基于分区建模的锌液痕量铜离子光谱检测方法

朱红求，吴书君，李勇刚*，阳春华
中南大学信息科学与工程学院，湖南 长沙 410083

摘要 炼锌溶液中痕量铜离子的光谱信号被掩蔽、干扰严重，以及铜离子的非线性特性在高、低浓度区间的显著差异，都会导致痕量铜离子的浓度检测比较困难。针对该问题，提出了一种基于分区建模的锌液中痕量铜离子的光谱检测方法。该方法采用导数光谱结合小波去噪的方法对光谱信号进行预处理，重现待测铜离子的谱峰。以相关系数-稳定性值作为变量的评价指标对波长变量进行排序，并结合支持向量回归(SVR)模型选取最佳波长变量，在此基础上，根据混合溶液中铜离子光谱信号非线性特性将浓度划分区间，并分别针对每个区间建立粒子群优化支持向量回归(PSO-SVR)模型，计算出铜离子的质量浓度。将所提方法与现有多种回归方法进行比较，结果表明：所提方法将预测方均根误差降低至 0.0678，模型决定系数提高至 99.61%，该方法的最大相对误差为 6.94%，平均相对误差为 2.74%。

关键词 光谱学；吸收光谱；相关系数-稳定性值；支持向量机分区建模；炼锌溶液；痕量铜离子
中图分类号 O433 **文献标识码** A **doi**：10.3788/AOS201939.0230002

A Spectrophotometric Detecting Method of Trace Copper Ion in Zinc Solution Based on Partition Modeling

Zhu Hongqiu, Wu Shujun, Li Yonggang*, Yang Chunhua
College of Information Science and Engineering, Central South University, Changsha, Hunan 410083, *China*

Abstract The mass concentration detection of trace copper ion in zinc solution is difficult because of trace copper spectral signals masking, serious interference and significant nonlinearity difference of copper ion in the high and low mass concentration intervals. Aiming at this issue, we propose a spectrophotometric detecting method of trace copper ion in zinc solution based on partition modeling. The derivative spectrum combined with wavelet denoising is used to preprocess the spectral signal and reproduce the spectral peak of the copper ion to be measured. The wavelength variables are ranked by the correlation coefficient-stability value, which serves as the evaluation index of the variables, and the support vector regression (SVR) model is used to select the optimal wavelength variables. On this basis, the mass concentration of the copper ions is divided into several intervals according to the nonlinear characteristics in the mixed solution. The particle swarm optimization support vector regression (PSO-SVR) model is respectively established for each interval to compute the concentration of copper ions. The proposed method is compared with many existing regression methods. The results show that the predicted root mean square error obtained with the proposed method is reduced to 0.0678, and the model determination coefficient is increased to 99.61%. The maximum relative error obtained with the method is 6.94% and the average relative error is 2.74%.

Key words spectroscopy; absorption spectroscopy; correlation coefficient-stability value; partition modeling of support vector machine; solution of zinc hydrometallurgy; trace copper ion
OCIS codes 300.1030; 300.6540; 300.6550; 300.6420

1 引 言

锌液中痕量铜离子的浓度是湿法冶锌净化工序的重要工艺参数指标，适量杂质铜离子可作为除钴反应的活化剂，但其浓度过高就会降低电解电效[1]，因此铜离子浓度的准确检测是后续工艺稳定的前提和基础[2]。紫外可见分光光度法(UV-Vis)具有准确度高、分析方法简单等优点，已被广泛应用于溶液

收稿日期：2018-08-08；**修回日期**：2018-09-14；**录用日期**：2018-09-29
基金项目：国家自然科学基金重点项目(61533021)、国家自然科学基金创新研究群体项目(61621062)、中南大学中央高校基本科研业务费专项资金(2018zzts063)

* **E-mail**：liyonggang@csu.edu.cn

中低浓度多金属离子的检测[3]。在锌液中，溶液基体锌离子 Zn(II)与痕量待测铜离子 Cu(II)的质量浓度之比高达 17 万～20 万，故而，Cu(II)光谱信号会被高浓度的 Zn(II)及钴离子 Co(II)等杂质金属离子信号严重掩蔽。此外，溶液中多种化学性质相近的金属离子相互干扰，光谱信号的重叠也很严重，这使得 Cu(II)浓度的检测比较困难。

随着分子光谱分析技术的发展，分析校正方法不断发展和完善，常用的线性回归方法有主成分分析法(PCR)[4]、偏最小二乘法(PLS)[5]，非线性校正方法有人工神经网络(ANN)[6]、支持向量回归(SVR)[7]等。锌液中金属离子光谱信号干扰、重叠严重，传统的基于全光谱波段的分析校正方法包含大量噪声和冗余信息，难以满足检测精度等的要求。针对这一局限性，近年来研究人员发展了结合波长变量选择的回归方法，主要有间隔偏最小二乘法(IPLS)[8]、移动窗口偏最小二乘法(MWPLS)[9]、蒙特卡罗无信息变量消除法(MC-UVE)[10]、竞争性自适应加权法(CARS)[11]等。这些方法可在满足线性、加和性、近似服从朗伯-比尔定律的情况下取得较高的精度。其中：IPLS 和 MWPLS 主要针对光谱区间进行选择[12]；MC-UVE 和 CARS 针对波长点进行选择；MC-UVE 可用于剔除光谱中的噪声波长点；CARS 可以减小共线性变量的影响，并去除无用的信息变量。当光谱变量和浓度之间由于高浓度基体离子的影响及各组分的相互作用而出现信号严重重叠、掩蔽或非线性时，IPLS 和 MWPLS 由于缺乏经验而难以选取区间宽度，并可能会遗漏所选区间以外的重要的特征波长变量，而 MC-UVE 和 CARS 在基体离子谱峰处的指标较好，容易选入基体离子信息波长点，从而影响痕量离子的模型精度[8]，难以得到理想的效果。

