相对误差越小，测量越准确。3 个指标的具体公式如式（13）~（15）所示。

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{13}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \tag{14}$$

$$MRE = \frac{\sum_{i=1}^{N} \left| \frac{(\hat{y}_i - y_i)^2}{y_i} \right|}{N} \tag{15}$$

式中：i 为测量样本序号，N 为测量样本数量，y_i 为第 i 个测量样本实际浓度值，$\hat{y}_i$ 为第 i 个测量样本预测浓度，$\bar{y}$ 为 N 个样本真实浓度的均值。

3　实验与讨论

3.1　不同相似性度量准则下的分析效果

为了对比不同相似性度量准则筛选相似样本对建模效果的影响，本文在“7 组分数据集”上对甲烷、乙烷、丙

烷开展定量分析测试,将其余组分均视为干扰组分,采用局部加权偏最小二乘法为基准模型,对比了 6 种单一相似性度量方法及本文提出的融合相似性度量准则下的预测效果,结果如表 2 所示。

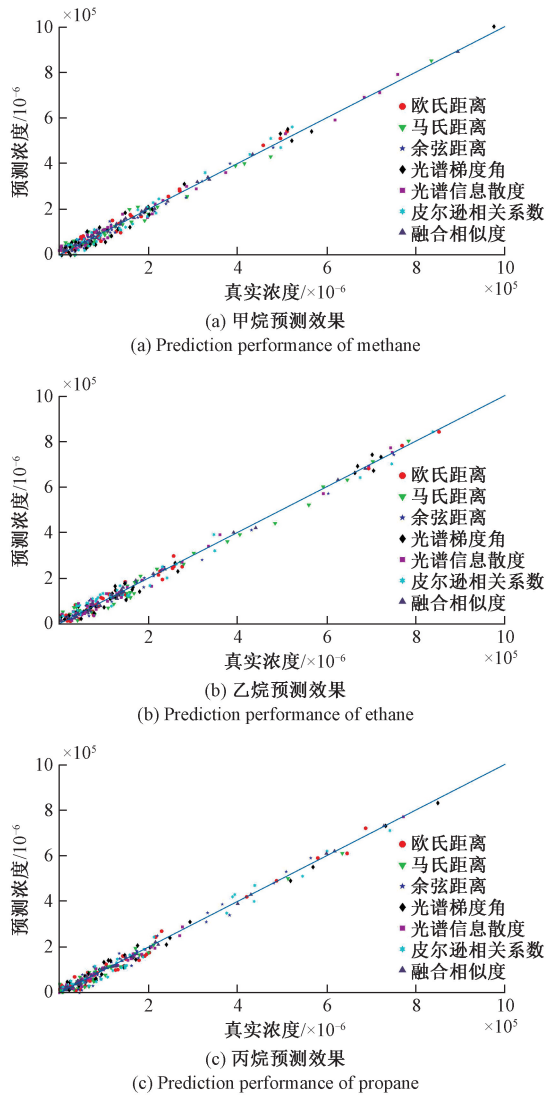
表 2　不同相似性度量准则预测结果

Table 2　The prediction results of different similarity measurement criteria

权重矩阵方式	C_1			C_2			C_3		
	MRE/%	*RMSE*/×10^{-6}	R^2	*MRE*/%	*RMSE*/×10^{-6}	R^2	*MRE*/%	*RMSE*/×10^{-6}	R^2
欧氏距离	3. 64	11 470	0. 987 5	3. 71	10 921	0. 986 9	3. 46	9 915	0. 989 1
马氏距离	3. 70	11 554	0. 987 4	4. 16	11 269	0. 985 9	3. 89	10 117	0. 987 9
余弦距离	4. 29	13 619	0. 981 5	4. 25	11 386	0. 985 6	4. 13	10 351	0. 986 2
光谱梯度角	3. 14	10 738	0. 988 8	3. 68	10 783	0. 987 6	3. 34	9 826	0. 989 7
光谱信息散度	2. 87	9 638	0. 993 9	3. 57	9 952	0. 989 6	3. 19	9 587	0. 992 4
皮尔逊相关系数	4. 01	12 743	0. 982 6	4. 97	11 092	0. 986 3	4. 34	10 730	0. 984 4
融合相似度	2. 56	8 410	0. 995 1	3. 07	9 228	0. 993 0	2. 97	9 053	0. 994 5

