文章编号:2095-6134(2020)06-0728-08

基于非参数模型的气体浓度的逆向预测*

吴栋,郭潇†

(中国科学技术大学管理学院国际金融研究院, 合肥 230026)

(2019 年 3 月 18 日收稿; 2019 年 5 月 20 日收修改稿)

Wu D, Guo X. Inverse prediction of gas concentration based on nonparametric model[J]. Journal of University of Chinese Academy of Sciences, 2020,37(6):728-735.

摘 要 气体传感器阵列是一种重要且强大的检测气体和测量浓度的技术。传统的描述传感器响应与气体浓度之间关系的策略是使用一些特定的非线性参数模型。本文使用非参数模型描述传感器响应随气体浓度的变化,有效避免了模型的错误假定。提出一种基于非参数模型逆向预测气体浓度的方法。还提出通过数据驱动选择可调参数的方法。数值模拟结果表明,当传感器阵列的实际模型未知或模型假定错误时,非线性参数模型的性能劣于非参数模型,实际数据分析也验证了这一点。

关键词 高斯牛顿法; 气体浓度; 逆向预测; 非线性参数模型; 非参数模型

中图分类号:O212.7 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2020.06.002

Inverse prediction of gas concentration based on nonparametric model

WU Dong, GUO Xiao

(*International Institute of Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract Gas sensor array is an important and powerful technique for detecting gas and measuring gas concentrations. The conventional strategy to describe the relationship between the response of the sensor and the actual gas concentration is to use some specific nonlinear parametric models. In this work, we use the nonparametric model to depict the change in the gas sensor response with the gas concentrations, which effectively avoids model misspecification. Furthermore, we propose an inverse prediction method based on the nonparametric model to predict gas concentrations. Data-driven selection of tuning parameters is also developed. The simulation results reveal that, when the real model of the sensor array is unknown or misspecified, the nonlinear parametric model is inferior to the nonparametric model in performance. Meanwhile, we verify this with the real data analysis.

Keywords Gauss Newton method; gas concentration; inverse prediction; nonlinear parametric model; nonparametric model

随着科学技术的进步和工业的发展,有毒有害气体的使用范围在不断扩大。这些气体既可能是生产之初需要的原材料,比如大多数的有机化学物质,也可能是生产过程中各个环节产生的副

* 中央高校基本科研业务费专项资金和国家自然科学基金(11601500,11671374,11771418)资助

† 通信作者,E-mail:xiaoguo@ustc.edu.cn

产品,比如一氧化碳、氨、硫化氢等等。有害气体影响人类的身体健康,所以有害气体的检测在我们的生产和生活中十分重要。

在检测气体的过程中用到的最关键的元器件就是气体传感器,气体传感器是一种能够感知外部环境中某种气体及其浓度变化的气体敏感元器件。针对不同检测任务,例如气体浓度预测(回归)[1-2],气体识别(分类)[3],类似气体的分组(聚类)[4],实验人员会使用气体传感器阵列设计不同的实验[5]。本文主要关注气体浓度预测(回归),实验中获得的气体传感器阵列的多变量响应都携带着有关气体浓度变化的信息[6],可以通过传感器的响应分析混合气体的种类和浓度。

为预测气体浓度,首先有必要建立一个联系气体传感器响应与真实气体浓度的校准模型。传统的方法是建立一些特定的参数模型,包括线性的或者非线性的[7]。比如,文献[2,5]使用线性模型对气体传感器数据进行定量的多元分析;文献[8-9]根据实验经验提出使用幂函数模型来刻画气体传感器的相对导电率与气体浓度变化之间的关系;文献[10]经过理论上的考虑提出另一种形式的参数模型;文献[11]则提出使用对数模型建立传感器与气体浓度之间的联系。尽管这些文献提出的模型都是联系气体传感器响应与气体浓度之间关系的,但是它们彼此之间还是存在着一定的差异。在实际应用中,真实模型的形式会因为气体种类和传感器类型的不同而不同。因此,如果选错了对应的模型,可能会导致最后无法得到较好的预测结果。

为了避免模型设定的错误,本文采用非参数模型建立每个传感器与气体浓度之间的联系。采用局部线性核回归的方法对非参数模型进行拟合。对于每个非参数模型中的窗宽参数,采用交叉验证的方法进行选取。在得到非参数模型拟合结果的同时,还可以得到非参数模型函数关于气体浓度这个协变量一阶偏导数的估计。这个一阶偏导数的估计在后面对气体浓度进行逆向预测的时候会起到很大的作用。

这里使用非参数模型建模,主要是考虑到非参数模型本质上包括线性模型与非线性模型这些特殊情况。采用非参数模型进行建模也降低了由于模型选择错误带来的气体浓度预测误差较大的风险,具有弱假设条件的非参数模型在面对数据污染的情况时会表现得更加稳健[12]。

紧接着,需要对气体的浓度进行预测。这里面临的预测问题与传统的预测问题不同,传统的预测问题往往是在已知协变量的情况下预测响应变量的值。但在实际情况中,只能观察到气体传感器的响应,也即回归模型中的因变量。然后利用前面拟合好的模型,在已知气体传感器响应的情况下,逆向预测气体的浓度。为了得到更精确的预测结果,在实验过程中,往往会使用比混合气体种类更多的气体传感器,这样会得到比气体浓度更多的气体传感器数据。通过代入已经拟合好的非参数模型,建立多个新的气体传感器响应与未知浓度气体之间的联系,最小化它们之间的误差,得到未知气体浓度的预测值[13]。在后面的数值模拟分析和实际数据分析的过程中,我们发现非参数模型的表现都好于特定的参数模型。