针对锌液光谱信号干扰、掩蔽、重叠严重、非线性强等特点，本课题组提出了一种基于变量排序选择的支持向量机(SVM)分区建模方法，建立混合溶液光谱信号与痕量 Cu(II)浓度之间的模型。该方法利用去噪后的导数光谱重现被掩蔽的待测Cu(II)光谱信号，并提出以相关系数-稳定性值作为重要性指标对波长变量进行排序、选择，从而剔除噪声信息和空白信息，减小 Zn(II)与Co(II)的干扰，最大程度地保留 Cu(II)的灵敏区域，并通过减少变量个数来提高模型的效率。同时，分析光谱信号在不同浓度铜离子区间呈现的非线性特性，采用先分区再回归的建模方法，通过支持向量分类(SVC)对离子浓度进行分区，分别针对每一个区间重建 SVR 模型，以增强区间内部特征信息的聚合，提高模型精度。

2　实验部分

2.1　仪器和试剂

仪器选用 T9 双光束紫外可见分光光度仪(北京普析通用仪器有限责任公司)。试剂包括：显色剂 4-亚硝基-3-羟基-2，7-萘二磺酸二钠(亚硝基 R 盐)溶液，其质量分数为0.4%；醋酸-醋酸钠缓冲液，其 pH＝5.5；Zn(II)标准溶液，其质量浓度为 180 g/L；Cu(II)、Co(II)标准溶液，质量浓度均为 50 mg/L。以上所用试剂均为分析纯。

2.2　实验方法

混合溶液中的 Zn(II)是高质量浓度的基体离子，Co(II)为干扰离子，Cu(II)为痕量待测离子。按照实验要求移取一定量上述 Zn(II)、Cu(II)、Co(II)标准溶液于 25 mL 容量瓶中，其中 Zn(II)的质量浓度范围为 70～100 g/L，间隔为 10 g/L；Cu(II)、Co(II)的质量浓度范围均为 0.5～4.0 mg/L，间隔均为 0.5 mg/L。然后，依次加入质量分数为 0.4%的亚硝基 R 盐 2 mL，醋酸-醋酸钠缓冲液 5 mL，再用去离子水稀释至刻度定容。以空白试剂(不含有金属离子)作为参比，以 1 nm 为间隔，以 400～800 nm 为扫描波长范围，测量单离子和 62 组混合溶液的光谱信号。

Zn(II)、Cu(II)、Co(II)单离子的原始光谱信号如图 1 所示。从图 1 可以看出：在 400～500 nm 波段，痕量 Cu(II)的信号被 Zn(II)严重掩蔽；500 nm 波段以后，Zn(II)的信号较弱，但干扰离子 Co(II)在 400～600 nm 波段的谱峰较高，基本掩蔽了 Cu(II)的信号。在 Zn(II)、Co(II)两种离子的掩蔽作用下，Cu(II)光谱信号的检测比较困难。此外，由于溶液中含有大量的基体离子 Zn(II)，溶液环境复杂，离子间的相互影响显著，离子质量浓度与光谱信号之间不再满足线性关系。

为解决原始光谱信号中 Cu(II)信号完全被 Zn(II)、Co(II)信号重叠、掩蔽的问题，重现 Cu(II)谱峰，采用导数光谱结合小波去噪的方法对光谱信号进行预处理。图 2 为 3 种离子的一阶导数去噪光谱信号，与图 1 相比，Cu(II)在 500～520 nm 范围内出现了独立的谱峰，受 Zn(II)、Co(II)的干扰远远小于原始光谱信号。针对 62 组混合溶液在 400～800 nm全波段的导数去噪光谱信号，采用 Kennard-

Stone(KS)算法对数据集进行划分，将其中的48组混合溶液作为校正集，14组作为验证集。

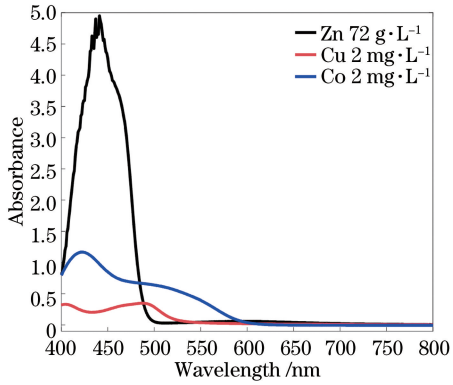

图1　Zn(II)、Cu(II)、Co(II)单离子的原始光谱信号

Fig. 1　Original spectral signals of Zn(II), Cu(II), Co(II)

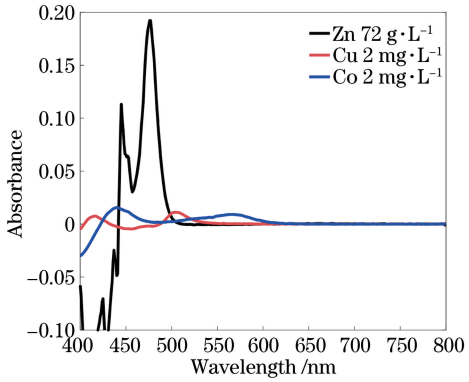

图2　Zn(II)、Cu(II)、Co(II)单离子的导数去噪光谱信号

Fig. 2　Derivative denoising spectral signals of Zn(II), Cu(II), Co(II)

