改进的 FastICA-SVR 结合荧光光谱技术测定 1-萘酚、2-萘酚

王玉田[1]，刘凌妃[1*]，张立娟[1,2]，张正帅[1]，刘婷婷[1]，王书涛[1]，商凤凯[1]

1. 燕山大学河北省测试计量技术及仪器重点实验室，河北 秦皇岛 066004
2. 河北环境工程学院，河北 秦皇岛 066102

摘 要 水作为生命之源与人类的生存息息相关，近年来关于水环境污染的报道越来越多，不容忽视。实验以萘酚的两种同分异构体1-萘酚、2-萘酚的混合物作为研究对象，提出了一种新的算法，通过对混合物的三维荧光光谱进行分析来实现水中萘酚的定性定量分析。利用FS920稳态荧光光谱仪对配制的混合溶液进行扫描得到荧光光谱数据，并对数据进行一系列的预处理去除拉曼散射和瑞利散射的影响。将解决盲源分离(BSS)问题的独立成分分析(ICA)算法应用到荧光光谱定性定量分析问题当中，盲源分离技术就是将测量得到的混合信号作为处理对象进行分解，实现未知系统中源信号的求解，并得到混合矩阵。对混合物中单一物质的识别与测量与盲源分离问题类似。采用基于负熵最大的快速独立成分分析(FastICA)算法对实验数据进行分解，将所有样本的三维荧光光谱数据沿发射波长方向展开成为向量，得到一个大小为($N\times M$)的矩阵(N为样本数，M为波长数)，将该矩阵作为快速独立成分分析的输入进行独立分量提取，输出分别为单组分物质的展开荧光光谱和混合矩阵。FastICA算法的关键是利用牛顿迭代算法得到解混矩阵，但迭代过程中复杂的求导问题会使计算量增大、迭代速度减慢，针对该算法存在的问题，提出用差分法(又称为双点弦截法)代替求导的解决方法。为了验证算法的可行性，用改进后的算法和原有算法分别对荧光光谱数据进行了五次独立分量提取实验，原有算法平均运行时间为17.78 s，而改进后的算法平均运行时间为3.22 s，比原有算法提高了14.56 s，有效地减少了计算量，改善了FastICA算法的迭代速度并且使其收敛性更加稳定。通过实验结果可以看出改进后的算法得到的光谱更接近真实的光谱。利用快速独立成分分析算法分解得到的混合矩阵与物质浓度相关，这是物质定量分析的依据，但它们之间的关系可能是非线性的，采用能实现非线性拟合的支持向量回归机(SVR)进行回归预测，将混合矩阵和实际浓度矩阵分别作为SVR的输入和输出，利用遗传算法(GA)对支持向量回归机的参数进行优化选择，并选择径向基核函数(RBF函数)作为SVR的核函数，建立回归模型，实现对荧光光谱的定量分析。1-萘酚的拟合相关系数(r)为0.998 6，样品回收率(Recovery rate)为96.75%～104.2%，预测均方根误差(RMSEP)为0.119 $\mu g\cdot L^{-1}$；2-萘酚的拟合相关系数为0.998 8，样品回收率为96.8%～105.5%，预测均方根误差为0.1 $\mu g\cdot L^{-1}$，预测结果比较令人满意，符合预测要求。实验证明改进的基于负熵最大的FastICA-SVR算法能实现对混合物中1-萘酚、2-萘酚准确有效的识别和测量，并且改进之后加快了算法的分解速度。

关键词 萘酚；光谱分解；独立成分分析；支持向量回归机；样品回收率

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)01-0142-08

引 言

1-萘酚、2-萘酚作为萘酚的两种同分异构体，是重要的化工原料，但同时萘酚容易被皮肤吸收[1]。随着工业的发展，大量萘酚化合物被排放到环境中，萘酚成为土壤、水体中常见的污染物之一，如果长期饮用含有萘酚的饮用水，可能会导致人体出现头晕、瘙痒、贫血等症状，严重威胁人身健康[2]。因此，研究对水中的萘酚化合物准确、快速且灵敏的检测方法十分重要。

常用的检测萘酚的方法有色谱分析法和光谱法。王英等用Diamonsil$^{TM}C_{18}$色谱柱，成功的实现了人尿中1-萘酚和2-

收稿日期：2017-11-24，**修订日期**：2018-03-19

基金项目：国家自然科学基金项目(61471312，61771419)，河北省自然科学基金项目(F2015203240，F2015203072)资助

作者简介：王玉田，1952年生，燕山大学河北省测试计量技术及仪器重点实验室教授 e-mail：y.t.wang@163.com

*通讯联系人 e-mail：ysuliulingfei@163.com

萘酚的同时检测[3]；叶存玲等采用分散液相微萃取-液相色谱联用技术，实现了自来水、地下水和湖水样品中 1-萘酚和 2-萘酚的分析测定[4]；周纯等基于 1-萘酚、2-萘酚三维荧光光谱的差异，利用荧光光度法同时测定痕量 1-萘酚、2-萘酚；王凡凡等建立了 POSCWPTPLS 程序，用于测定多组分中的 1-萘酚、2-萘酚，效果良好[5]。