1　模型与求解

在使用传感器进行实验的过程中,可以观察到 $\{y_{1t},\cdots,y_{mt};\ x_{1t},\cdots,x_{kt}\}_{t=1}^{n}$,其中 $\{y_{1t},\cdots,y_{mt}\}_{t=1}^{n}$ 表示 m 个传感器的响应读数,$\{x_{1t},\cdots,x_{kt}\}_{t=1}^{n}$ 表示 k 种气体的真实浓度。但是,在真实环境中,得到的仅仅是 m 个传感器的响应读数。因此,我们的目标是通过已知的 m 个传感器的响应读数逆向预测 k 种气体的真实浓度。

为避免模型假定的错误,首先使用非参数模型对气体传感器响应与真实的气体浓度之间的关系进行建模,模型如下所示:

$$y_{it}=f_i(\boldsymbol{X}_t)+\epsilon_{it},t=1,\cdots,n,i=1,\cdots,m, \tag{1}$$

式中:y_{it} 表示第 i 个传感器的第 t 次观察值,$\boldsymbol{X}_t=(x_{1t},\cdots,x_{kt})^{\mathrm{T}}$ 表示第 t 次观察时对应的 k 种气体浓度,ϵ_{it} 表示随机误差。这里假定误差是独立同分布的,且均值为 0。同时假定 $x_{jt},j=1,\cdots,k$ 这些协变量有紧凑的支撑 $\Omega=[0,1]$。在实际应用中,如果 $x_{jt}\notin[0,1]$,需要对所有的协变量进行归一标准化,如下所示

$$\frac{x_{jt}-\min_{t=1,\cdots,n}(x_{jt})}{\max_{t=1,\cdots,n}(x_{jt})-\min_{t=1,\cdots,n}(x_{jt})}, \tag{2}$$

式中:$\min\limits_{t=1,\cdots,n}(x_{jt})$ 和 $\max\limits_{t=1,\cdots,n}(x_{jt})$ 分别表示 n 次观察中第 j 种气体的最小和最大浓度。接下来需要通过局部线性核回归的估计方法来拟合 $f_i(\cdot)$[14-15]。对于未知的气体浓度 $\boldsymbol{x}=(x_1,x_2,\cdots,x_k)^{\mathrm{T}}$,可以同时得到 $f_i(\boldsymbol{x})$ 的估计和 $f_i(\boldsymbol{x})$ 关于气体浓度 $\boldsymbol{x}$ 一阶偏导 $\beta_i(\boldsymbol{x})$ 的估计。其中一阶偏导 $\beta_i(\boldsymbol{x})$ 的估计

对后面进行气体浓度的逆向预测会起到很大的作用。模型 $f_i(\cdot)$ 及其偏导的估计如下：

$$(\hat{f}_i(\boldsymbol{x}),\hat{\boldsymbol{\beta}}_i(\boldsymbol{x})^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_{x_i}\boldsymbol{X}_x)^{-1}(\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_{x_i}\boldsymbol{Y}_i), \quad (3)$$

式中：$\boldsymbol{Y}_i = (y_{i1},\cdots,y_{in})^{\mathrm{T}}$，$\boldsymbol{W}_{x_i} = \mathrm{diag}\{K_{h_i}(\boldsymbol{X}_1 - \boldsymbol{x}),\cdots,K_{h_i}(\boldsymbol{X}_n - \boldsymbol{x})\}$，

$$\boldsymbol{X}_x = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^{\mathrm{T}} \end{pmatrix},$$

$K_{h_i}(\boldsymbol{x}) = (h_{i_1}h_{i_2}\cdots h_{i_k})^{-1}\prod_{j=1}^{k} g(x_j \cdot h_{i_j}^{-1})$，$g(\cdot)$ 是核函数，$\boldsymbol{h}_i = (h_{i_1},h_{i_2},\cdots,h_{i_k})^{\mathrm{T}}$ 为窗宽参数。

通过上面的处理，对每个气体传感器与气体浓度之间的非参数模型进行了拟合。因为我们的目的是预测未知的气体浓度，所以需要根据拟合好的非参数模型，对 k 种气体浓度进行逆向预测。

假设观察到 m 个新的传感器的响应为 $\boldsymbol{Y}_{\mathrm{new}} = (y_{\mathrm{new}_1},\cdots,y_{\mathrm{new}_m})^{\mathrm{T}}$，需要预测未知的 k 种气体浓度记为 $\boldsymbol{X}_{\mathrm{new}} = (x_{\mathrm{new}_1},\cdots,x_{\mathrm{new}_k})^{\mathrm{T}}$。根据前面提出的非参数模型(1)，可以得到如下等式

$$\begin{aligned} y_{\mathrm{new}_1} &= f_1(\boldsymbol{X}_{\mathrm{new}}) + \epsilon_1, \\ &\vdots \\ y_{\mathrm{new}_m} &= f_m(\boldsymbol{X}_{\mathrm{new}}) + \epsilon_m. \end{aligned}$$

