高技术通讯 2014 年 第24 卷 第1 期:80 ~85

doi:10.3772/j.issn.1002-0470.2014.01.012

基于支持张量机回归的三维荧光光谱法水体有机污染物浓度检测①

杜树新② 蒋丹红 李林军

(浙江大学宁波理工学院信息科学与工程分院 浙江宁波 315100)

摘 要 根据三维荧光光谱二阶张量的数据模式特点,提出了应用基于核函数的支持张量机回归进行三维荧光光谱定量分析的方法,并用其实现了水体有机污染物浓度的检测。在建立回归校正模型中,将二阶张量数据作为模型的输入,充分利用了二阶张量原有的流形结构信息以及数据的内在相关性,提高了模型的推广能力,同时也克服了平行因子法(PARAFAC)、多维偏最小二乘算法等常规二阶校正法需要预先估计组分数、对所预估组分数敏感、要求光谱数据服从三线性分解模型的缺点。对水体中的有机污染物浓度化学耗氧量、总有机碳的检测实验,表明了基于支持张量机的三维荧光光谱分析方法提高了模型的性能,对需预先设置的模型参数不是很敏感。

关键词 光谱学,光谱分析,三维荧光光谱,有机污染物浓度检测,支持张量机(STM)

0 引 言

三维荧光光谱在水体有机污染物浓度检测中有重要应用。水体有机污染物浓度反映了化学耗氧量(chemical oxygen demand,COD)、总有机碳(total organic carbon,TOC)、生物耗氧量(biological oxygen demand,BOD)等污染物综合指标,有机污染物浓度检测是地表水和重点污染源监测的主要依据。三维荧光光谱反映了光谱强度随激发波长和发射波长变化的情况,能完整地展示光谱信息,具有检测灵敏度、选择性高,检测速度快以及无需化学试剂等优点,因而利用三维荧光光谱检测水体有机污染物浓度的研究备受关注[1,2]。本文根据三维荧光光谱二阶张量的数据模式特点,提出了一种应用基于核函数的支持张量机回归实现三维荧光光谱定量分析的方法,并用其进行了水体有机污染物综合指标的检测。

1 相关知识

三维荧光光谱数据存放在一个二维矩阵中,矩阵的行对应于激发波长,矩阵的列对应于发射波长,元素值对应于荧光强度,因此,三维荧光光谱数据也称为激发发射矩阵。目前,根据三维荧光光谱数据建立定量分析模型的方法分为基于向量的一阶校正法和基于张量的二阶校正法[3]。基于向量的一阶校正法首先将三维光谱数据展开成向量,再以该向量为输入,采用传统的光谱定量分析方法(如统计数据分析中的多元线性回归、主成分回归、偏最小二乘算法以及机器学习中的神经网络方法、支持向量机)建立光谱分析模型。基于张量的二阶校正法采用的是线性分解进技术,即通过对二阶张量的光谱数据进行线性分解而进行模型计算的。二阶校正法由于具有“二阶优势”,在三维荧光光谱分析中得到了广泛研究和应用,常见的二阶校正法是平行因子法(parallel factor analysis,PARAFAC)[4]和多维偏最小二乘方法[5]。一阶校正法将三维荧光光谱向量化后,破坏了三维光谱固有的结构信息以及数据间的内在相关性,掩盖数据原本存在的冗余信息,从而降低了定量分析模型的推广能力,降低了检测精度。二阶校正法要求光谱数据服从三线性分解模型以及需要预先估计组分数,因此对于无法预先知道水体中有机污染物组分、三维荧光光谱与有机污染物浓度之间呈现非线性关系的有机污染物综合指标检测,应用常规的二阶校正法得不到满意的建模效果。

① 国家863 计划(2009AA04Z123)和国家自然科学基金(60974111)资助项目。

② 男,1967 年生,博士,副教授;研究方向:基于机器学习的光谱在线分析技术,方法及应用;联系人,E-mail:shxdu@iipc.zju.edu.cn

(收稿日期:2013-05-06)