3　基于变量排序选择的分区回归建模方法

为解决锌液中高浓度Zn(II)背景对痕量Cu(II)信号造成的掩蔽、非线性以及Co(II)等痕量杂质离子对Cu(II)信号造成的重叠、干扰等问题，筛选出包含Cu(II)信息量大的波长变量，以提高模型精度，运用基于变量排序选择的SVM分区建模方法，建立混合溶液光谱信号与痕量Cu(II)质量浓度之间的模型。首先，获取Zn(II)、Cu(II)、Co(II)混合溶液的导数光谱信号；之后，以相关系数-稳定性值作为重要性指标对波长变量进行排序，按序选取不同个数的波长变量建立SVR模型，以交叉验证均方误差(CVmse)最小、变量个数最少为原则选取最佳波长变量；然后，基于混合溶液光谱信号在高、低浓度铜离子区间呈现不同非线性特性的数据统计现象，采取先分区后建模的方法，通过粒子群优化-支持向量分类(PSO-SVC)方法对离子浓度进行分区，并在每一个浓度区间重建粒子群优化-支持向量回归(PSO-SVR)模型；最后，采用建立的模型预测待测痕量离子的质量浓度区间，并计算其质量浓度。

3.1　分区建模

锌液中含有高浓度的基体离子Zn(II)以及多种金属离子，如Co(II)、Cu(II)等，溶液成分复杂，多种金属离子的化学特性相近，在紫外可见分光光度法测定过程中，受电荷相互作用的影响，摩尔吸收系数会发生改变，严重偏离朗伯-比尔定律成立所要求的稀溶液、离子之间无相互干扰等前提假设条件，并且由于仪器色散元件杂散光、光源能量等的限制，待测痕量Cu(II)的质量浓度与光谱信号不再满足良好的线性关系。

在高浓度Zn(II)背景以及痕量Co(II)干扰的情况下，Cu(II)的质量浓度在0～4.2 mg/L范围内变化时，光谱信号的变化曲线如图3所示。13组混合溶液的导数光谱信号如图3(a)中的大图所示，根据图2中单离子光谱信号的分析可知，Cu(II)在495～525 nm波段的波峰显著，且受Zn(II)、Co(II)的干扰较小，因此13组混合溶液光谱信号的差异主要集中在该波段。为凸显光谱信号的差异，使用495～525 nm波段的放大示意图，即图3(a)中右上角小图，描述不同质量浓度的Cu(II)在495～525 nm波段内的光谱信号变化。

从图3(a)中选取具有象征意义的9个波长点观察Cu(II)质量浓度与光谱信号之间的关系，如图3(b)所示：当Cu(II)的质量浓度较低时，Cu(II)光谱信号微弱，并极大程度地被Zn(II)、Co(II)掩蔽，离子相互影响，干扰严重，非线性强；当Cu(II)的质量浓度较高时，Zn(II)与Cu(II)的质量浓度比有所降低，Zn(II)对Cu(II)信号的干扰、掩蔽作用降低，Cu(II)质量浓度与光谱信号之间的相关性增强，线性度增强。

根据Cu(II)的不同质量浓度区间对光谱信号影响程度的不同，将Cu(II)的浓度划分成两个区间，并分别在低浓度区间、高浓度区间对光谱信号进行拟合，表1列出了500，520，540 nm波长点下的全区间线性拟合、全区间非线性拟合和分区拟合的结果。与全区间线性拟合相比，分区拟合的方均根误差(RMSE)降低了37.9%～59.2%，9个波长点的RMSE平均降低了50.1%；与全区间非线性拟合相比，RMSE降低了11.0%～35.1%，9个波长点的RMSE平均降低了23.4%。由此可见，分区拟合的

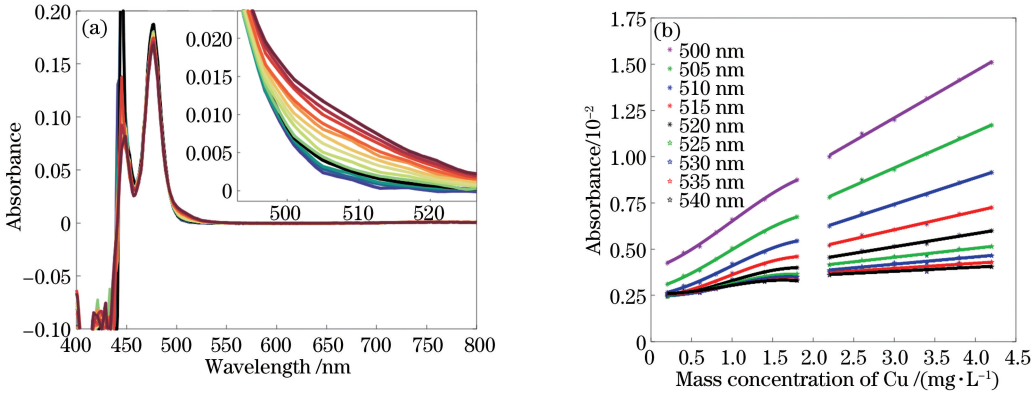

图 3　(a)高质量浓度 Zn(II)背景以及痕量 Co(II)干扰下,不同质量浓度 Cu(II)的光谱信号;(b) 9 个波长点下 Cu(II)质量浓度与光谱信号的关系