其中 $\boldsymbol{\epsilon} = (\epsilon_1,\cdots,\epsilon_m)^{\mathrm{T}}$ 表示随机误差。

接下来，采用最小二乘法得到未知浓度的气体的逆向预测值。通过代入拟合好的 m 个传感器对应的非参数模型，需要最小化如下残差平方和

$$S(\boldsymbol{X}_{\mathrm{new}}) = \sum_{i=1}^{m} (y_{\mathrm{new}_i} - \hat{f}_i(\boldsymbol{X}_{\mathrm{new}}))^2. \quad (4)$$

将式(4)的极小值当做是 $\boldsymbol{X}_{\mathrm{new}}$ 的逆向预测值。这里采用高斯牛顿法来求极小值[16]。前面介绍的非参数模型 $f_i(\cdot)$ 关于气体浓度的偏导数估计在这里发挥了关键的作用，下面介绍如何应用高斯牛顿法解决非参数模型下气体浓度的逆向预测问题。

首先需要给未知的气体浓度设定一个初始值，记为 $\boldsymbol{X}_0$，然后将非参数模型 $\hat{f}_i(\cdot)$ 在 $\boldsymbol{X}_0$ 这点做一阶泰勒展开，那么残差平方和 $S(\boldsymbol{X}_{\mathrm{new}})$ 的表达式约等于如下形式

$$\sum_{i=1}^{m} [y_{\mathrm{new}_i} - \{\hat{f}_i(\boldsymbol{X}_0) + (\boldsymbol{X}_{\mathrm{new}} - \boldsymbol{X}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i(\boldsymbol{X}_0)\}]^2. \quad (5)$$

对式(5)进行简单的改写，再使用普通最小二乘法可以得到 $\boldsymbol{X}_{\mathrm{new}}$ 的首次更新值，记为 $\boldsymbol{X}_0^{(1)}$，具体表达式如下

$$\boldsymbol{X}_0^{(1)} = \boldsymbol{X}_0 + (\hat{\boldsymbol{\beta}}(X_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{X}_0))^{-1} \{\hat{\boldsymbol{\beta}}(\boldsymbol{X}_0)^{\mathrm{T}}(\boldsymbol{Y}_{\mathrm{new}} - \hat{\boldsymbol{f}}(\boldsymbol{X}_0))\}. \quad (6)$$

其中 $\hat{\boldsymbol{f}}(\boldsymbol{X}_0) = (\hat{f}_1(\boldsymbol{X}_0),\cdots,\hat{f}_m(\boldsymbol{X}_0))^{\mathrm{T}}$，$\hat{\boldsymbol{\beta}}(\boldsymbol{X}_0) = (\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}(\boldsymbol{X}_0),\cdots,\hat{\boldsymbol{\beta}}_m^{\mathrm{T}}(\boldsymbol{X}_0))^{\mathrm{T}}$.

然后判断 $\| \boldsymbol{X}_0^{(1)} - \boldsymbol{X}_0 \| < \varepsilon$ 是否成立，其中 ε 是一个很小的数，本文令 $\varepsilon = 10^{-3}$。如果不成立，需要把第一次更新的值 $\boldsymbol{X}_0^{(1)}$ 赋给 $\boldsymbol{X}_0$，然后通过迭代式(6)得到 $\boldsymbol{X}_0^{(2)}$。不断重复上面的步骤，直到 $\| \boldsymbol{X}_0^{(q)} - \boldsymbol{X}_0^{(q-1)} \| < \varepsilon$，就得到最终的估计，其中 $\boldsymbol{X}_0^{(q)}$ 表示第 q 次迭代的结果。最终将 $\boldsymbol{X}_0^{(q)}$ 当做浓度未知气体 $\boldsymbol{X}_{\mathrm{new}}$ 的预测值，记为 $\hat{\boldsymbol{X}}_{\mathrm{new}}$。

如果前面对协变量进行了归一化处理，则还需要进行逆归一化的操作才能得到气体的逆向预测值 $\hat{\boldsymbol{X}}_{\mathrm{new}}$，具体步骤如下

$$\hat{x}_{\mathrm{new}_j} = x_{0_j}^{(q)}(\max_{t=1,\cdots,n}(x_{jt}) - \min_{t=1,\cdots,n}(x_{jt})) + \min_{t=1,\cdots,n}(x_{jt}), \quad (7)$$

其中 $\hat{x}_{\mathrm{new}_j}$ 是 $\hat{\boldsymbol{X}}_{\mathrm{new}}$ 的第 j 项值，$x_{0_j}^{(q)}$ 是 $\boldsymbol{X}_0^{(q)}$ 的第 j 项值。

2 窗宽选择

在前面介绍的基于非参数模型对气体浓度进行逆向预测的方法中，首先是对每个传感器对应的非参数模型进行拟合，当时采用的拟合方法是局部线性核回归。采用核方法比较关键的一步是确定核函数的窗宽。为了使得下一步能够更好地对气体浓度进行预测，需要根据已有的数据提高非参数模型的拟合精度，因此需要尽量减小第一步所产生的拟合误差。下面介绍本文采用的窗宽选择方法。

关于非参数模型窗宽选择的方法，很多文献有过讨论。比如，文献[17]证明在具有单个连续协变量的回归模型中，在使用局部多项式方法拟合时，通过交叉验证选择得到的带宽具有渐近最优性。文献[18-19]指出利用数据驱动选择窗宽的方法在非参数模型设定的情形下应用十分广泛，并证明在具有多个协变量的非参数模型中，通过交叉验证选择得到的窗宽在代入模型之后的渐进正态性。以上研究都表明在非参数模型设置下