在数学上三维光谱数据是一个二阶张量,本文应用机器学习中的张量学习方法,提出了基于支持张量机回归的三维荧光光谱定量分析方法,并应用于水体中的化学耗氧量和总有机碳的检测。

2 基于支持张量机回归的三维荧光光谱分析方法

机器学习中的支持张量机(support tensor machine,STM)方法是支持向量机扩展到张量模式的一种监督学习方法,是由 Tao 等[6]针对张量数据的分类问题于 2005 年提出的,Guo 等[7]在 2012 年将支持张量机用于线性模型的回归估计。光谱定量分析的本质就是建立回归估计模型,因此可以将用于回归估计的支持张量机方法应用于校正模型的建立。本文针对水体三维荧光光谱数据,将基于核函数的非线性支持张量机回归方法应用于校正模型的建立。

给定训练样本及其输出 $\{X_i,y_i\},i=1,2,\cdots,M$,其中 $X_i\in\boldsymbol{R}^{n_1\times n_2}$ 为输入的三维荧光光谱,是二阶张量,$y_i\in\boldsymbol{R}$ 为输出的目标值即水体中有机物污染物浓度,M 为训练样本数量。所构造的回归估计函数为

$$f(X)=u^{\mathrm{T}}\varphi(X)v+b \tag{1}$$

式中通过非线性映射函数 $\varphi(\cdot)\in\boldsymbol{R}^{n_1\times N\times n_2}$ 将非线性光谱矩阵数据 X 映射到 N-维线性空间中,$u\in\boldsymbol{R}^{n_1},v\in\boldsymbol{R}^{n_2},b\in\boldsymbol{R}$ 为在训练过程需要计算的量(即模型待定参数)。类似于回归支持向量机方法,通过引入 ε 不敏感损失函数,将回归模型的确定转化成如下优化问题:

$$\min_{u,v,b,\xi,\xi^*}\frac{1}{2}\|u^{\mathrm{T}}v\|^2+C\sum_{i=1}^{M}(\xi_i+\xi_i^*)$$
$$\text{s.t.}\quad\begin{cases}y_i-u^{\mathrm{T}}\varphi(X_i)v-b\leqslant\varepsilon+\xi_i\\ u^{\mathrm{T}}\varphi(X_i)v+b-y_i\leqslant\varepsilon+\xi_i^*\\ \xi_i\geqslant0,\xi_i^*\geqslant0\end{cases},i=1,2,\cdots,M \tag{2}$$

其中 C,ε 为预先给定的常数,ξ_i,ξ_i^* 为松弛变量。该优化问题通过以下步骤迭代计算得到。

步骤 1:初始化向量 u,如 u 的所有元素值为 1。

步骤 2:计算 v。u 已知,采用拉格朗日乘子法求解式(2)的二次规划问题,即

$$\max_{\alpha_i,\alpha_i^*,\eta_i,\eta_i^*}\min_{v,b,\xi,\xi^*}\Big\{\frac{1}{2}\|uv^{\mathrm{T}}\|^2+C\sum_{i=1}^{M}(\xi_i+\xi_i^*)-\sum_{i=1}^{M}(\eta_i\xi_i+\eta_i^*\xi_i^*)-\sum_{i=1}^{M}\alpha_i(\varepsilon+\xi_i-y_i+u^{\mathrm{T}}\varphi(X_i)v+b)-\sum_{i=1}^{M}\alpha_i^*(\varepsilon+\xi_i^*+y_i-u^{\mathrm{T}}\varphi(X_i)v-b)\Big\}$$
$$\text{s.t.}\quad\alpha_i,\alpha_i^*,\eta_i,\eta_i^*\geqslant0,i=1,2,\cdots,M \tag{3}$$