三维荧光光谱法具有良好的选择性、信息更加完整，被广泛应用于多组分体系的分析中。目前的研究中多用二阶校正算法来实现多组分三维荧光光谱解析，如平行因子(parallel factor analysis，PARAFAC)，而 PARAFAC 算法存在计算量大，分解时间长等缺陷，而且该算法要求数据严格遵循三线性模型，使其适用范围受到限制。为了使测量结果更加精确，需要不断寻找新的方法来实现混合物的三维荧光光谱分解。

独立成分分析(independent component analysis，ICA)最早应用于盲源分离，能从一组混合观察信号中分离出独立信号，其具有较高的收敛速度。本文采用快速独立成分分析法[6](fast independent component analysis，FastICA)并对该算法进行改进，用差分法代替迭代算法中的求导问题以减少计算量、加快迭代速度，对混合物中的 1-萘酚和 2-萘酚进行定性分析，同时得到两种物质的浓度得分矩阵。常用于分类的支持向量机，逐渐在回归预测中广泛应用[7]。本文利用支持向量回归机[8](support vector regression，SVR)建立回归预测模型，实现多组分系统中 1-萘酚、2-萘酚定量分析。

1　原理知识

1.1　独立成分分析(ICA)算法

ICA 能够将测量得到的混合信号分解成相互独立的源信号，其数学模型表达式可以表示为

$$\boldsymbol{X}=\boldsymbol{AS} \tag{1}$$

将实验中测量得到的 n 个样本的三维荧光光谱按发射波长的方向展开为 n 个行向量，得到矩阵 $\boldsymbol{X}=[x_1, x_2, \cdots, x_n]^{\mathrm{T}}$，$\boldsymbol{S}=[s_1, s_2, \cdots, s_m]^{\mathrm{T}}$ 为 m 个待测量独立成分三维荧光光谱展开组成的光谱阵，$\boldsymbol{A}$ 为混合矩阵，该混合矩阵与样本中各独立成分浓度相关，维数为 $n\times m$。一般的 $m\leqslant n$，$r(\boldsymbol{A})=m$。

ICA 算法首先假设各成分相互独立，在这个基础上，从混合的观测信号 X 中分解出源信号 S 及混合矩阵 $\boldsymbol{A}$，即找解混矩阵 $\boldsymbol{W}$，使得

$$\overline{S}=\boldsymbol{WX} \tag{2}$$

其中，$\overline{S}$ 为计算得到的独立源信号 S 的估计信号。

1.2　改进的 FastICA 算法

FastICA 是一种基于定点递推的独立成分分析算法[9]，其中一种形式是基于负熵最大化。对于独立分量估计值 y，其负熵目标函数为

$$J(y)=[E\{g(y)\}-E\{g(y_{\mathrm{guass}})\}]^2 \tag{3}$$

其中 $g(.)=\tanh(1.5(.))$为非线性函数，y_{guass}和 y 协方差相同，为高斯信号。FastICA 是用牛顿迭代算法来寻找使 $J(\boldsymbol{W})$ 最大时的解混矩阵 $\boldsymbol{W}$。式(4)为对方程 $f(\lambda)=0$ 求解的牛顿迭代公式

$$\lambda_{k+1}=\lambda_k-[f(\lambda_k)/f'(\lambda_k)] \tag{4}$$

根据式(3)和式(4)可知基于负熵的 FastICA 算法的迭代公式为

$$\begin{cases}\boldsymbol{W}_{k+1}=\boldsymbol{E}\{\boldsymbol{X}g(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{Z})\}-\boldsymbol{E}\{g'(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{W}_k\\ \boldsymbol{W}_{k+1}=\boldsymbol{W}_{k+1}/\|\boldsymbol{W}_{k+1}\|\end{cases} \tag{5}$$

式中 $\boldsymbol{W}_k$ 和 $\boldsymbol{W}_{k+1}$分别表示迭代前后的解混矩阵，$\boldsymbol{X}$ 为 n 个样本按发射波长方向展开后组成的光谱矩阵，$\boldsymbol{Z}$ 为 $\boldsymbol{X}$ 白化后的矩阵。

式(5)对 $\boldsymbol{W}$ 迭代求解过程中有求导问题，这里的求导问题比较复杂，而且在每一次迭代中都无法避免，导致计算量增大，迭代速度减慢。为了减少计算量，加快迭代速度，用差分法代替求导，对迭代公式进行改进，如式(6)

$$f'(\lambda_k)\approx\frac{f(\lambda_k)-f(\lambda_{k-1})}{\lambda_k-\lambda_{k-1}} \tag{6}$$

由式(5)和式(6)得到改进的 *FastICA* 的迭代公式为

$$\begin{cases}\boldsymbol{W}_{k+1}=\boldsymbol{E}\{\boldsymbol{X}g(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{W}_{k-1}-\boldsymbol{E}\{g(\boldsymbol{W}_{k-1}^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{W}_k\\ \boldsymbol{W}_{k+1}=\boldsymbol{W}_{k+1}/\|\boldsymbol{W}_{k+1}\|\end{cases} \tag{7}$$