使用交叉验证进行窗宽选择具有良好的表现,因此这里也采用交叉验证的方法进行窗宽选择。下面需要最小化交叉验证(CV)得分,表达式为

$$CV(h_{i_1},\cdots,h_{i_k})=\frac{1}{n}\sum_{t=1}^{n}(y_{it}-\hat{f}_i^{(-t)}(\boldsymbol{X}_t))^2,\quad(8)$$

式中 $\hat{f}_i^{(-t)}(\boldsymbol{X}_t)$ 表示去除第 t 次观察 $(\boldsymbol{X}_t^{\mathrm{T}},y_{it})$ 之后得到的 $f_i(\boldsymbol{X}_t)$ 的留一估计。

在实际应用中,如果直接最小化式(8)来选择窗宽,需要进行 n 次非参数模型的拟合,计算成本很大。文献[20]指出 $CV(h_{i_1},\cdots,h_{i_k})$ 可以进一步简化为

$$CV(h_{i_1},\cdots,h_{i_k})=\frac{1}{n}\sum_{t=1}^{n}\frac{(y_{it}-\hat{f}_i(\boldsymbol{X}_t))^2}{(1-\boldsymbol{S}_{h_i}(t,t))^2},\quad(9)$$

其中 $\boldsymbol{S}_{h_i}$ 表示局部线性光滑矩阵,定义如下

$$\boldsymbol{S}_{h_i}=\begin{pmatrix}\boldsymbol{e}_1^{\mathrm{T}}\cdot\boldsymbol{S}_{d_i}(\boldsymbol{X}_1)\\ \vdots\\ \boldsymbol{e}_1^{\mathrm{T}}\cdot\boldsymbol{S}_{d_i}(\boldsymbol{X}_n)\end{pmatrix},$$

式中:$\boldsymbol{S}_{d_i}(\boldsymbol{x})=(\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_{x_i}\boldsymbol{X}_x)^{-1}(\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_{x_i})$,$\boldsymbol{e}_i=(0,\cdots,1,\cdots,0)^{\mathrm{T}}$ 是一个长度为 $(k+1)$ 的单位列向量,其中除第 i 个元素为 1 其余都为 0。

显而易见,最小化式(9)比最小化式(8)节省很大的计算量。因此我们将通过最小化式(9)来选择 $\hat{f}_i(\boldsymbol{x})$ 的窗宽。

3　数值模拟

本节将通过数值模拟检验基于非参数模型的气体浓度逆向预测方法的效果。首先介绍一下以前的学者提出的一些描述气体传感器响应与气体浓度之间关系的校准模型。Clifford 和 Tuma[8-9] 根据实验观察,提出将传感器相对电导率与气体浓度联系起来的经验公式。该模型可以通过幂函数来描述,表达式如下

$$y_{\mathrm{pow}}(x)=(1+b\cdot x)^{\beta},\quad(10)$$

式中:$y_{\mathrm{pow}}(x)=G_0/G_g(x)$ 表示传感器的相对导电率,G_0 表示传感器在空气中的基础导电率,$G_g(x)$ 表示传感器在气体浓度为 x 下的导电率,b 和 β 是未知的参数。

除此之外,Chaiyboun 等[11] 提出拟合传感器响应的对数模型,表达式如下

$$y_{\log}(x)=a-b\cdot\ln(x+0.5),\quad(11)$$

式中:a 和 b 为未知的模型参数,x 表示气体浓度。

与此同时,Chaiyboun 等[11] 将式(10)和式(11)都推广到两种混合气体的情形,表达式如下

$$y_{\mathrm{pow}}(x_1,x_2)=\left(1+b_1\cdot\left[x_1+\frac{\{(1+b_2\cdot x_2)^{\beta_2}\}^{1/\beta_1}}{b_1}\right]\right)^{\beta_1},\quad(12)$$

式中:b_1,β_1,b_2,β_2 为未知的模型参数;x_1,x_2 分别表示两种混合气体各自的气体浓度。

$$y_{\log}(x_1,x_2)=a_0-b_0\cdot\ln\left(x_1+\exp\left[\frac{a_0-\{a_1-b_1\cdot\ln(x_2+0.5)\}}{b_0}\right]\right),\quad(13)$$

式中:a_0,b_0,a_1,b_1 为对数模型的参数;x_1,x_2 分别表示两种混合气体各自的气体浓度。

接下来,使用 Clifford 和 Tuma[8-9] 提出的幂函数模型生成模拟数据。然后分别使用对数模型、幂函数模型和非参数模型构建校准模型。通过给定新的传感器阵列响应,使用这些拟合好的模型对气体浓度进行逆向预测,并比较它们的表现。下面分别考虑单一气体和两种混合气体的情况。

3.1　单一气体

这里假设有单一气体和 8 个半导体气体传感器,气体浓度由范围为[0,500]的均匀分布产生。不同传感器的幂函数模型(10)的预定义系数如表 1 所示。每次模拟生成的样本数为 1 000。

表 1　8 个传感器对应的幂函数模型(10)的预定义系数

Table 1　Predefined coefficients in power function model (10) for 8 sensors

sensor	1	2	3	4	5	6	7	8
b	0.4	0.7	0.4	0.8	0.7	0.7	0.5	0.4
β	0.8	0.6	0.6	0.7	0.5	0.7	0.8	0.8