Fig. 3　(a) Spectral signals of different mass concentrations of Cu(II) in the background of high mass concentration of Zn(II) and trace concentration of Co(II); (b) relationship between Cu(II) mass concentration and spectral signal at 9 wavelength points

表 1　不同波长点下 Cu(II)质量浓度 c 与光谱信号的拟合结果

Table 1　Fitting functions of Cu(II) mass concentration c and spectral signal at different wavelengths

Wavelength /nm	Fitting method	Fitting function
500	Linear fitting	$2.8\times10^{-3}c+3.8\times10^{-3}$
	Nonlinear fitting	$-2.2\times10^{-5}c^3+6.1\times10^{-5}c^2+2.9\times10^{-3}c+3.6\times10^{-3}$
	Partition fitting	1) $-6.1\times10^{-4}c^3+1.8\times10^{-3}c^2+1.4\times10^{-3}c+3.9\times10^{-3}$; 2) $2.5\times10^{-3}c+4.5\times10^{-3}$
520	Linear fitting	$9.1\times10^{-4}c+2.4\times10^{-3}$
	Nonlinear fitting	$-1.8\times10^{-6}c^3-6.0\times10^{-5}c^2+1.2\times10^{-3}c+2.2\times10^{-3}$
	Partition fitting	1) $-5.4\times10^{-4}c^3+1.5\times10^{-3}c^2-1.3\times10^{-4}c+2.4\times10^{-3}$; 2) $7.1\times10^{-4}c+3.0\times10^{-3}$
540	Linear fitting	$3.8\times10^{-4}c+2.6\times10^{-3}$
	Nonlinear fitting	$-1.3\times10^{-6}c^3-5.4\times10^{-5}c^2+6.3\times10^{-4}c+2.4\times10^{-3}$
	Partition fitting	1) $-4.5\times10^{-4}c^3+1.2\times10^{-3}c^2-3.2\times10^{-4}c+2.6\times10^{-3}$; 2) $2.2\times10^{-4}c+3.1\times10^{-3}$

效果更佳,能够凸显溶液特征,增强区间内部特征信息的聚合。

3.2　基于相关系数-稳定性值的波长变量排序选择方法

锌液中 Zn(II)、Co(II)、Cu(II)3 种金属离子的化学特性比较相似,在检测溶液中形成的络合物的吸收光谱也相近,高浓度的 Zn(II)对 Cu(II)信号掩蔽严重,Co(II)信号的灵敏度高,谱峰宽,Cu(II)信号的灵敏度弱,谱峰窄。根据图 3(b)以及表 1 可知,不同的波长点下,Cu(II)光谱信号的灵敏度显著不同,例如:在 500 nm 处,Cu(II)显现出独立的谱峰,且受 Zn(II)、Co(II)信号掩蔽、干扰的作用较小,因此光谱信号变化灵敏;在 540 nm 处,Cu(II)信号的灵敏度低,Cu(II)、Co(II)光谱重叠严重,因此光谱信号变化缓慢。为了筛选出包含 Cu(II)信息量大的变量,并剔除对 Zn(II)和 Co(II)信号敏感、干扰严重的冗余变量,本研究提出了相关系数-稳定性值,并以此为重要性指标对波长变量进行排序选择。

复杂的溶液环境以及离子间的相互影响会造成 Cu(II)的质量浓度与光谱信号之间呈现非线性,但二者仍存在正相关性,光谱信号与待测组分之间的相关系数越大,表示波长变量包含的 Cu(II)的质量浓度信息越多,信号灵敏度越高,受其他离子的干扰越小,含有的空白信息和冗余噪声也越少。但是,普通相关系数法对样本的依赖性高,加入或丢失一个随机样本都会造成数值变动,尤其是在样本分布不均衡的情况下。MC-UVE 是一种基于模型变量稳定性值的无信息变量剔除方法,通过蒙特卡罗随机采样尽可能充分利用样本集,减少样本不均衡带来的误差。为提高波长变量选取的稳定性和模型可靠

性，将相关系数与 MC-UVE 相结合，提出相关系数-稳定性值这一指标对波长变量的重要性进行评价，并按照该值从大到小的顺序对波长变量进行排序。每个波长变量 j 的相关系数-稳定性值 I_j 为

$$I_j=\frac{f_{\mathrm{mean}}(R_{kj})}{f_{\mathrm{std}}(R_{kj})},\quad j=1,2,\cdots,J;k=1,2,\cdots,K,\tag{1}$$

式中：J 为波长变量总数；K 为蒙特卡罗采样次数；f_{mean} 为求平均值的函数；f_{std} 为求标准差的函数；R_{kj} 为第 k 次蒙特卡罗采样的 n 个采集样本中，波长变量 j 对应的光谱信号 $\boldsymbol{x}_j$ 和待测痕量 Cu(II)质量浓度 $\boldsymbol{y}$ 之间的相关系数，即

$$R_{kj}=\frac{\sum_{i=1}^{n}(\boldsymbol{x}_{ij}-\bar{x}_j)(\boldsymbol{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}(\boldsymbol{y}_i-\bar{y})^2}},\tag{2}$$

其中

$$\begin{cases}\bar{x}_j=\dfrac{\left(\sum_{i=1}^{n}\boldsymbol{x}_{ij}\right)}{n},i=1,2,\cdots,n\\ \bar{y}=\dfrac{\left(\sum_{i=1}^{n}\boldsymbol{y}_i\right)}{n},i=1,2,\cdots,n\end{cases}。\tag{3}$$