这样就通过差分法避免了求导，使迭代过程简化，减少了计算量，节省了计算时间，图 1 为改进的 *FastICA* 算法流程图。通过迭代运算求得解混矩阵 $\boldsymbol{W}$，将得到的 $\boldsymbol{W}$ 代入式(2)，即让解混矩阵 $\boldsymbol{W}$ 与混合光谱信号 $\boldsymbol{X}$ 相乘就能得到单一组分的光谱信号的估计值，实现对多组分物质的定性分析。

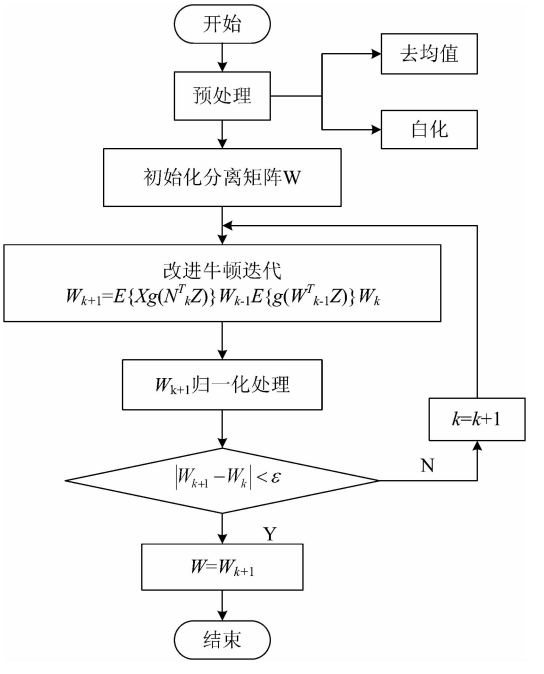

图 1　改进的 FastICA 算法流程图

Fig. 1　Flow chart of Improved FastICA algorithm

2　实验部分

2.1　样品制备及仪器参数

样品配制：(1)称取 1-萘酚、2-萘酚各 0.1 g，用少量 0.1 mol·L^{-1}的 KOH 水溶液溶解，分别倒入 100 mL 的容量瓶

中，并用 KOH 水溶液稀释成浓度为 1 g·L^{-1}的储备液 1；(2)分别量取两种储备液 1 各 0.1 mL 加入到两只 10 mL 的容量瓶中，用超纯水定容，配制成浓度为 10 mg·L^{-1}的两种溶液的储备液 2；(3)分别量取两种储备液 2 各 0.1 mL 于两个 10 mL 的容量瓶中，用超纯水定容成浓度为 100 μg·L^{-1}的标准液；(4)准备 20 个 10 mL 容积的容量瓶，分别加入 1 mL KOH 溶液用来控制溶液的 pH 值，加入不同体积的标准液，用超纯水定容，配制成 20 个混合溶液样本，浓度见表 1；(5)量取体积为 1 mL 的 KOH 水溶液于 10mL 容量瓶中，并用超纯水定容，作为实验的空白溶液。其中样本 1～11 为校正样本，12～20 为预测样本。用 FLS920 荧光光谱仪测量得到所有样本的三维荧光光谱，450 W 的 Xe900 氙灯为激发光源，信噪比 6 000∶1。设置荧光光谱仪 FS920 激发波长为 200～370 nm，发射波长为 320～550 nm，间隔都为 5 nm。对 20 个样品及空白溶液进行扫描，得到三维荧光光谱数据。

表 1 样品真实浓度(μg·L^{-1})

Table 1 Real concentration of samples (μg·L^{-1})

sample	1-naphthol	2-naphthol	sample	1-naphthol	2-naphthol
1	0.50	2.00	11	10.00	8.00
2	1.50	2.00	12	1.00	1.50
3	2.00	1.00	13	2.50	2.50
4	3.00	4.00	14	3.50	3.50
6	6.00	7.00	16	5.00	4.50
7	7.00	5.00	17	5.50	5.00
8	8.50	10.00	18	6.50	4.50
9	9.00	8.50	19	7.50	2.00
10	9.50	7.50	20	8.00	6.00

2.1 数据处理

由于溶剂的干扰会产生散射，实验中测量得到的三维荧光光谱不是真实的荧光谱。其中，瑞利散射的发射波长等于激发波长，进行三维荧光光谱实验时，设置发射波长始终滞后激发波长 20 nm 以消除瑞利散射对荧光的影响[10]；通过混合溶液光谱减去 KOH 水溶液光谱(扣除空白溶液)的方法消除拉曼散射[10]；扫描得到的光谱可能还存在二级瑞利散射，可以通过 Delaunay 三角形内插值法[11]消除。图 2 为样本 1 的三维荧光光谱图，图 2(a)和(c)为实际测量得到的原始光谱，(b)和(d)为消除散射后的校正光谱。

2.2 改进的 FastICA 光谱分解

对光谱进行 ICA 分析前，需要将测量得到的三维荧光光谱按发射波长方向展开成向量，图 3 给出了两种单组分物质的展开光谱图，混合样本以校正后样本 1，3，8 和 15 为例，展开如图 4 所示。