根据式(10)产生对应的模拟数据,具体地,对于气体浓度 x_j,第 k 个传感器的响应可以根据下面这个表达式产生:

$$y_{kj}=y_{\mathrm{pow}}(x_j)+\sigma\epsilon_{kj},\quad(14)$$

式中:ϵ_{kj} 为独立同分布的标准正态随机变量,$k=1,\cdots,8$,$j=1,\cdots,1\,000$。误差的方差 σ^2 根据信躁比的大小来确定,信噪比(SNR)的定义为 $y_{\mathrm{pow}}(x_1),\cdots,y_{\mathrm{pow}}(x_{1000})$ 的样本方差与误差 $\epsilon_1,\cdots,\epsilon_{1000}$ 的样本方差的比值。我们分别考虑 SNR=4 和 SNR=8 这两种情况。

根据 6∶4 的比例将生成的数据分成两个部分,其中 60% 的数据进行训练,40% 用于测试。接下来,通过对训练数据集进行交叉验证来确定不同传感器对应的非参数模型(1)的窗宽参数。然

后使用非参数模型对单一气体的浓度进行逆向预测。本文采用高斯核作为加权函数。同样,首先使用训练数据拟合对数模型(11)和幂函数模型(10)。紧接着根据拟合好的 8 个传感器对应的对数模型和幂函数模型,通过非线性最小二乘的方法,对气体浓度进行逆向预测。

分别在训练集和测试集上使用 3 种不同的模型对气体浓度进行逆向预测。为评估最终预测结果的好坏,通过计算预测气体浓度与真实气体浓度之间的根均方误差(RMSE)来衡量,表达式为 $\mathrm{RMSE}=\sqrt{(1/n)\sum_{i=1}^{n}(x_i-\hat{x}_i)^2}$,其中 x_i 是气体浓度的真实值,$\hat{x}_i$ 为气体浓度的预测值。

最后,将上述模拟过程重复 200 次,并计算 200 次模拟结果的 RMSE 的均值和标准差(SD)。同时使用对数模型(11)、幂函数模型(10)和非参数模型(1)对气体浓度进行逆向预测,结果汇总在表 2 中。

表 2 使用不同校准模型对单一气体浓度进行逆向预测的结果对比

Table 2 Comparison of inverse prediction results of concentration for single gas among different calibration models

SNR	模型	训练集	测试集
		gas_1(SD)	gas_1(SD)
4	对数	184.78(24.96)	185.55(29.75)
	幂函数	32.47(1.10)	32.68(1.46)
	非参	31.58(1.16)	32.03(1.31)
8	对数	169.98(21.22)	170.19(25.09)
	幂函数	23.05(0.84)	23.06(0.99)
	非参	22.28(1.00)	22.63(1.06)

从表 2 可以看出,使用对数模型对气体浓度进行逆向预测的效果不如幂函数模型和非参数模型。同时可发现使用非参数模型对气体浓度进行逆向预测能够取得与正确模型幂函数模型相同的效果,且使用这两种模型得到的 RMSE 的标准差相比对数模型也是比较小的。

3.2 混合气体

这一部分,考虑存在两种混合气体和 8 个半导体气体传感器的情况。每种气体浓度同样由范围为[0,500]的均匀分布产生。表 3 列出 8 个不同传感器对应的幂函数模型(12)的预定义系数。每次模拟生成的样本数为 1 000。

因此采用适合两种气体状况的模型(12)产生数据,对于给定的气体浓度 x_{1j} 和 x_{2j},第 k 个传感器的响应由如下表达式产生:

$$y_{kj}=y_{\mathrm{pow}}(x_{1j},x_{2j})+\sigma\epsilon_{kj}, \quad (15)$$

式中:ϵ_{kj} 是服从独立同分布的标准正态的随机变量,$k=1,\cdots,8$,$j=1,\cdots,1\,000$。噪音的方差σ^2 同样由 SNR 来确定。这里也考虑 SNR=4 和 SNR=8 两种情况。

表 3 8 个传感器对应的幂函数模型(12)的预定义系数

Table 3 Predefined coefficients of power function model (12) for 8 sensors

sensor	1	2	3	4	5	6	7	8
b_1	0.4	0.7	0.4	0.8	0.7	0.7	0.5	0.4
b_2	0.3	0.5	0.5	0.6	0.5	0.3	0.4	0.7
β_1	0.8	0.6	0.6	0.7	0.5	0.7	0.8	0.8
β_2	0.9	0.7	0.9	0.5	0.8	0.2	0.6	0.6

同样地,将数据分成 60% 的训练集和 40% 的测试集。运用对数模型(13)、幂函数模型(12)和非参数模型(1)在训练集和测试集上分别对气体浓度进行逆向预测,然后通过计算气体的预测浓度与真实浓度之间的 RMSE 比较这 3 种模型的表现。整个模拟过程重复 200 次。接下来对比对数模型(13)、幂函数模型(12)和非参数模型(1)在两个数据集上的表现,结果如表 4 所示。

表 4 使用不同校准模型对混合气体浓度进行逆向预测的结果对比

Table 4 Comparison of inverse prediction results of concentrations for mixed gases among different calibration models

SNR	模型	训练集		测试集	
		gas_1(SD)	gas_2(SD)	gas_1(SD)	gas_2(SD)
4	对数	50.92(8.23)	80.52(17.87)	51.56(8.36)	81.88(18.19)
	幂函数	43.11(1.46)	51.46(2.07)	43.29(1.74)	51.75(2.02)
	非参	42.06(1.49)	51.29(2.20)	42.61(1.76)	51.79(2.04)
8	对数	36.99(4.59)	61.17(10.97)	36.97(4.44)	60.85(10.28)
	幂函数	30.43(1.08)	36.53(1.31)	30.72(1.17)	36.90(1.45)
	非参	29.71(1.06)	36.18(1.34)	30.37(1.16)	36.81(1.50)