式中 $\alpha_i,\alpha_i^*,\eta_i,\eta_i^*$ 为拉格朗日乘子,u 已知,因此不是优化变量。分别对 v,b,ξ_i,ξ_i^* 求偏导并将其置零可得

$$\begin{cases}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)=0\\ v=\frac{1}{\|u\|^2}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)(u^{\mathrm{T}}\varphi(X_i))^{\mathrm{T}}\\ \eta_i=C-\alpha_i,\quad i=1,2,\cdots,M\\ \eta_i^*=C-\alpha_i^*,\quad i=1,2,\cdots,M\end{cases} \tag{4}$$

将上式代入式(3)中可得

$$\max_{\alpha_i,\alpha_i^*}\Big\{\sum_{i=1}^{M}y_i(\alpha_i-\alpha_i^*)-\sum_{i=1}^{M}(\alpha_i+\alpha_i^*)\varepsilon-\frac{1}{2\|u\|^2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)u^{\mathrm{T}}K(X_i,X_j)u\Big\}$$
$$\text{s.t.}\quad\begin{cases}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)=0\\ 0\leqslant\alpha_i,\alpha_i^*\leqslant C,\quad i=1,2,\cdots,M\end{cases} \tag{5}$$

其中 $K(X_i,X_j)=\varphi(X_i)\varphi^{\mathrm{T}}(X_j)\in\boldsymbol{R}^{n_1\times n_1}$ 称为核函数矩阵,式(5)是一个二次凸优化问题,可求解得到最优值 α_i,α_i^*,从而由式(4)计算得到 v,即

$$\|v\|^2=\frac{1}{\|u\|^4}\sum_{i=1}^{M}\sum_{j=1}^{M}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)u^{\mathrm{T}}K(X_i,X_j)u \tag{6}$$

步骤 3:求 u 和 b。令

$$x''_j=v^{\mathrm{T}}\varphi(X_j)^{\mathrm{T}}=\frac{1}{\|u\|^2}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)u^{\mathrm{T}}K(X_i,X_j) \tag{7}$$

式中 α_i,α_i^*,u 为步骤 2 中计算得到,因此 x''_j 也可计算得到。同样,采用拉格朗日乘子法求解式(2)的二次规划问题,得到

$$\max_{\alpha_i,\alpha_i^*,\eta_i,\eta_i^*}\min_{u,b,\xi,\xi^*}\Big\{\frac{1}{2}\|uv^{\mathrm{T}}\|^2+C\sum_{i=1}^{M}(\xi_i+\xi_i^*)-\sum_{i=1}^{M}(\eta_i\xi_i+\eta_i^*\xi_i^*)-\sum_{i=1}^{M}\alpha_i(\varepsilon+\xi_i-y_i+x''_iu+b)$$

$$-\sum_{i=1}^{m}\alpha_i^*(\varepsilon+\xi_i^*+y_i-x''_iu-b)\Big\}$$

$$\text{s.t.}\quad \alpha_i,\alpha_i^*,\eta_i,\eta_i^*\geqslant 0,i=1,2,\cdots,M \tag{8}$$

分别对 u,b,ξ_i,ξ_i^* 求偏导后将其置零可得

$$\begin{cases}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)=0\\ u=\dfrac{1}{|v|^2}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)(x''_i)^{\mathrm{T}}\\ \eta_i=C-\alpha_i,i=1,2,\cdots,M\\ \eta_i^*=C-\alpha_i^*,i=1,2,\cdots,M\end{cases} \tag{9}$$

将式(9)代入式(8)得

$$\min_{\alpha_i,\alpha_i^*}\Big\{\sum_{i=1}^{M}y_i(\alpha_i-\alpha_i^*)-\sum_{i=1}^{M}(\alpha_i+\alpha_i^*)\varepsilon -\frac{1}{2|v|^2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)x''_i(x''_j)^{\mathrm{T}}\Big\}$$

$$\text{s.t.}\quad\begin{cases}\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)=0\\ 0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,2,\cdots M\end{cases} \tag{10}$$