对波长变量排序后，波长变量选取个数对模型的稳定性和泛化能力尤为重要。过少的变量个数虽然会降低冗余噪声的选入，但同时会丢失样本的部分信息而无法充分反映样本的特征；若变量个数过多，则会引入大量噪声、空白信息和干扰波段，降低整个变量集合对模型的贡献度。通常，MC-UVE 可通过经验设定稳定性阈值来合理地选择变量个数，将波长变量稳定性值在阈值中间的无信息变量剔除。但高浓度背景下痕量离子浓度的研究很少，缺乏经验支撑，因此波长变量个数的选择需要结合样本特性和回归模型，以增强其适应性和针对性。采用回归的方法分别按序以不同个数波长变量建立回归模型，以回归模型 CVmse 最小为指标确定最佳的波长变量；同时，为尽可能减少建模波长变量，减少运行时间，在满足上述选取指标的同时，应选取尽可能少的波长变量，这样模型的效率就会越高。

3.3　基于粒子群优化-支持向量机(PSO-SVM)的分区回归方法

SVM 是基于统计学习理论发展起来的一种模式识别方法，已成功应用于 SVC[13] 和 SVR[14]，是光谱分析中最常用的一种非线性校正方法，在分子光谱分析和化学计量学中应用广泛。

以 UV-Vis 获取的 N 个样本的光谱信号 $\boldsymbol{x}$ 作为输入，待测痕量 Cu(II)的质量浓度 $\boldsymbol{y}$ 作为输出，则，SVR 的最优化问题为

$$\begin{aligned}&\min\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+C\sum_{i=1}^{N}(\xi_i+\hat{\xi}_i),\\&\text{s.t.}\quad f(\boldsymbol{x}_i)-\boldsymbol{y}_i\leqslant\xi_i+\varepsilon,\\&\qquad\boldsymbol{y}_i-f(\boldsymbol{x}_i)\leqslant\hat{\xi}_i+\varepsilon,\\&\qquad\xi_i,\hat{\xi}_i\geqslant0,i=1,2,\cdots,N,\end{aligned}\tag{4}$$

式中：$\boldsymbol{\omega}$ 为权重向量；C 为惩罚系数；ξ_i 和 $\hat{\xi}_i$ 为松弛变量；ε 为偏差。引入 Lagrange 函数将(4)式中的问题转化为对偶问题，求解得到痕量 Cu(II)质量浓度的 SVR 回归函数为

$$f(\boldsymbol{x})=\sum_{i=1}^{N}(\boldsymbol{\alpha}_i-\hat{\boldsymbol{\alpha}}_i)k(\boldsymbol{x}_i,\boldsymbol{x})+b,\tag{5}$$

式中：$k(\boldsymbol{x}_i,\boldsymbol{x})$ 为内核函数，使用复杂度低、灵活性高的径向基(RBF)核函数；$\boldsymbol{\alpha}_i$ 和 $\hat{\boldsymbol{\alpha}}_i$ 为拉格朗日乘数；b 为阈值。

同理可得到 SVC 的表达式：

$$f(\boldsymbol{x})=f_{\mathrm{sgn}}\left[\sum_{i=1}^{N}(\boldsymbol{\alpha}_i-\hat{\boldsymbol{\alpha}}_i)k(\boldsymbol{x}_i,\boldsymbol{x})+b\right],\tag{6}$$

式中：f_{sgn} 为符号函数。

惩罚系数 C 和 RBF 核函数参数 σ 的选取对 SVC、SVR 模型的学习精度和泛化能力起着决定性作用，因此有必要对这两个参数进行优化，使用收敛速度快、调整参数少、全局搜索的粒子群优化(PSO)算法对这两个参数进行寻优[15]。

基于变量排序选择的 SVM 分区建模方法的基本步骤为：

1) 计算导数光谱分离重叠掩蔽的光谱信号，重现待测离子谱峰，并用小波函数对导数光谱信号去噪；

2) MC-UVE 按照 75%的比例随机采集样本 n 个，采样次数为 K($K=100$)次，采样后根据(1)式计算每个波长变量的相关系数-稳定性值 I_j，并按该值从大到小的顺序对波长变量进行排列；

3) 分别按序以前 z($z=1,2,\cdots,J$)个波长点作为变量，将其输入 PSO-SVR 模型；

4) 初始化 PSO-SVR 的粒子群大小、粒子初始速度、位置、最大迭代次数 $t_{\max}=100$；以 CVmse 作为适应度，选取当前适应度最好的粒子作为初始全局最优值；更新粒子速度和位置，计算当前种群粒子的适应度，更新全局最优；直至达到迭代终止条件；