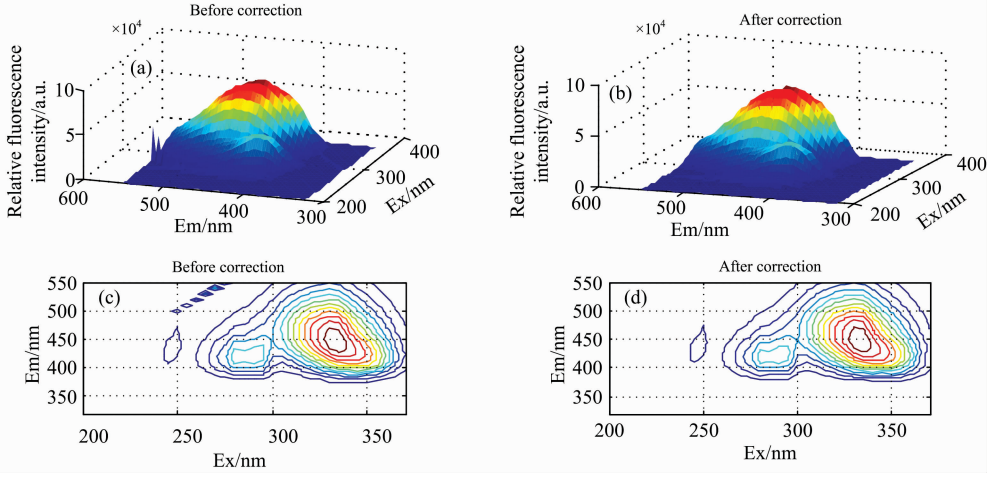

图 2 样本 1 消除散射前后荧光光谱图

Fig. 2 Fluorescence Spectra of sample 1 before and after eliminating scattering

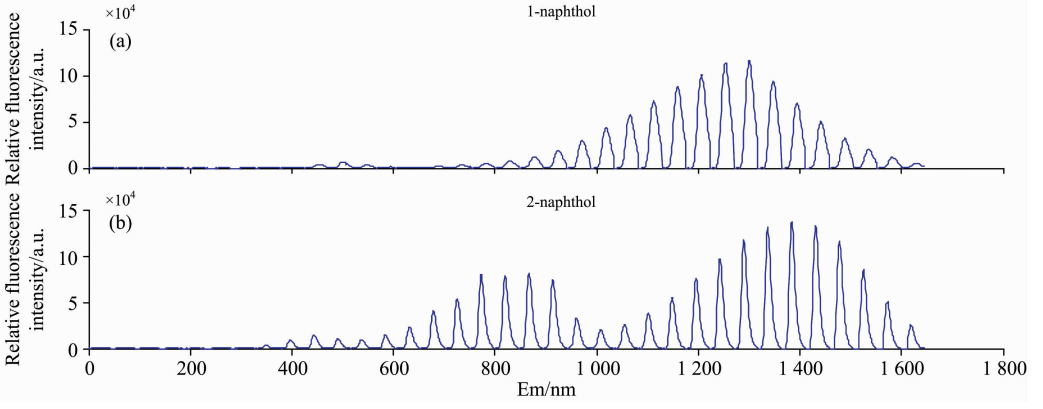

图 3 1-萘酚(a)、2-萘酚(b)沿发射波长展开发射谱

Fig. 3 Emission spectrum spreading along the emission wavelength of 1-naphthol(a) and 2-naphthol (b)