式中 $\|v\|^2,x''_i$ 已知,因此可针对该二次凸优化问题求解得到最优值 α_i,α_i^*,从而由式(9)计算得到 u。

在求解式(10)最优二次规划问题时,由最优化的充要条件,在最优点,拉格朗日乘子与约束的乘积为 0,即

$$\begin{aligned}&\alpha_i(y_i-x''_iu-b-\varepsilon-\xi_i)=0\\ &\alpha_i^*;(x''_iu+b-y_i-\varepsilon-\xi_i^*)=0\\ &\eta_i\xi_i=0\rightarrow(C-\alpha_i)\xi_i=0\\ &\eta_i^*\xi_i^*=0\rightarrow(C-\alpha_i^*)\xi_i^*=0\end{aligned} \tag{11}$$

由上式,对于标准支持向量 $0<\alpha_i<C,\xi_i=0$,因此有

$$y_i-x''_iu-b-\varepsilon=0 \tag{12}$$

即

$$b=y_i-x''_iu-\varepsilon \tag{13}$$

步骤 4:循环执行步骤 2 和步骤 3,直到上次循环与本次循环得到的 $u\in\boldsymbol{R}^{n_1}$、$v\in\boldsymbol{R}^{n_2}$ 充分接近。到达最优化后,根据所计算的 u、v 和 b,由式(1)得到回归模型,即校正模型。

从上述基于核函数的支持张量机回归方法中可看出,核函数实际上是一个矩阵,这是从核函数角度来将区别于支持向量机的所在。同时,核函数矩阵的计算仅限于步骤 2 式(5)中的 v 优化计算,不涉及到式(10)中的优化计算。核函数矩阵的计算如下:

$$\begin{aligned}K(X_i,X_j)&=\varphi(X_i)\varphi^{\mathrm{T}}(X_j)\\ &=\begin{bmatrix}\varphi(x_i^1)\\ \varphi(x_i^2)\\ \vdots\\ \varphi(x_i^{n_1})\end{bmatrix}\cdot[\varphi^{\mathrm{T}}(x_j^1)\quad\varphi^{\mathrm{T}}(x_j^2)\quad\cdots\quad\varphi^{\mathrm{T}}(x_j^{n_1})]\\ &=\begin{bmatrix}\varphi(x_i^1)\cdot\varphi^{\mathrm{T}}(x_j^1) & \varphi(x_i^1)\cdot\varphi^{\mathrm{T}}(x_j^2) & \cdots & \varphi(x_i^1)\cdot\varphi^{\mathrm{T}}(x_j^{n_1})\\ \varphi(x_i^2)\cdot\varphi^{\mathrm{T}}(x_j^1) & \varphi(x_i^2)\cdot\varphi^{\mathrm{T}}(x_j^2) & \cdots & \varphi(x_i^2)\cdot\varphi^{\mathrm{T}}(x_j^{n_1})\\ \vdots & \vdots & & \vdots\\ \varphi(x_i^{n_1})\cdot\varphi^{\mathrm{T}}(x_j^1) & \varphi(x_i^{n_1})\cdot\varphi^{\mathrm{T}}(x_j^2) & \cdots & \varphi(x_i^{n_1})\cdot\varphi^{\mathrm{T}}(x_j^{n_1})\end{bmatrix}\\ &=\begin{bmatrix}k(x_i^1,x_j^1) & k(x_i^1,x_j^2) & \cdots & k(x_i^1,x_j^{n_1})\\ k(x_i^2,x_j^1) & k(x_i^2,x_j^2) & \cdots & k(x_i^2,x_j^{n_1})\\ \vdots & \vdots & \vdots & \vdots\\ k(x_i^{n_1},x_j^1) & k(x_i^{n_1},x_j^1) & \cdots & k(x_i^{n_1},x_j^{n_1})\end{bmatrix}\\ &\in\boldsymbol{R}^{n_1\times n_1}\end{aligned} \tag{14}$$