正如从表 4 中看到的,对于两种气体的情况,采用非参数模型对气体浓度进行逆向预测的结果也好于对数模型。同时,非参数模型同样也达到了与使用正确模型幂函数模型一样的效果。其中幂函数模型和非参数模型对应的 RMSE 的标准差都是比较小的,因此可以认为非参数模型体现了更加稳定的逆向预测能力。

4 实际数据分析

这一节主要是通过对实际的气体传感器数据

进行分析,对比对数模型、幂函数模型和非参数模型对气体浓度逆向预测的表现。本文使用 Fonollosa 等[21]在 UCI 机器学习网站上分享的气体传感器数据。这个数据集记录了浓度变化的混合气体暴露在 16 个气体传感器的环境下,气体浓度与传感器响应的时间序列数据。其中作者分别对两种混合气体的组合进行了实验,一种为乙烯与甲烷在空气中的组合,另一种为乙烯和一氧化碳在空气中的组合。每组实验持续 12 h,不断改变混合气体的浓度,分别记录 16 个气体传感器的读数。这 16 个传感器包含 4 种类型:TGS-2602,TGS-2600,TGS-2610,TGS-2620,其中每种传感器有 4 个。实验中气体的浓度每间隔 80~120 s 会随机改变一次,其中乙烯的浓度变化范围为(0~20)×10^{-6},一氧化碳浓度的变化范围为(0~600)×10^{-6},甲烷浓度的变化范围为(0~300)×10^{-6}。

首先,分析乙烯和一氧化碳这个气体组合。为了更好地了解气体传感器数据,随机截取一段时间的传感器数据,按照气体传感器的种类分组展示传感器读数的变化,如图 1 所示。

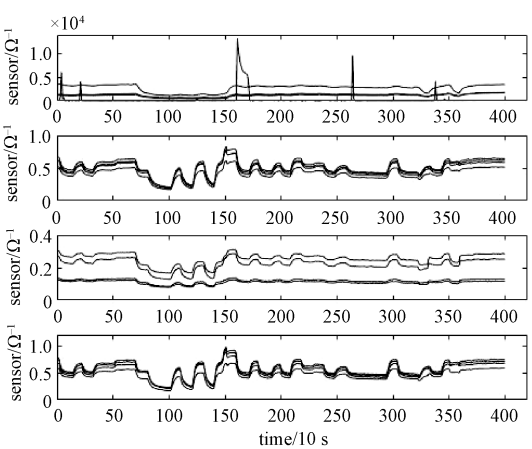

图 1　传感器阵列的读数

Fig. 1　Readings on the sensor array

图 1 依次展示 TGS-2602,TGS-2600,TGS-2610,TGS-2620 这 4 种类型的传感器的读数。发现 TGS-2602 传感器在某些时刻会产生突变,无法对气体浓度变化做出规律的反应。因此,在后面的分析步骤中,会去掉这个传感器的数据。同时,Fonollosa 等[21]在网站上提供的气体浓度数据是设定时刻的浓度值,而我们在模型中考虑的是传感器响应与气体真实浓度之间的关系。因此需要考虑设定时刻的气体浓度到达传感器过程中的时间延迟。根据 Fonollosa 等[21]的数据处理代码,得到气体到达传感器的不同时间,一氧化碳和甲烷为 17.82 s,乙烯为 26.73 s。为了将真实的气体浓度与传感器的读数进行匹配,需要根据前面计算出来的延迟时间对气体浓度数据进行平移。例如:在 t 秒时的一氧化碳的实际浓度需用(t − 17.82) s 处的设定浓度值代替,乙烯在 t 秒时的实际浓度需用(t − 26.73) s 处的设定浓度值代替。

为对比气体浓度平移变换前后的不同,分别画出未经平移变换和经过平移变换后的乙烯和一氧化碳的浓度与气体传感器阵列的读数变化图,如图 2 和图 3 所示。图 2 为未经平移处理的情况,第 1 个面板图展示 15 个传感器的读数变化,垂直虚线标示对应的浓度变化时刻,第 2 和第 3 个面板图展示未经平移变换的乙烯和一氧化碳浓度的变化。图 3 为经过平移处理的情况,第 1 个面板图展示 15 个传感器的读数变化,垂直虚线标示对应的浓度变化时刻,第 2 和第 3 个面板图展示经过平移变换的乙烯和一氧化碳浓度的变化。

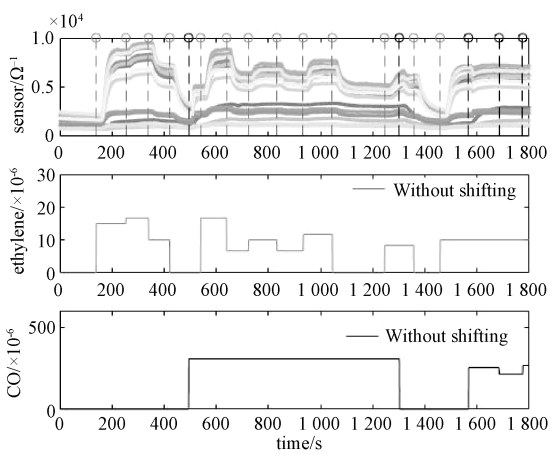

图 2　传感器阵列的读数与未经平移变换的乙烯和一氧化碳浓度

Fig. 2　Readings on sensor array and concentrations of ethylene and CO without shifting