表 2 的实验数据表明,采取不同的相似性度量准则,其预测效果各不相同。其中,本文提出的融合相似度准则在 7 组分数据集中的分析效果最佳,其拟合优度在 3 种烷烃气体中均达到 0. 99 以上,C_1、C_2、C_3 3 种物质的 *RMSE* 值分别为 8 410×10^{-6}、9 228×10^{-6} 和 9 053×10^{-6},取得了最低的均方根误差,均低于配气系统的输出误差(标准气体浓度为 100%,配气系统输出误差为 10 000×10^{-6}),在所有相似性度量准则中表现最优,其次为光谱信息散度。可以看出,单一的相似性度量方法在相似样本筛选方面表现不稳定,烃类气体红外谱线的相互干扰影响了有效信息的提取。本文所提方法结合距离、角度、形状、信息熵等多种角度进行相似样本筛选用于即时学习建模,在烷烃多组分混合气体数据集上仍能达到较高的分析精度,这有利于在线建模及模型更新,更适用于红外光谱气测录井等复杂工况环境。

不同相似性度量准则下烷烃气体预测浓度和真实浓度结果如图 5 所示。从图 5 中样品点相对于参考线的离散程度可以明显看出,选择不同相似性度量准则建立预测模型所预测的浓度误差有不同程度的离散。本文所提基于融合相似性度量准则的局部加权方式建模预测结果基本都处于参考线附近,可以更有效提取光谱数据样本的相似性,对待测数据的浓度预测结果更接近真实值。本文构建的融合相似性度量准则可以有效提升模型的预测精度且表现稳定,是一种可靠的样本相似性度量准则。



(a) 甲烷预测效果
(a) Prediction performance of methane

(b) 乙烷预测效果
(b) Prediction performance of ethane

(c) 丙烷预测效果
(c) Prediction performance of propane

图 5　不同相似性度量准则的预测效果

Fig. 5　The prediction performance based on different similarity measurement criteria

3. 2　不同非线性核函数的影响

为了验证不同类型的非线性核函数对模型预测效果的影响,本文进一步在小样本数据集“6 组分数据集”中开展甲烷、乙烷和丙烷定量分析实验,以本文所提的融合相似性度量准则加权偏最小二乘为基础模型,引入了 5 种不同非线性核函数,比较核函数的引入对模型精度的影响,优选出最适合预测烷烃气体的核函数。

基于不同核函数的模型预测效果如表 3 所示,可以看出,并不是所有的核函数都有助于提高模型精度。在

5种核函数中，基于Morlet核函数和高斯核函数的局部建模在甲烷、乙烷、丙烷中的预测结果均明显优于无核函数时的模型，可用于烃类气体分析。相比于高斯核函数，Morlet核函数的预测结果在R^2、*RMSE*和*MRE*值上均最优，预测C_1、C_2、C_3时的*RMSE*值分别为57.17×10^{-6}、56.31×10^{-6}、71.08×10^{-6}，均远低于配气系统的输出误差（标准气体浓度为1%，配气系统输出误差为100×10^{-6}）。因Morlet核函数本身的平移不变性质，对时变非线性数据的拟合能力更强，因此优选Morlet核函数，可以有效改善偏最小二乘局部模型对非线性信息的处理能力，提高对烷烃气体的预测精度。