5) 输出 $z(z=1,2,\cdots,J)$个波长变量下模型的CVmse，选取CVmse最小、变量个数最少的波长变量作为SVM模型的输入；

6) 训练PSO-SVC和PSO-SVR模型，以留一交叉验证方均根误差(RMSECV)最小为寻优目标，对模型参数 C 和 σ 进行寻优。

4　实验结果与讨论

按照3.3节所述的基于变量排序选择的SVM分区建模方法的步骤，将48组校正集样本按照75%的比例随机采集样本36个，得到每个波长变量的相关系数，采样100次后得到相关系数-波长矩阵，然后根据(1)式计算每个波长变量的相关系数-稳定性值，结果如图4所示。在图4中，500～541 nm的波长变量具有相当高的相关系数-稳定性值，综合图2可知，此波长段内Cu(II)谱信号波峰明显，且Zn(II)谱信号急剧降低至吸光度0附近，Co(II)谱信号平缓地稳定在吸光度0附近，说明这些波长变量包含的待测离子的有用信息多，且受其他干扰离子的影响小，而且谱信号光滑，含有的噪声少。将所提方法与MC-UVE PLS方法进行比较，结果如图5所示，500～541 nm波长变量的稳定性值 h 高，与图4分析效果相近，但是其在400～420 nm的稳定性值 h 较高，甚至一部分高于Cu(II)谱信号明显的波段，图2中显示400～420 nm波段基体离子Zn(II)有非常大的负光谱信号，且谱线起伏波动大，含有大量噪声，因此MC-UVE PLS在波长变量选择过程中很可能大量选入400～420 nm波段内的波长变量，这会影响模型的效果，不适合用于高浓度基体离子和杂质干扰离子影响下痕量Cu(II)的浓度检测。

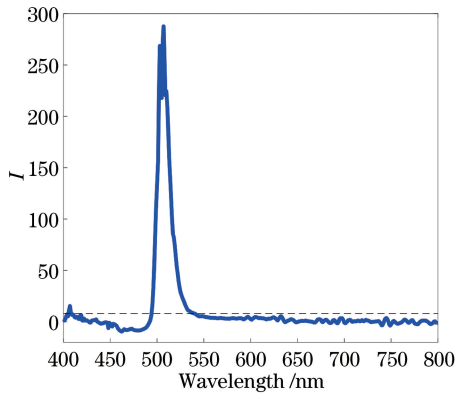

图4　痕量Cu(II)各波长变量的相关系数-稳定性值

Fig. 4　Correlation coefficient-stability value of wavelength variables of trace Cu(II)

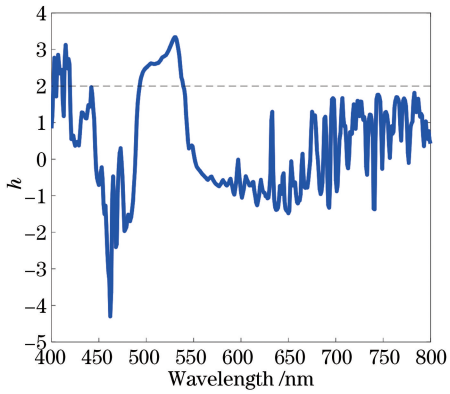

图5　MC-UVE PLS方法下各波长变量的稳定性值

Fig. 5　Stability value of wavelength variables with MC-UVE PLS method

将波长变量按照图4中相关系数-稳定性值从大到小的顺序排列，为尽可能提取Cu(II)的有用信息并减小干扰信息和噪声的影响，优先选取稳定性值大的波长变量；然后按序分别选取1，2，3，…，401个不同数量的波长变量，分别建立PSO-SVR模型，每个PSO-SVR模型以CVmse作为适应度，更新粒子速度和位置，计算种群粒子的适应度并更新全局最优，达到迭代终止条件后，确定模型并输出401种波长变量选择下模型的CVmse。在图6中，变量个数为50时，CVmse值达到最低，之后随着变量个数增加基本保持不变，但当变量个数达到150以后，CVmse急剧增加。在提高模型精度的前提下，为了简化模型复杂度，减少运行时间，选取最佳波长变量个数为50个，所选波长变量为406～408 nm以及495～541 nm。

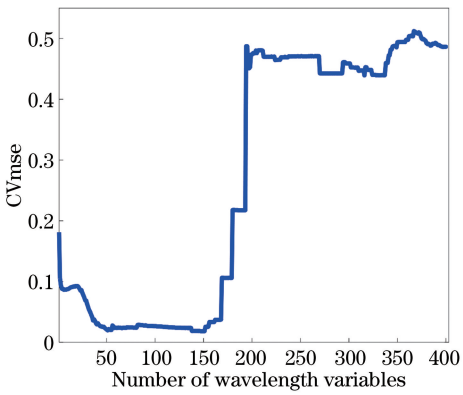

图6　不同波长变量个数下模型的CVmse

Fig. 6　CVmse of models with different number of wavelength variables

确定波长变量后，根据图3的分析讨论将样本分为Cu(II)高浓度区间和低浓度区间，将50个波长变量的光谱信号输入PSO-SVC模型中训练模

型，得到浓度区间预测模型，然后分别针对不同浓度区间建立 PSO-SVR 模型。表 2 为 PSO-SVC、低浓度区间 PSO-SVR 及高浓度区间 PSO-SVR 模型中的惩罚系数 C 和 RBF 核函数参数 σ 的取值。

表 2　SVM 模型参数

Table 2　Parameters of SVM model

Ion	SVM model	Penalty parameter C	Kernel function parameter σ
Cu(II)	PSO-SVC	4.1095	0.10
	PSO-SVR with low concentration interval	8.8250	0.01
	PSO-SVR with high concentration interval	4.1604	0.01

表 3 为 14 组测试样本质量浓度区间预测结果的混淆矩阵，可见：预测正确率为 100%。在实验中如若出现少数预测错误的情况，也不会对整体模型造成较大影响，但会在一定程度上增大该样本的误差，说明该样本的特异性强。与不进行区间预测的模型相比，区间预测模型的误差低一些。

表 3　Cu(II)质量浓度区间预测结果混淆矩阵

Table 3　Confusion matrix of mass concentration interval prediction results for Cu(II)