通过图 2 和图 3 的对比,发现经过平移处理后传感器读数的变化与气体浓度的变化更为一致,因此在分析数据之前需要对气体的浓度数据进行平移。同时,从图 3 可以看出,在气体浓度转换的阶段,传感器的读数都会发生突变,但这段突变并不能显示传感器与对应气体浓度的关系。因此后面的分析过程中,去掉了这些突变的部分,处理之后的数据如图 4 所示。

采用对数模型(13)、幂函数模型(12)和非参数模型(1)对气体浓度进行逆向预测,其中对数模型和幂函数模型的响应变量是传感器的相对

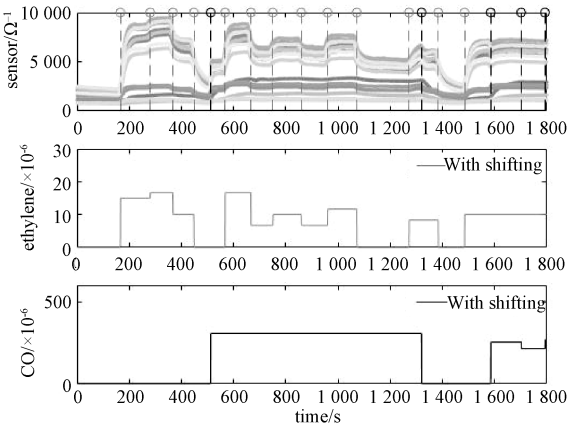

图 3 传感器阵列的读数与经过平移变换的乙烯和一氧化碳浓度

Fig. 3 Readings on sensor array and concentrations of ethylene and CO with shifting

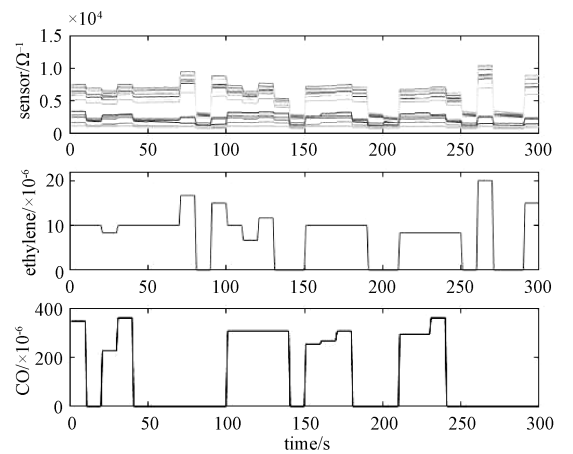

图 4 截取后稳定阶段的传感器阵列读数与对应的气体浓度

Fig. 4 Readings on sensor array in the stable phase after interception and the corresponding gas concentrations

导电率。而 Fonollosa 等[21]给出的传感器读数为其导电率,需要对原始数据进行处理得到相对导电率。而我们提出的非参数模型对响应变量的具体含义并没限制。

在使用非参数模型 (1) 时分别考虑两种不同的情况:第 1 种情况考虑模型(1)的响应变量 y 为相对导电率;第 2 种情况考虑模型(1)的响应变量 y 为作者给定的原始读数。然后比较使用对数模型 (11)、幂函数模型(10)和非参数模型(1) 进行气体浓度逆向预测的表现,结果如表 5 所示。括号中的 RC 表示模型因变量使用的是传感器的相对导电率,OR 表示模型因变量使用的是传感器的原始读数。

表 5 使用不同校准模型对乙烯和一氧化碳浓度逆向预测的 RMSE

Table 5 RMSE of inverse prediction for ethylene and CO concentrations among different calibration models

$\times 10^{-6}$

模型	训练集		测试集	
	乙烯	一氧化碳	乙烯	一氧化碳
对数(RC)	16.23	97.52	14.02	150.02
幂函数(RC)	6.82	173.64	6.23	161.07
非参(RC)	5.63	105.12	3.94	117.25
非参(OR)	1.56	53.68	2.09	79.09

根据同样的步骤分析甲烷和乙烯这个气体组合的数据,结果如表 6 所示。

表 6 使用不同校准模型对乙烯和甲烷浓度逆向预测的 RMSE

Table 6 RMSE of inverse prediction for ethylene and methane concentrations among different calibration models

$\times 10^{-6}$

模型	训练集		测试集	
	乙烯	甲烷	乙烯	甲烷
对数(RC)	12.43	24.98	11.01	27.34
幂函数(RC)	6.03	87.20	5.47	64.34
非参(RC)	2.93	55.93	3.21	70.61
非参(OR)	0.83	17.33	0.85	21.54

从表 5 和表 6 的结果可以看出非参数模型的表现优于对数模型和幂函数模型。进一步发现直接使用传感器的原始读数建立模型,得到的 RMSE 更小。通过实际分析结果可以看出使用非参数模型对气体浓度进行逆向预测具有更高的灵活性,同时也可以得到比非线性模型更好的预测结果。