表3　基于不同核函数的模型预测结果

Table 3　The prediction results based on different kernel functions

核函数	C_1			C_2			C_3		
	MRE/%	*RMSE*/$\times10^{-6}$	R^2	*MRE*/%	*RMSE*/$\times10^{-6}$	R^2	*MRE*/%	*RMSE*/$\times10^{-6}$	R^2
无	7.28	92.23	0.982 6	9.13	98.83	0.981 9	9.71	89.02	0.983 1
多项式核	9.14	94.47	0.982 2	8.99	82.26	0.982 2	23.09	92.96	0.978 2
高斯核	6.75	68.972	0.991 1	6.17	71.957	0.986 2	7.82	83.952	0.989 1
Sigmoid核	13.78	99.59	0.972 9	10.37	99.57	0.979 1	27.05	96.17	0.971 7
拉普拉斯核	8.82	92.29	0.981 9	6.69	77.40	0.985 4	5.40	82.58	0.989 6
Morlet核	5.42	57.17	0.994 1	5.44	56.31	0.991 7	3.39	71.08	0.991 5

为了进一步地验证引入核函数后模型的稳定性，图6展示了6种核函数以及无核函数时100次重复性实验中决定系数R^2的变化情况。可以看出，基于不同核函数模型的R^2都不会产生较大波动。在其余组分的干扰下，基于Morlet核函数的模型在小样本数据集中对甲烷、乙烷和丙烷的预测均表现出最优性能，进一步证明了基于Morlet核函数建模的可行性。

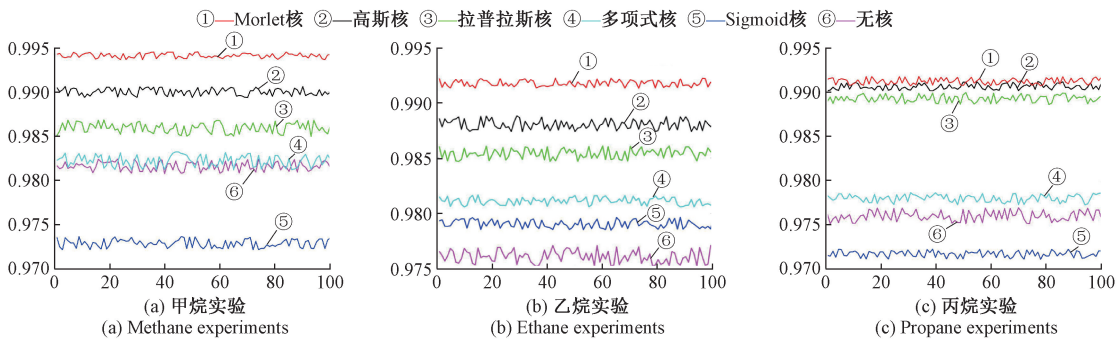


图6　不同核函数的100次重复性实验

Fig. 6　100 repetitions of experiments with different kernel functions

3.3　改进模型对比实验

为了进一步确定本文模型的性能，在“6组分数据集”上开展改进模型对比实验。将偏最小二乘(PLS)作为基础对比模型[17]；在PLS模型上结合本文提出的融合相似性度量，记为SF-PLS；在PLS模型中引入本文优选出的Morlet核函数，记为MPLS；针对相似样本以及局部模型两个方面结合改进，本文提出的融合相似性度量加权核偏最小二乘的建模策略，记为SF-KPLS。4种模型均只对甲烷、乙烷、丙烷开展定量分析测试，将其余组分均视为干扰组分，实验结果统计如表4所示。

表4　4种定量分析模型实验结果

Table 4　Experimental results of four quantitative analysis models

模型	C_1			C_2			C_3		
	MRE/%	*RMSE*/$\times10^{-6}$	R^2	*MRE*/%	*RMSE*/$\times10^{-6}$	R^2	*MRE*/%	*RMSE*/$\times10^{-6}$	R^2
PLS	9.38	98.87	0.981 6	11.48	99.89	0.980 6	23.93	91.49	0.976 0
SF-PLS	7.28	92.23	0.982 6	9.13	98.83	0.981 9	9.71	89.02	0.983 1
MPLS	6.18	64.88	0.992 1	5.87	65.88	0.989 2	5.25	82.87	0.989 7
SF-KPLS	5.42	57.17	0.994 1	5.44	56.31	0.991 7	3.39	71.08	0.991 5