式中 $x_i^p\in\boldsymbol{R}^{n_2}$ 表示第 i 个三维荧光光谱数据 $X_i\in\boldsymbol{R}^{n_1\times n_2}$ 中的第 p 个行向量,$k(x_i^p,x_j^q)$ 为核函数,如果采用径向基核函数,则该核函数为

$$k(x_i^p,x_j^q)=\exp\Big(-\frac{\|x_i^p-x_j^q\|^2}{\sigma^2}\Big) \tag{15}$$

本文所给出的支持张量机方法与文献[6,7]的线性支持张量机方法相比,由于引入了核函数矩阵,因此可适用于非线性数据的回归估计,与文献[8]的核支持张量机回归方法相比,尽管思路相同,但本文详细推导了 u、v 和 b 的计算过程。同时本文将支持张量机应用于三维荧光光谱定量分析,提出了一类基于支持张量机的二阶校正法,从而在二阶校正法中引入机器学习方法,克服了目前二阶校正法中通过张量分解进行求解的缺陷。

3 实 验

3.1 水样、光谱采集和分析值测量

用于实验的水样采集自某市地表水及生活排污水,共 32 个水样。对水样进行三维荧光光谱的采集、化学耗氧量和总有机碳指标化学法参考值的测量。三维荧光光谱通过日本日立公司的 F-4500 型荧光光谱仪测量,激发波长为 225~400nm,发射波长为 250~700nm,采样波长间隔为 5nm,扫描速度

为2400nm/min。总有机碳采用日本岛津公司的TOC－VCSH总有机碳分析仪测量得到,化学耗氧量则委托环境检测站专业实验室采用国家标准测定方法GB11914－89测定。图1为32个水样的COD和TOC图。

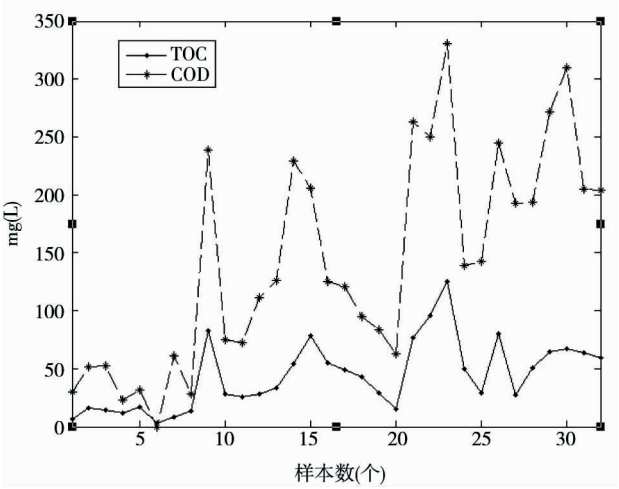

图1　32个水样的TOC和COD值

3.2　实验结果及其分析

(1)为了比较支持张量机与其他建模方法的模型性能,除了采用支持张量机(STM)进行建模外,还分别采用一阶校正法中的主元回归方法(principal component regress,PCR)、偏最小二乘算法(partial least squares,PLS)、支持向量机(support vector machine,SVM),以及二阶校正法中的平行因子法(PARAFAC)、多维偏最小二乘(multi-way partial least squares,N-PLS)进行实验,支持向量机和支持张量机中选取的核函数为径向基核函数。比较结果如表1所示,其中RMSEP为分析误差均方根,R为模型预测值与化学分析值之间的相关系数,其计算参见文献[1]。表1的建模方法中SVM1是激发发射矩阵按行展开得到向量后再采用支持向量机方法进行建模,SVM2是激发发射矩阵按列展开得到向量后再采用支持向量机方法进行建模。由表1可知,对于COD,STM相对于PCR、PLS、SVM1、SVM2、PARAFAC、N-PLS,相关系数分别提高了1.74%、11.04%、2.63%、4.78%、12.05%、1.15%,分析误差均方根降低了1.76%、37.26%、5.49%、27.98%、41.48%、2.75%;对于TOC,STM相对于PCR、PLS、SVM1、SVM2、PARAFAC、N-PLS分别提高了7.94%、8.91%、5.73%、4.57%、2.56%、0.42%,分析误差均方根降低了38.62%、39.94%、34.98%、32.66%、5.00%、2.74%。因此模型性能有较大的提高,尤其对于TOC的检测。