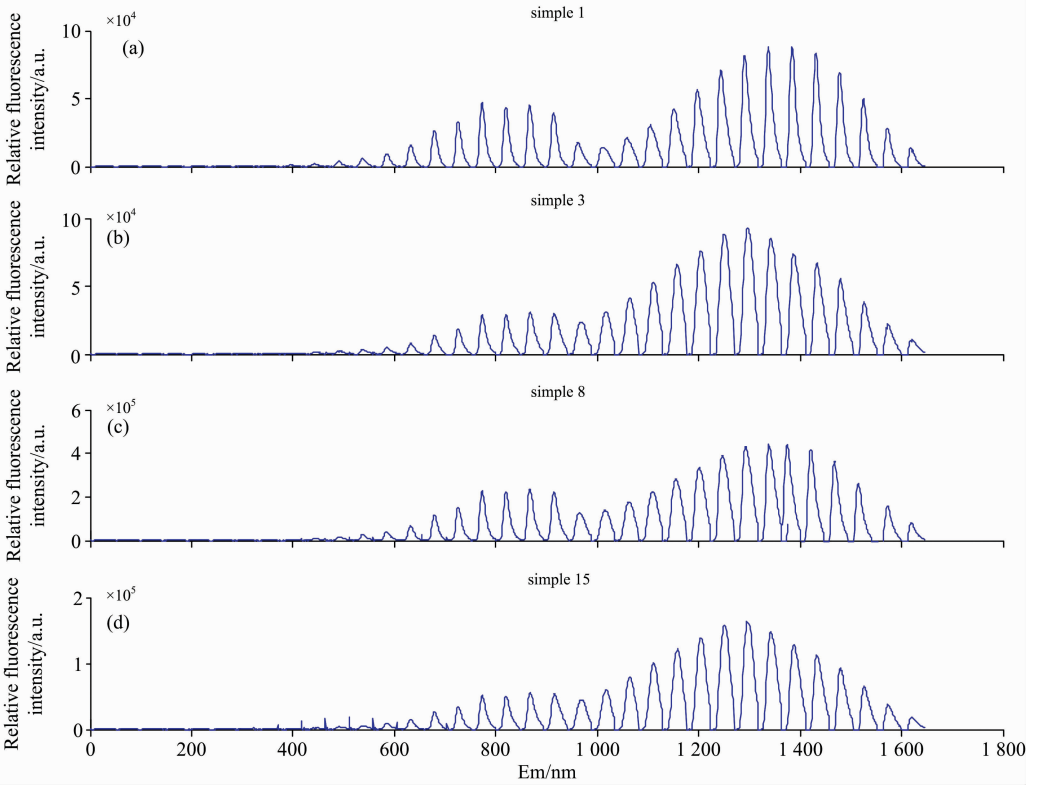

图 4　样本 1(a)，3(b)，8(c)，15(d)沿发射波长展开发射谱

Fig. 4　Emission spectrum spreading along the emission wavelength of sample 1(a), 3(c), 8(c), 15(d)

所有样本展开，并组成矩阵 $\boldsymbol{X}$，大小为(20×1 645)。矩阵 $\boldsymbol{X}$ 作为 ICA 模型的输入进行分离，当独立分量数 ICs=2 时，能量贡献率为 99.97%，ICs=3 时，能量贡献率为 99.98%，与独立分量数为 2 时差别不大，所以选取 ICs=2。为了比较改进的 FastICA 算法和原有的 FastICA 算法的分解性能，分别利用两种算法对混合光谱矩阵 $\boldsymbol{X}$ 进行独立分量提取，结果如图 5 和图 6 所示。由于 ICA 算法中迭代时随机性会导致分解出来的光谱顺序发生变化，识别物质主要依靠波形，从图 5 和图 6 中可以判断出经典 FastICA 算法分解出来的独立分量 1 为 1-萘酚，独立分量 2 为 2-萘酚，改进的 FastICA 算法分解出来的独立分量 1 为 2-萘酚，独立分量 2 为 1-萘酚。两种算法都能把单组分光谱从混合光谱中分离出来，但是改进的算法得到的光谱图更接近原始光谱图。

为了验证改进后的算法是否减少了计算量，加快了迭代速度，分别用两种算法对矩阵 $\boldsymbol{X}$ 进行 5 次分解，记录每次运行用的时间，如表 2 所示。