从表 4 的实验结果中可以看出,相较于 PLS 模型,SF-PLS 模型、MPLS 模型性能均有提升,说明融合相似性度量与 Morlet 核函数,均有助于提高模型的分析精度。其中 Morlet 核函数引入后,模型的性能提升更加明显。将两者有效结合,则进一步提升了模型的性能。因此,在本文提出的 SF-KPLS 模型中,引入即时学习策略,通过融合相似性度量方法筛选相似样本进行局部建模,基于 Morlet 核函数尽可能地提取非线性特征,能够在复杂组分干扰的情况下取得较高的定量分析精度,对烷烃气体具有优秀的分析能力。

4　结　论

本文提出了一种融合相似性度量加权核偏最小二乘的建模策略,首先针对即时学习中的相似性度量方法改进,为解决相似性度量单一问题,从距离、角度、形状、信息熵等多种角度提出一种新的融合相似性度量方法,有效筛选出相似样本,构成局部建模数据集,再引入非线性核函数的方法并优选出 Morlet 核函数用于偏最小二乘局部模型中,增强模型的非线性数据处理能力。本文所提方法的有效性和可靠性在采集的烷烃气体混合组分数据集上进行了充分验证,结果表明本文所提融合相似性度量方法加权结合到核偏最小二乘模型能够在小样本数据集及多组分干扰的情况下,有效提高光谱定量分析精度,为红外光谱气测录井在线建模提供了有效的参考,能够促进红外光谱技术在烷烃气体定量分析检测中的应用。

参考文献

[1] 李国欣, 朱如凯. 中国石油非常规油气发展现状、挑战与关注问题[J]. 中国石油勘探, 2020, 25(2): 1-13.
LI G X, ZHU R K. Progress, challenges and key issues of unconventional oil and gas development of CNPC[J]. China Petroleum Exploration, 2020, 25(2): 1-13.

[2] 张斌. 空气钻井条件下的综合录井技术[J]. 石化技术, 2019, 26(1): 293.
ZHANG B. Comprehensive mud logging technology during air drilling [J]. Petrochemical Industry Technology, 2019, 26(1): 293.

[3] 周建立, 姚金志. 红外光谱技术在录井气体检测中的应用分析与展望[J]. 录井工程, 2016, 27(3): 36-38,97.
ZHOU J L, YAO J ZH. Infrared spectrum technique in mud logging gas detection [J]. Mud Logging Engineering, 2016, 27(3): 36-38, 97.

[4] 荆文峰, 阎荣辉, 陈中普, 等. 红外光谱录井技术在长庆油田的创新应用[J]. 录井工程, 2019, 30 (3): 124-130.
JING W F, YAN R H, CHEN ZH P, et al. Innovative application of infrared spectroscopy logging technology in Changqing Oilfield [J]. Mud Logging Engineering, 2019, 30(3): 124-130.

[5] CYBENKO G. Just-in-time learning and estimation[J]. Nato ASI Series F Computer and Systems Sciences, 1996, 153(12): 423-434.

[6] SHAO W, GE Z, SONG Z. Bayesian just-in-time learning and its application to industrial soft sensing[J]. IEEE Transactions on Industrial Informatics, 2020, 16(4): 2787-2798.

[7] KIM S, KANO M, NAKAGAWA H. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection[J]. International Journal of Pharmaceutics, 2011, 421(2): 269-274.

[8] 施锦涛, 陈磊, 秦凯, 等. 基于时空相似性的即时学习在线建模[J]. 仪器仪表学报, 2022, 43(6): 185-193.
SHI J T,CHEN L, QIN K, et al. Online modeling of just-in-time learning based on spatial-temporal similarity[J]. Chinese Journal of Scientific Instrument, 2022, 43(6): 185-193.

[9] KANO M, FUJIWARA K. Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications[J]. Journal of Chemical Engineering of Japan, 2013, 46: 1-17.

[10] KE T, LV H, SUN M, et al. A biased least squares support vector machine based on Mahalanobis distance for PU learning[J]. Physica A: Statistical Mechanics and its Applications, 2018, 509: 422-438.