表1　不同建模方法的模型性能比较

建模方法	COD		TOC	
	R	RMSEP	R	RMSEP
PCR	0.9426	14.7842	0.8726	22.9823
PLS	0.8534	23.1475	0.8634	23.4872
SVM1	0.9341	15.3647	0.8936	21.6948
SVM2	0.9134	20.1648	0.9046	20.9497
PARAFAC	0.8437	24.8132	0.9236	14.8487
N-PLS	0.9483	14.9326	0.9439	14.5042
STM	0.9593	12.6528	0.9479	13.9257

(2)支持张量机方法事先需要确定的参数是ε、C以及核函数参数σ,其中$\varepsilon>0$为与模型估计精度直接相关的设计参数,$C>0$为惩罚系数,C越大表示对超出ε管道数据点的惩罚越大。表2给出了ε和σ固定、C取不同值对应的模型性能比较,表3给出了C和σ固定、ε取不同值对应的模型性能比较,表4给出了ε和C固定、σ取不同值对应的模型性能比较。由表2可知,在C由5变化到95过程中,模型性能会有些波动,用标准差描述波动情况,则对于COD,相关系数的标准差为0.01,分析误差均方根的标准差为1.17,对于TOC,相关系数的标准差为0.006,分析误差均方根的标准差为0.84。由表3可知,在ε由0.002变化到0.02过程中,对于COD,相关系数的标准差为0.009,分析误差均方根

表2　不同的C对模型性能的影响($\varepsilon=0.01,\sigma=0.1$)

C	COD		TOC	
	R	RMSEP	R	RMSEP
5	0.9276	16.6797	0.9326	15.3214
15	0.9368	15.8036	0.9423	14.7036
25	0.9593	12.6528	0.9479	13.9257
35	0.9309	16.0286	0.9367	15.8354
45	0.9429	15.2087	0.9372	15.6578
55	0.9267	16.8614	0.9254	16.9615
65	0.9426	15.2147	0.9429	14.6793
75	0.9389	15.6431	0.9354	15.9735
85	0.9419	15.3764	0.9402	14.9764
95	0.9498	14.9126	0.9379	15.2126

表 3 不同的 ε 对模型性能的影响(C = 25,σ = 0.1)

ε	COD		TOC	
	R	RMSEP	R	RMSEP
0.002	0.9329	16.3197	0.9318	16.3246
0.004	0.9431	15.7091	0.9332	16.1032
0.006	0.9394	16.1629	0.9297	16.5669
0.008	0.9298	16.8359	0.9367	15.8354
0.01	0.9593	12.6528	0.9479	13.9257
0.012	0.9454	15.1563	0.9354	15.9648
0.014	0.9471	15.0793	0.9437	14.4791
0.016	0.9364	15.8435	0.9361	15.8439
0.018	0.9491	14.9764	0.9406	14.9763
0.02	0.9329	16.1349	0.9399	14.9975

表 4 不同的 σ 对模型性能的影响(C = 25,ε = 0.01)

σ	COD		TOC	
	R	RMSEP	R	RMSEP
0.02	0.9306	16.7214	0.9294	16.4216
0.04	0.9346	16.2896	0.9338	15.9894
0.06	0.9432	15.2398	0.9428	14.6286
0.08	0.9389	15.7698	0.9395	15.7732
0.1	0.9593	12.6528	0.9479	13.9257
0.12	0.9428	15.3278	0.9416	14.9140
0.14	0.9312	16.4319	0.9327	16.0976
0.16	0.9416	15.8426	0.9425	14.7638
0.18	0.9325	16.4649	0.9346	15.7048
0.2	0.9412	15.8856	0.9436	14.6796

的标准差为 1.16,对于 TOC,相关系数的标准差为 0.006,分析误差均方根的标准差为 0.86。由表 4 可知,在 σ 由 0.02 变化到 0.2 过程中,对于 COD,相关系数的标准差为0.0085,分析误差均方根的标准差为 1.16,对于 TOC,相关系数的标准差为 0.006,分析误差均方根的标准差为 0.91。由此可见,C、ε 和 γ 的取值变化对模型性能影响不大,即模型对 C、ε 和 σ 的变化不是很敏感。