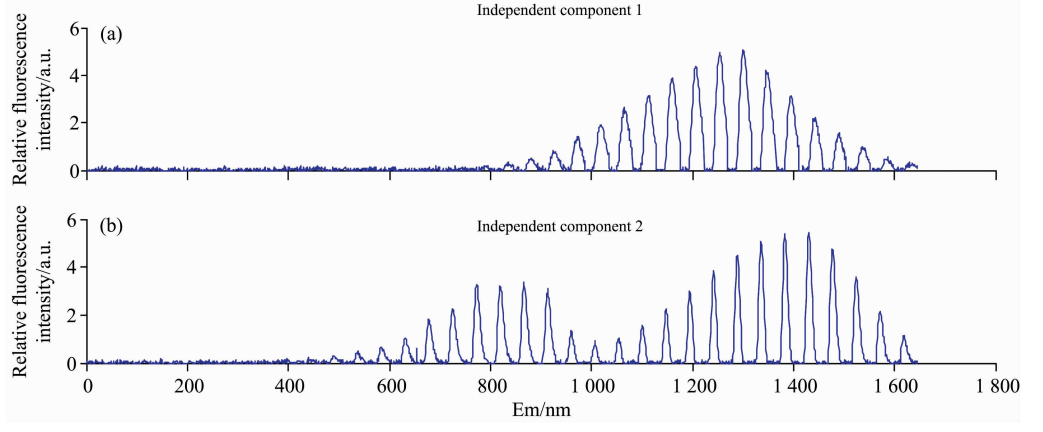

图 5　经典的 FastICA 算法分解光谱图

Fig. 5　Spectrum separated by classic FastICA

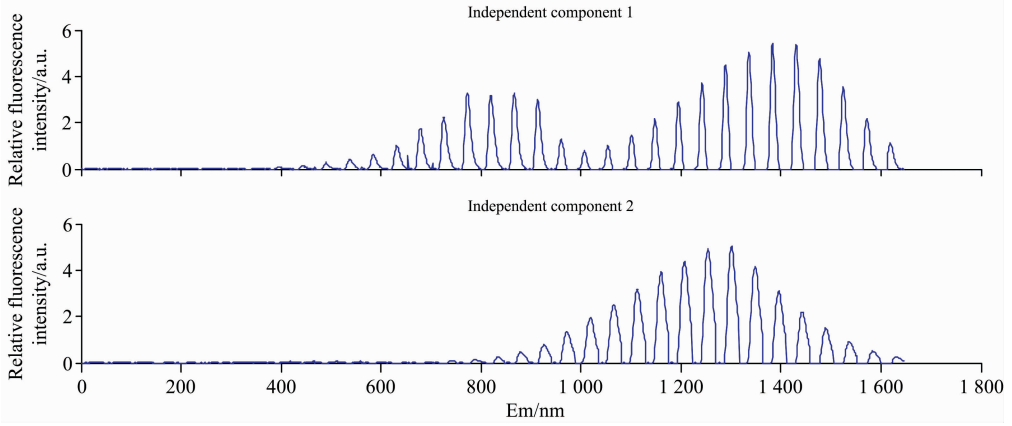

图 6 改进的 FastICA 算法分解光谱图

Fig. 6 Spectrum separated by improved FastICA

表 2 两种算法运行时间对比

Table 2 Running time comparison of two algorithms

Algorithm	Running time/s					Average time/s
	1	2	3	4	5	
FastICA	13.21	20.46	21.33	16.49	17.42	17.78
Improved FastICA	3.22	3.14	3.39	2.99	3.35	3.22

从表 2 中可以看出改进的 FastICA 算法对荧光光谱矩阵 X 进行分解时所用的时间明显少于原有的算法，而且多次实验分解时间波动比较小，比原有的算法稳定。通过两种算法的比较可知对 FastICA 算法中牛顿迭代公式进行改进能够有效的改善原有算法中迭代速度慢的缺陷，使数据处理过程更加快速。

2.3 SVR 浓度回归预测

利用 ICA 将混合信号 X 分解得到各独立成分及混合矩阵 ***A***，根据 ICA 模型可知混合矩阵 ***A*** 与浓度矩阵相关，但它们之间的关系可能是非线性的，支持向量机可以实现非线性回归预测。采用支持向量回归机需要对核函数和参量进行选择。遗传算法(genetic Algorithm，GA)以生物进化为原型，是模仿自然界生物进化机制发展起来的随机全局搜索和优化方法，通过选择、交叉、变异等获得全局最优值，具有很好的收敛性，计算时间少，鲁棒性高[12]。本文采用 RBF 核函数，利用 GA 对 SVR 的参数 c 和 g 进行优化，设置遗传算法的终止代数＝100，种群数量＝20，得到 SVR 最优参数值，然后利用支持向量回归机对预测样本中几种物质的浓度进行测定，选择的参数值及预测性能见表 3，其中 RMSEP 为预测均方根误差，r 为相关系数。

表 3 SVR 的参数值及预测性能

Table 3 Parameter values and prediction performance of SVR

Component	Parameter value		Performance index	
	c	g	RMSEP/(μg·L^{-1})	r
1-naphthol	86.4	0.025	0.119	0.998 6
2-naphthol	93.9	0.011	0.100	0.998 6

3 结果与讨论

利用改进的 FastICA-SVR 算法对 1-萘酚、2-萘酚浓度的预测结果如表 4 所示，并用样品回收率和预测均方根误差(RMSEP)作为性能指标，对预测效果进行评价。两种物质的预测浓度与实际浓度拟合曲线见图 7，其中 1-萘酚的拟合相

表 4 预测样本的预测结果

Table 4 Prediction results for test sample

sample	1-naphthol			2-naphthol		
	Actual /(μg·L^{-1})	Predicted /(μg·L^{-1})	Recovery Rate/%	Actual /(μg·L^{-1})	Predicted /(μg·L^{-1})	Recovery Rate/%
13	2.50	2.57	102.8	2.50	2.42	96.8
14	3.50	3.48	99.4	3.50	3.61	103.1
15	4.00	3.87	96.6	2.00	2.11	105.5
16	5.00	5.21	104.2	4.50	4.38	97.3
17	5.50	5.60	101.8	5.00	5.02	100.4
18	6.50	6.45	99.2	4.50	4.62	102.7
19	7.50	7.37	98.3	2.00	1.94	97.0
20	8.00	7.87	98.4	6.00	6.09	101.5
REMSP/(μg·L^{-1})		0.119			0.100	

关系数 $r=0.998\ 6$，2-萘酚的相关系数 $r=0.998\ 8$。FastICA-SVR 模型对萘酚混合物进行浓度预测，预测精度较高，线性拟合度良好。

为了验证本文提出算法的可行性，表 5 中列出了利用 PARAFAC 算法对混合物的预测结果，并利用相关系数(r)、检出限(LOD)和运行时间(Running Time)作为指标来对两种算法的性能进行对比，如表 6 所示。