5 结论与展望

本文提出一种基于非参数模型的气体浓度的逆向预测方法,通过数值模拟和实际数据分析两个维度对比非参数模型、对数模型和幂函数模型的表现。从数值模拟的结果可以看出,当模型选择错误时,对数模型的结果不如非参数模型和正确的幂函数模型,同时在使用非参数模型对气体浓度进行逆向预测时也取得了与幂函数模型一样好的结果。实际分析结果也表明非参数模型对气体浓度进行逆向预测的效果好于对数模型和幂函数模型。

在构建传感器响应与气体浓度之间关系的过程中,一些学者提出的线性模型和非线性模型过

于具体。在实际应用中,如果假定的模型与实际的模型有偏差,气体的预测结果将会受到影响。因此,我们建议采用非参数模型构建校准模型。

在对气体浓度进行逆向预测的步骤中,需要采用高斯牛顿法最小化传感器响应与模型之间的误差。高斯牛顿法以前一般是应用于非线性的最小二乘问题,而我们在使用非参数模型建模对气体浓度进行逆向预测时,需要解决的是非参数的最小二乘问题。由于在使用局部线性核回归方法拟合非参数模型时,也得到了非参数模型关于自变量偏导数的估计,这里创造性地将这个偏导数应用到高斯牛顿法中,也解决了非参数的最小二乘问题。

虽然本文提出的方法是应用于气体浓度的预测问题,但是它对解决生物学和药学等其他领域的一些问题也具有一定的借鉴意义。例如,当遇到只能观察到响应变量、但是感兴趣的协变量和两者之间的关系都未知的情况时,就可以使用基于非参数模型对协变量进行逆向预测的方法。

在实际应用这些方法的过程中,我们发现非参数模型进行运算需要的时间多于非线性的参数模型。主要有如下两个原因:一方面,使用式(3)拟合非参数模型需要复杂的计算;另一方面,在使用非参数回归进行逆向预测步骤中,重复迭代式(6)会增加计算负担。这也是使用非参数模型进行逆向预测方法的缺陷之一。此外,在拟合非参数模型的步骤中,认为误差是独立同分布的。可以进一步研究相关误差的情况,最终的预测准确率可能会有所改善。

参考文献

[1] Gujral P, Amrhein M, Bonvin D. Drift correction in multivariate calibration models using on-line reference measurements[J]. Analytica Chimica Acta, 2009, 642: 27-36.

[2] Gujral P, Amrhein M, Wise B M, et al. Framework for explicit drift correction in multivariate calibration models[J]. Journal of Chemometrics, 2010, 24: 534-543.

[3] Trincavelli M, Coradeschi S, Loutfi A. Odour classification system for continuous monitoring applications[J]. Sensors and Actuators B: Chemical, 2009, 139(2): 265-273.

[4] Marcelloni F. Recognition of olfactory signals based on supervised fuzzy C-means and k-NN algorithms[J]. Pattern Recognition Letters, 2001, 22(9): 1007-1019.

[5] Marco S, Gutiérrez-Gálvez A. Signal and data processing for machine olfaction and chemical sensing: a review[J]. IEEE Sensors Journal, 2012, 12(11): 3189-3214.

[6] Llobet E, Brezmes J, Vilanova X, et al. Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick-film tin oxide gas sensor array[J]. Sensors and Actuators B: Chemical, 1997, 41: 13-21.

[7] Marco S, Pardo A, Davide F A M, et al. Different strategies for the identification of gas sensing systems[J]. Sensors and Actuators B: Chemical, 1996, 34: 213-223.

[8] Clifford P K, Tuma D T. Characteristics of semiconductor gas sensors. I. Steady state gas response[J]. Sensors and Actuators, 1982, 3: 233-254.

[9] Clifford P K, Tuma D T. Characteristics of semiconductor gas sensors. II. Transient response to temperature change[J]. Sensors and Actuators, 1983, 3: 255-281.

[10] Madou M J, Morrison S R. Chemical sensing with solid state devices[M]. Boston: Academic Press, 1989: 547-556.

[11] Chaiyboun A, Traute R, Kiesewetter O, et al. Modular analytical multicomponent analysis in gas sensor aarrays[J]. Sensors, 2006, 6(4): 270-283.

[12] Faraway J J. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models[M]. Boca Raton, Florida: Chapman and Hall, 2006.

[13] Pardo A, Marco S, Samitier J. Nonlinear inverse dynamic models of gas sensing systems based on chemical sensor arrays for quantitative measurements[J]. IEEE Transactions on Instrumentation and Measurement, 1998, 47(3): 644-651.

[14] Fan J, Gijbels I. Local polynomial modeling and its applications[M]. London: Chapman and Hall, 1996.

[15] Ruppert D, Wand M P. Multivariate locally weighted least squares regression[J]. The Annals of Statistics, 1994, 22: 1346-1370.

[16] Björck A. Numerical methods for least squares problems[M]. Philadelphia: Society for Industrial and Applied Mathematics, 1996.

[17] Xia Y C, Li W K. Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting[J]. Journal of Multivariate Analysis, 2002, 83(2): 265-287.

[18] Li Q, Racine J. Cross-validated local linear nonparametric regression[J]. Statistica Sinica, 2004, 14: 485-512.

[19] Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data[J]. Journal of Econometrics, 2004, 119(1): 99-130.

[20] Zhang C M. Calibrating the degrees of freedom for automatic data smoothing and effective curve checking[J]. Journal of the American Statistical Association, 2003, 98(463): 609-628.

[21] Fonollosa J, Sheik S, Huerta R, et al. Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring[J]. Sensors and Actuators B: Chemical, 2015, 215: 618-629.