Item		Predicted interval of mass concentration	
		High	Low
True interval of mass concentration	High	5	0
	Low	0	9

对 14 组测试样本分别采用全波段 PLS、全波段 PSO-SVR、CARS PLS、MC-UVE PLS、MC-UVE LS SVM 以及基于变量排序选择的支持向量机分区回归(VR-S C-SVR)进行建模。此外，为证明分区的效果，采用没有进行浓度分区的基于变量排序选择的支持向量回归(VR-S SVR)的方法进行建模，以波长变量个数、最大相对误差、预测方均根误差(RMSEP)、决定系数 R^2 作为 7 种模型的评价指标，建模方法的比较结果如表 4 所示。从表 4 中可以看出：全波段 PLS 和全波段 PSO-SVR 的变量数目庞大，模型复杂度高，并且由于存在大量噪声、冗余和干扰信息，模型的精度较低；CARS PLS、MC-UVE PLS 和 MC-UVE LS SVM 减少了波长变量个数，模型精度有所提高，但由于挑选了 400～420 nm 波段内的部分波长变量，引入了基体离子带来的干扰信号及大量噪声，导致模型的精度不理想，而且最大相对误差达到了 10%左右，不能满足实验要求和工业现场的需求；VR-S SVR 和 VR-S C-SVR 结合相关系数和 MC-UVE 的优势，选取 Cu(II)谱信号特征明显的波长变量，减小了高浓度 Zn(II)和杂质 Co(II)的干扰，剔除噪声和无用信息，波长变量个数较少，最大相对误差、预测方均根误差 RMSEP 和决定系数 R^2 均可达到要求。VR-S C-SVR 由于先预测质量浓度区间而后分别针对两个浓度区间分开建立模型，体现了混合溶液中Cu(II)谱信号的特性，最大相对误差和预测方均根误差 RMSEP 更低，决定系数 R^2 更高。显然，基于变量排序选择的支持向量机分区回归算法效果更优。

表 4　7 种建模方法的结果

Table 4　Modeling results based on seven methods

Ion	Model	Wavelength number	Maximum relative error /%	R^2/%	RMSEP
Cu(II)	PLS	401	13.78	98.16	0.1424
	PSO-SVR	401	97.53	93.23	0.2835
	CARS PLS	46	10.88	99.06	0.1020
	MC-UVE PLS	67	8.13	98.99	0.1057
	MC -UVE LS SVM	67	14.79	99.18	0.0986
	VR-S SVR	50	8.32	99.47	0.0794
	VR-S C-SVR	50	6.94	99.61	0.0678

图 7 为采用 VR-S C-SVR 模型对锌液中痕量 Cu(II)质量浓度建模后，验证集样本预测值和实际值的散点图。在 14 个样本中，最大相对误差为 6.94%，平均相对误差为 2.74%，相对误差在 0～5%之间的样本有 13 个，相对误差在 5%～10%之间的样本有 1 个。相对其他的线性回归和非线性回归方法而言，基于变量排序选择的支持向量机分区回归(VR-S C-SVR)方法在高浓度 Zn(II)背景和杂质 Co(II)的干扰下，波长数减少了 25.37%～87.53%，模型精度提高 21.10%～76.08%。

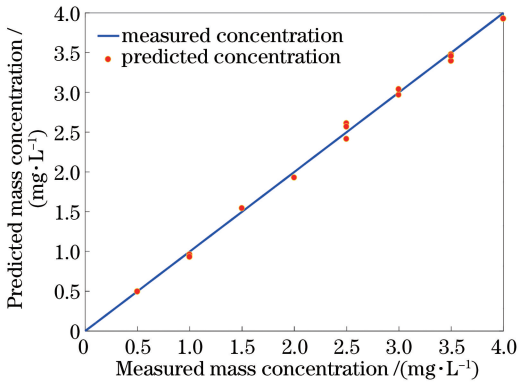

图 7　Cu(II)的质量浓度预测值与实际值的散点图

Fig. 7　Scatter plots of predicted and measured mass concentrations of Cu(II)

5　结　　论

针对在高质量浓度 Zn(II)背景以及杂质离子 Co(II)的干扰下，痕量 Cu(II)浓度难以检测的问题，分析了 Zn(II)、Co(II)、Cu(II)在亚硝基 R 盐化学体系中的紫外可见光谱信号；针对 Cu(II)光谱信号被掩蔽、干扰严重并呈现一定非线性等问题，提出了一种基于变量排序选择的 SVM 分区建模方法。与传统的全波段 PLS 和全波段 SVR 建模方法相比，所提方法有效剔除了噪声大、谱线重叠严重的波长点；与现有基于 MC-UVE 或 CARS 波长变量选择的线性、非线性建模方法相比，该方法通过对波长变量排序选择，提高了波长变量筛选的稳定性和适用性，选取与 Cu(II)相关性大的波长变量，快速有效剔除了 Zn(II)、Co(II)光谱信号掩蔽、重叠、干扰的波段，并根据混合溶液中 Cu(II)在高、低浓度区间呈现不同非线性特性的数据统计现象进行分区建模，增强模型的针对性。实验结果表明，所提方法取得了较好的结果，Cu(II)质量浓度的 RMSEP 降低至 0.0678，最大相对误差降至 7%以下，平均相对误差降至 2.74%，模型决定系数提高至 99.61%，模型波长变量个数有所减少，模型复杂度降低，运行时间缩短，模型精度明显提高，较好地解决了锌液中高浓度 Zn(II)背景以及杂质离子 Co(II)干扰下痕量Cu(II)浓度检测的问题。所提方法也可以应用于荧光光谱、激光诱导击穿光谱、红外吸收光谱等领域中，结合光谱特性进行数据分析并建立定量回归分析模型，解决测试过程中的背景组分交叉干扰问题。

参　考　文　献

[1] Laatikainen K, Lahtinen M, Laatikainen M, *et al*. Copper removal by chelating adsorption in solution purification of hydrometallurgical zinc production[J]. Hydrometallurgy, 2010, 104(1): 14-19.