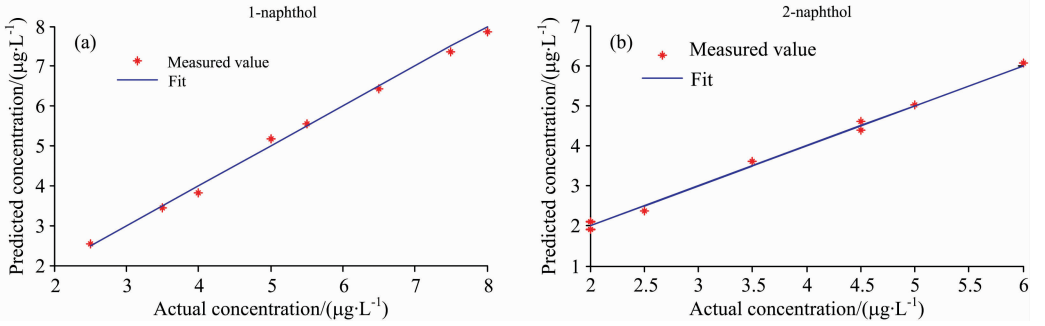

图 7　1-萘酚和 2-萘酚实际浓度和预测浓度拟合曲线

Fig. 7　Fitting curve between actual concentration and predicted concentration of 1-naphthol and 2-naphthol

表 5　PARAFAC 对预测样本的预测结果

Table 5　Prediction results for test samples by PARAFAC

sample	1-naphthol			2-naphthol		
	Actual /(μg·L^{-1})	Predicted /(μg·L^{-1})	Recovery Rate/%	Actual /(μg·L^{-1})	Predicted /(μg·L^{-1})	Recovery Rate/%
13	2.50	2.60	104.0	2.50	2.44	97.6
14	3.50	3.46	98.9	3.50	3.62	103.4
15	4.00	3.83	95.8	2.00	2.14	107.0
16	5.00	4.73	96.6	4.50	4.33	96.2
17	5.50	5.57	101.3	5.00	5.11	102.2
18	6.50	6.39	98.3	4.50	4.47	99.3
19	7.50	7.64	101.9	2.00	1.92	96.0
20	8.00	7.76	97.0	6.00	6.03	100.5
REMSP/(μg·L^{-1})		0.161			0.104	

表 6　两种算法性能指标

Table 6　Performance indicators of two algorithms

Algorithm	Composition	r	LOD/(μg·L^{-1})	Running Time/s
PARAFAC-SVR	1-naphthol	0.996 7	0.071	42.74
	2-naphthol	0.997 2	0.046	
FastICA-SVR	1-naphthol	0.998 6	0.053	3.22
	2-naphthol	0.998 8	0.044	

对比表 4、表 5 和表 6 可以看出两种算法都能对混合物浓度实现良好的预测，本文提出的算法稍好于 PARAFAC，而且 PARAFAC 分解所用的时间为 42.74 s，改进的 FastICA 算法所用时间为 3.22 s，证明本文提出的算法能实现混合物的快速分解，大大提高了分析速率。

4　结　论

将用于信号“盲源分离”的 ICA 算法应用到混合物三维荧光光谱分解当中，并对原有的 FastICA 算法进行改进，用差分法代替迭代过程中的求导问题，算法运行平均时间为 3.22 s，比原来时间减少了 14.56 s，实验证明改进的 FastICA 算法有效地减少了计算量，加快了迭代速度，且分解的独立分量更接近单组分荧光光谱图。ICA 算法得到的混合矩阵与浓度矩阵相关，利用 SVR 对预测样本的浓度进行预测，GA 算法对 SVR 模型参数进行选择。1-萘酚的样品回收率为 96.6%～104.2%，2-萘酚的样品回收率为 96.8%～105.5%。实验证明本文提出的改进的 FastICA-SVR 算法能实现多组分三维荧光光谱的定性定量分析，且能得到良好的预测效果。

References

[1] Sidney J Stohs, Sunny Ohia, Debasis Bagchi. Toxicology, 2002, 180(1): 97.

[2] Zang Shuyan, Lian Bin. Journal of Hazardous Materials, 2009, 166(1): 33.

[3] WANG Ying, WANG Yong-sheng, CAO Xiao-juan, et al(王 英, 王永生, 曹晓娟, 等). Chinese Journal of Health Laboratory Technology(中国卫生检验杂志), 2009, 19(3): 565.

[4] YE Cun-ling, LIU Qing-ling, WANG Zhi-ke(叶存玲, 刘清玲, 王治科). Chinese Journal of Analysis Laboratory(分析实验室), 2010, 29(8): 40.

[5] WANG Fan-fan, REN Shou-xin, MENG He, et al(王凡凡, 任守信, 孟 和, 等). Chinese Journal of Analytical Chemistry(分析化学), 2011, 39(6): 915.

[6] ZHENG Cheng-zhi, GAO Jin-liang, HE Wen-jie(郑成志, 高金良, 何文杰). Journal of Zhejiang University • Engineering Science(浙江大学学报 • 工学版), 2016, 5(50): 977.

[7] GU Yan-ping, ZHAO Wen-jie, WU Zhan-song(顾燕萍, 赵文杰, 吴占松). Journal of Tsinghua University • Science and Technology(清华大学学报 • 自然科学版), 2015, 55(4): 396.

[8] CHEN Jin-dong, PAN Feng(陈进东, 潘 丰). Control and Decision(控制与决策), 2014, 29(3): 460.

[9] WANG Bin, WANG Nian, JIANG Yun-zhi, et al(王 斌, 王 年, 蒋云志, 等). Electric Power Automation Equipment(电力自动化设备), 2011, 31(3): 135.