4 结 论

根据三维荧光光谱用二阶张量描述的特点,论文提出了应用基于核函数的支持张量机回归实现三维荧光光谱定量分析的方法,并应用于水体中有机污染物浓度化学耗氧量(COD)和总有机碳(TOC)的检测。

在建立回归校正模型中,将二阶张量数据作为模型的输入,充分利用了二阶张量原有的流形结构信息以及数据的内在相关性,提高了模型的推广能力,同时在建模过程中应用核函数技术使得定量分析方法适用于非线性光谱数据,同时分析方法不需要预先估计组分数,因此克服了常规二阶校正法需要预先估计组分数、对所预估组分数敏感、要求光谱数据服从三线性分解模型的缺点。在对水体中化学耗氧量和总有机碳的实验中,通过不同建模方法建立的模型的性能比较表明,基于核函数的支持张量机方法所建立的模型的性能要优于采用现有的主元回归方法、偏最小二乘算法、支持向量机、平行因子法、多维偏最小二乘所建立的模型的性能。同时,对不同预设参数的比对实验说明,本文所提出的三维荧光光谱定量分析方法对需预先设置的模型参数不是很敏感。

参考文献

[1] 杜树新,杜阳锋,武晓莉. 基于三维荧光导数光谱的水体有机污染物浓度检测. 光谱学与光谱分析,2010,30(12):3268-3271

[2] Hudson N,Baker A,Reynolds D. Fluorescence analysis of dissolved organic matter in natural,waste and polluted waters—a review. *River Research and Applications*,2007,23:631-649

[3] 聂瑾芳. 二阶张量校正新算法研究及其在三维荧光分析中的应用:博士学位论文,长沙:湖南大学化学化工学院,2010.1-10

[4] Bro R. PARAFAC:tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*,1997,38:149-171

[5] Bro R. Multiway calibration:multilinear PLS. *Journal of Chemometrics*,1996,10:47-61

[6] Tao D,Li X,Hu W,Maybank S,Wu X. Supervised tensor learning. In:Proceedings of the Fifth International Conference on Data Mining, Houston,Texas,USA,2005,450-457

[7] Guo W,Kotsia I,Patras I. Tensor learning for regression. *IEEE Transactions on Image Processing*,2012,21(2):816-827

[8] Gao C,Wu X-J. Kernel support tensor regression,*Procedia Engineering*,2012,29:3986-3990

[9] 杜树新,杜阳锋,武晓莉. 基于 Savizky-Golay 多项式的三维荧光光谱的曲面平滑方法,光谱学与光谱分析,2011,31(2):440-443

Detection of dissolved organic matter using three-dimensional fluorescence spectrometry based on support tensor machine regress

Du Shuxin, Jiang Danhong, Li Linjun
(School of Information Science and Engineering, Ningbo Institute of Technology
Zhejiang University, Ningbo 315100)

Abstract

The kernel function based support tensor machine regression was used to detect dissolved organic matter in water by using the three-dimensional fluorescence spectrometry. The fluorescence spectrometry with two-order tensor was taken as the input of the calibration model during the model's establishing, and the original manifold structural information and the intrinsic data relationship were fully utilized to increase the calibration model's generalization capability. Moreover, the disadvantages of the traditional methods such as the parallel factor analysis (PARAFAC) and the multi-way partial least squares (N-PLS) in the aspects of need of the component number estimation and sensitivity of the estimated component number and requirements of trilinear decomposition were overcomed. The results of the experiment on detecting total organic carbon (TOC) and chemical oxygen demand (COD) in water showed that the model performance was improved by the proposed method and the model was unsensitive to model parameters variation.

Keywords: spectroscopy, spectroscopic analysis, three-dimensional fluorescence spectrometry, detection of dissolved organic matter, support tensor machines (STM)