[2] Wang G W, Yang C H, Zhu H Q, *et al*. State-transition-algorithm-based resolution for overlapping linear sweep voltammetric peaks with high signal ratio[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 151: 61-70.

[3] Xu D, Fan W, Lv H Y, *et al*. Simultaneous determination of traces amounts of cadmium, zinc, and cobalt based on UV-Vis spectrometry combined with wavelength selection and partial least squares regression [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2014, 123: 430-435.

[4] Zhi T X, Shang L P, Deng H, *et al*. Simultaneous determination of mixed systems using principal components regression by fluorescence spectroscopy [J]. Applied Chemical Industry, 2008, 37(10): 1231-1234.
职统兴, 尚丽平, 邓琥, 等. 主成分回归荧光光谱法同时分析多组分混合体系[J]. 应用化工, 2008, 37(10): 1231-1234.

[5] Zhang C H, Yun Y H, Zhang Z M, *et al*. Simultaneous determination of neutral and uronic sugars based on UV-vis spectrometry combined with PLS [J]. International Journal of Biological Macromolecules, 2016, 87: 290-294.

[6] Hu Y, Li Z H, Lü T. Quantitative measurement of iron content in geological standard samples by laser-induced breakdown spectroscopy combined with artificial neural network[J]. Laser & Optoelectronics Progress, 2017, 54(5): 053003.
胡杨, 李子涵, 吕涛. 激光诱导击穿光谱结合人工神经网络测定地质标样中的铁含量[J]. 激光与光电子学进展, 2017, 54(5): 053003.

[7] Liu X, Chen H C, Liu T A, *et al*. Application of PCA-SVR to NIR prediction model for tobacco chemical composition[J]. Spectroscopy and Spectral Analysis, 2007, 27(12): 2460-2463.
刘旭, 陈华才, 刘太昂, 等. PCA-SVR 联用算法在近红外光谱分析烟草成分中的应用[J]. 光谱学与光谱分析, 2007, 27(12): 2460-2463.

[8] Zhu H Q, Gong J, Li Y G, *et al*. A spectrophotometric detecting method of trace cobalt under high concentrated zinc solution [J]. Spectroscopy and Spectral Analysis, 2017, 37(12): 3882-3888.
朱红求, 龚娟, 李勇刚, 等. 一种高锌背景下痕量钴离子浓度分光光度测量法[J]. 光谱学与光谱分析, 2017, 37(12): 3882-3888.

[9] Chen H Z, Pan T, Chen J M, *et al*. Waveband

selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods [J]. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 139-146.

[10] Li C, Zhao T L, Li C, *et al*. Determination of gossypol content in cottonseeds by near infrared spectroscopy based on Monte Carlo uninformative variable elimination and nonlinear calibration methods [J]. Food Chemistry, 2017, 221: 990-996.

[11] Liu S S, Zhang J, Lin S H, *et al*. Quantitative analysis of copper element in pig feed using laser induced breakdown spectroscopy combined with CARS algorithm [J]. Laser & Optoelectronics Progress, 2018, 55(2): 023001.
刘珊珊, 张俊, 林思寒, 等. 激光诱导击穿光谱结合竞争自适应重加权采样算法对猪饲料中铜元素的定量分析[J]. 激光与光电子学进展, 2018, 55(2): 023001.

[12] Zhu H Q, Zou S N, Yang C H, *et al*. Simultaneously measuring method for Zn(II), Co(II) based on feature interval association-partial least squares [J]. Acta Optica Sinica, 2017, 37(6): 0630004.
朱红求, 邹胜男, 阳春华, 等. 基于特征区间联合-偏最小二乘的 Zn(II)、Co(II)同时测量方法[J]. 光学学报, 2017, 37(6): 0630004.

[13] Chen Q S, Zhao J W, Zhang H D, *et al*. Indentification of authenticity of tea with near infrared spectroscopy based on support voctor machine [J]. Acta Optica Sinica, 2006, 26(6): 933-937.
陈全胜, 赵杰文, 张海东, 等. 基于支持向量机的近红外光谱鉴别茶叶的真伪[J]. 光学学报, 2006, 26(6): 933-937.

[14] Wang C L, Liu J G, Zhao N J, *et al*. Quantitative analysis of laser-induced breakdown spectroscopy of heavy metals in water based on support-vector-machine regression[J]. Acta Optica Sinica, 2013, 33(3): 0330002.
王春龙, 刘建国, 赵南京, 等. 基于支持向量机回归的水体重金属激光诱导击穿光谱定量分析研究[J]. 光学学报, 2013, 33(3): 0330002.

[15] Zhang Y Z, Liu Y, Hou H Y, *et al*. Intrinsic tissue fluorescence spectrum recovery based on particle swarm optimization algorithm[J]. Chinese Journal of Lasers, 2016, 43(5): 0504001.
张元志, 刘勇, 侯华毅, 等. 基于粒子群优化算法的生物组织固有荧光光谱复原方法[J]. 中国激光, 2016, 43(5): 0504001.