[10] YANG Li-li, WANG Yu-tian, LU Xin-qiong(杨丽丽, 王玉田, 鲁信琼). Chinese Journal of Laser(中国激光), 2013, 40(6): 0615002-1.

[11] Morteza B, Rasmus B, Colin S. Journal of Chemometries, 2007, 20: 99.

[12] FANG Rui, ZHU Bi-ying, SU Fan-chen(方 睿, 朱碧颖, 粟藩臣). Journal of Computer Applications(计算机应用), 2014, 34(S1): 114.

Determination of 1-Naphthol and 2-Naphthol Based on Fluorescence Spectrometry Combined with Improved FastICA-SVR

WANG Yu-tian[1], LIU Ling-fei[1*], ZHANG Li-juan[1,2], ZHANG Zheng-shuai[1], LIU Ting-ting[1], WANG Shu-tao[1], SHANG Feng-kai[1]

1. Measurement Technology and Instrument Key Lab of Hebei Province, Yanshan University, Qinhuangdao 066004, China
2. Hebei University of Environmental Engineering, Qinhuangdao 066102, China

Abstract As the source of life, water is closely related to the survival of human beings. In recent years, there have been more and more reports on water pollution. Water pollution has become a serious problem, which can not be ignored. Two isomers of naphtol, 1-naphthol and 2-naphthol, were used as the research object in the experiment, and a new algorithm, which was used for qualitative and quantitative analysis of naphthol in water by analyzing the three-dimensional fluorescence spectrum of the mixture, was proposed. Using FS920 steady-state fluorescence spectrometer to scan the mixed solution and get the required experimental data. Then, a series of preprocessing steps for data are needed to remove the effects of Raman scattering and Rayleigh scattering. Independent component analysis (ICA) which is always used to solve the problem of blind source separation (BSS) will be applied to solve the problem in quantitative and qualitative analysis of fluorescence spectrum. BBS is an algorithm that uses the measured mixed signals as the processing objects to realize the decomposion of the source signals in the unknown system, as well as, to get the mixed matrix. The problem in identification and measurement of a single substance in a mixture is similar to the problem in blind source separation. The fast independent component analysis (FastICA) algorithm based on the maximum negative entropy is used to decompose the experimental data. The three-dimensional fluorescence data of all samples need to be expanded into a vector along the direction of the emission wavelength, and a matrix whose size is $N\times M$ can be obtained (N is the number of samples and M is the number of wavelength). This matrix is used as the input of fast independent component analysis to extract independent component, and the output is the expansion fluorescence spectrum of the single component material and a mixed matrix. The key to the fast independent component analysis algorithm is using Newton iterative algorithm to obtain the solution matrix, but the complex derivation of iteration process makes this algorithm have some problems, such as large computation and slow iteration. In order to overcome the shortcomings of fast independent component analysis, the differential method, also called double point chord cut method, is proposed to replace the complex derivation problem in the iterative

process. In order to verify the feasibility of the algorithm, five times independent component extraction experiments were carried out on the spectral data with the improved algorithm and five times independent component extraction experiments were carried out on the spectral data with the original algorithm. The average running time of original FastICA algorithm is 17.78 seconds, and improved FastICA algorithm is 3.22 seconds, which is 14.56 seconds lower than original algorithm. The experiment result proves that differential method instead of the complex derivation problem in the iterative process can effectively reduce the amount of calculation and improve the speed of the iteration of the fast independent component analysis algorithm and the convergence is more stable. It can be seen from the experiment result that the fluorescence spectrum which was obtained by the decomposition are closer to the real spectrum. The mixture matrix obtained by FastICA is related to concentration matrix, which is the basis for quantitative analysis of materials. But the relationship between the mixture matrix and the concentration matrix may be nonlinear. Therefore, it is necessary to take the nonlinear fitting method to realize the fitting between the two. Support vector regression (SVR) machine can realize nonlinear regression, so SVR will be used to obtain predicted concentration. The mixed matrix decomposed and the actual concentration matrix are as the input and output of support vector regression machine respectively. The parameters of SVR are crucial to the prediction. Genetic algorithm (GA) is used to optimize the parameters and radial basis function (RBF function) is selected as the kernel function of SVR. Then the regression model is established by using the algorithm to realize quantitative analysis of the fluorescence spectrum. The fitting correlation coefficient (r) of 1-naphthol is 0.998 6 and 2-naphthol is 0.998 8; the recovery rate of 1-naphthol is 96.6%～104.2% and 2-naphthol is 96.8%～105.5%; the prediction of root mean square error (RMSEP) of 1-naphthol is 0.119 $\mu g \cdot L^{-1}$ and 2-naphthol is 0.100 $\mu g \cdot L^{-1}$. The results of the prediction are satisfactory and meet the requirements of the prediction. The experiment proved that the improved fast independent component analysis algorithm based on negative entropy combined with support vector regression algorithm can accurately identify and measure 1-naphthol and 2-naphthol in mixture, and this algorithm can also increase the speed of analysis for the hybrid system.

Keywords Naphthol; Spectral decomposition; Independent component analysis; Support vector regression; Sample recovery rate

(Received Nov. 24, 2017; accepted Mar. 19, 2018)

* Corresponding author