[11] PAN B, JIN H, WANG L, et al. Just-in-time learning based soft sensor with variable selection and weighting optimized by evolutionary optimization for quality prediction of nonlinear processes [J]. Chemical Engineering Research and Design, 2019, 144: 285-299.

[12] YUAN X, ZHOU J, WANG Y, et al. Multi-similarity measurement driven ensemble just-in-time learning for soft sensing of industrial processes [J]. Journal of Chemometrics, 2018, 32(9): 1-14.

[13] UCHIMARU T, KANO M. Sparse sample regression based just-in-time modeling (SSR-JIT): Beyond locally weighted approach [J]. IFAC PapersonLine, 2016, 49(7): 502-507.

[14] GUO F, HUANG B. A mutual information-based variational autoencoder for robust JIT soft sensing with abnormal observations[J]. Chemometrics and Intelligent Laboratory Systems, 2020, 204:104118.

[15] XIA P, ZHANG L, LI F. Learning similarity with cosine similarity ensemble [J]. Information Sciences, 2015, 307: 39-52.

[16] CHEN M, KHARE S R, HUANG B. A unified recursive just-in-time approach with industrial near infrared spectroscopy application [J]. Chemometrics and Intelligent Laboratory Systems, 2014, 135 (1): 133-140.

[17] KIM S, KANO M, HASEBE S, et al. Long-term industrial applications of inferential control based on just-in-time Soft-Sensors: Economical impact and challenges[J]. Industrial & Engineering Chemistry Research, 2013, 52 (35): 12346-12356.

[18] 潘贝, 金怀平, 杨彪,等. 基于多样性加权相似度的集成局部加权偏最小二乘软测量建模[J]. 信息与控制, 2019, 48(2): 217-223,231.
PAN B, JIN H P, YANG B, et al. Soft sensor development based on ensemble locally weighted partial least squares using diverse weighted similarity measures[J]. Information and Control, 2019, 48(2): 217-223,231.

[19] CHEN L, ZHAO Z, LIU F. A modified recursive locally weighted NIR modeling for fermentation process [C]. IEEE International Symposium on Advanced Control of Industrial Processes, 2017: 559-564.

[20] ROSIPAL R, TREJO L J. Kernel partial least squares regression in reproducing kernel Hilbert space [J]. Journal of Machine Learning Research, 2001, 2(2): 97-123.

[21] ZHAO M, MA S, REN J. An improved ensemble adaptive kernel PLS soft sensor model and its application [C]. Proceedings of the 37th Chinese Control Conference, 2018, 8098-8103.

[22] MELLO-ROMAN J D, HERNANDEZ A, MELLO-ROMAN J C. Improved predictive ability of KPLS regression with memetic algorithms [J]. Mathematics, 2021, 9(5): 506.

[23] ALAKENT B. Soft-sensor design via task transferred just-in-time learning coupled transductive moving window learner[J]. Journal of Process Control, 2021, 101: 52-67.

[24] 陆荣秀, 饶运春, 杨辉, 等. 基于改进即时学习算法的镨/钕元素组分含量预测[J]. 控制理论与应用, 2020, 37(8): 1846-1854.
LU R X, RAO Y CH, YANG H, et al. Prediction of Pr/Nd component content based on improved just-in-time learning algorithm[J]. Control Theory and Applications, 2020, 37(8): 1846-1854.

作者简介

李忠兵(通信作者),2009 年于武汉大学获得学士学位,2014 年于武汉大学获得博士学位,现为西南石油大学讲师,主要研究方向为图像处理技术、机器学习和光谱分析。
E-mail:lzb@ swpu. edu. cn

Li Zhongbing (Corresponding author) received his B. Sc. degree in 2009 from Wuhan University, Wuhan, China, received his Ph. D. degree in 2014 from Wuhan University, Wuhan, China. Now he is a lecturer in Southwest Petroleum University. His main research interests include image processing technology, machine learning and spectrum analysis.