结合平行因子分析算法和模式识别方法的三维荧光光谱技术用于石油类污染物的检测

孔德明[1,3]，宋乐乐[1]，崔耀耀[2*]，张春祥[1]，王书涛[1]

1. 燕山大学电气工程学院，河北 秦皇岛 066004
2. 燕山大学信息科学与工程学院，河北 秦皇岛 066004
3. Department of Telecommunications and Information Processing，Ghent University，B-9000 Ghent，Belgium

摘 要 随着海洋中石油资源的不断开发，泄漏到海洋环境中的石油也日益增多，它不仅威胁着海洋生态环境，同时也严重影响着人们的身体健康。因此，快速、有效地检测出海洋环境中的石油类污染物对于保护海洋生态环境和人类健康具有重要意义。石油产品中含有大量的多环芳烃，其具有较强的荧光特性。因此，荧光光谱技术成为检测石油类污染物的重要手段之一。利用三维荧光光谱技术结合平行因子分析算法和模式识别方法，对石油类污染物进行表征和分类。首先，以海水和十二烷基硫酸钠(SDS)配制的胶束溶液作为溶剂，分别配制不同浓度的柴油、航空煤油、汽油和润滑油溶液，最终得到80个实验样本；然后，利用FLS920型荧光光谱仪采集实验样本的三维荧光光谱数据，并通过Delaunay三角形内插值法对所获得的三维荧光光谱数据进行去散射处理；其次，利用平行因子分析(PARAFAC)算法分解去散射后的三维荧光光谱数据，通过运用核一致诊断法和残差分析法对组分数进行估计；最后，为了建立稳健的分类模型，利用Kennard-Stone算法将80个实验样本分为60个训练集样本和20个测试集样本，运用K最近邻(KNN)算法、主成分判别分析(PCA-LDA)算法以及偏最小二乘判别分析(PLS-DA)算法分别建立分类模型，并利用灵敏度、特异性和准确率对分类效果进行评估。研究结果表明：三种分类模型对测试集中样本的识别准确率分别为85%，90%和94%，其中，PLS-DA分类模型对测试集样本的识别准确率最高，具有最佳的分类效果。因此，在利用平行因子分析算法提取石油类污染物荧光光谱数据的基础上，结合模式识别方法可以很好的对不同种类油品进行分类研究。利用三维荧光光谱技术结合平行因子分析算法和模式识别方法快速、有效地检测油类污染物，为石油类污染物的快速检测提供了一种新的研究思路和重要参考。

关键词 光谱学；石油类污染物；三维荧光光谱；平行因子分析；模式识别
中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)09-2798-06

引 言

近几十年来，石油产品作为重要的能源及化工原料在现代社会中发挥着不可替代的作用。而随着对能源需求的持续增长，石油产品在开采、使用、运输及储存过程中不可避免地会存在发生泄露的可能性。石油类污染物严重影响附近水域的生态环境，造成附近水域范围内植物、鱼类和浮游生物等生物的大量死亡，间接影响人类的生命健康，而越来越多受到人们的关注[1]。针对石油类污染物的有效检测和识别是处理溢油污染问题的前提基础[2]。因此，研究一种快速、高效的石油类污染物成分识别和分类的检测手段，对于有关部门及时展开应急处理和后续生态环境的治理恢复工作具有重要的现实意义。

目前，针对石油类污染物进行检测的方法主要有红外光谱法、气相色谱法[3]、紫外分光光度法[4]、荧光光谱法[5]等。其中，三维荧光光谱法(excitation-emission matrix，EEM)具有分析速度快、灵敏度高、非破坏性，以及能够表征更多荧光光谱信息等优点，成为一种用于石油类污染物检测的重要手段[6]。程朋飞等[7]利用三维荧光光谱法结合自加权交替三线性分解算法对多种石油类污染物进行了分析，实现了对石

收稿日期：2019-08-06，**修订日期**：2019-12-16
基金项目：国家自然科学基金项目(61501394，61771419)和河北省自然科学基金项目(F2016203155)资助
作者简介：孔德明，1983年生，燕山大学电气工程学院副教授 e-mail：demingkong@ysu.edu.cn
*通讯联系人 e-mail：cuiyaoyao@stumail.ysu.edu.cn

油类污染物的成分识别和浓度预测。杨丽丽等[8]利用三维荧光光谱法结合二阶校正算法对石油类污染物进行了检测，实现了对石油类污染物的定性定量检测。但上述方法存在对噪声容忍能力较弱和收敛速度慢等不足，限制了在实际复杂环境下的应用。借助近年来发展的模式识别方法，在利用平行因子分析(parallel factor analysis，PARAFAC)算法提取石油类物质的荧光特征光谱的基础上，构建稳健的分类模型，解决了石油类物质难以准确识别和分类的问题，具有广阔的应用前景。

分别采集含有海水的四组单一油液的三维荧光光谱数据，利用 Delaunay 三角形内插值法对实验样本的三维荧光光谱数据进行去散射处理，并利用 PARAFAC 算法分解去散射后的三维荧光光谱数据，获得油品的荧光特征光谱，再通过模式识别方法对所提取的荧光特征光谱构建分类模型，从而建立针对石油类污染物的成分表征和油品种类分类的方法。

1　实验部分

1.1　仪器设置与样本配制

实验样本的三维荧光光谱数据由购自英国 Edinburgh Instruments 公司的 FLS920 型荧光光谱仪测得。激发波长的范围设定为 260～500 nm，发射波长的范围设定为 280～520 nm，激发和发射步长均为 5 nm；激发和发射端狭缝宽度设定为 0.44 nm。

选取市场购置的柴油(C)、航空煤油(H)、汽油(Q)和润滑油(R)作为污染物质，采用取自渤海秦皇岛海域的海水作为溶剂来配制实验样本。实验样本的配制步骤如下：(1)取适量海水和十二烷基硫酸钠(SDS)配制 0.1 mol·mL^{-1}的样本溶剂，其目的是为了使油类更充分的溶于海水中；(2)利用精密电子秤称取航空煤油、汽油、柴油和润滑油各 0.1 g，用样本溶剂溶解并分别定容于 10 mL 的容量瓶中，得到 10 mg·mL^{-1}的一级储备溶液并避光保存；(3)分别取 10 mL 的一级储备溶液，用样本溶剂稀释并定容于 10 mL 的容量瓶中，配制成 1 mg·mL^{-1}的标准溶液；(4)分别取不同体积的标准溶液，通过稀释配制成不同浓度的实验样本。

1.2　数据处理方法

1.2.1　平行因子分析算法(PARAFAC)

平行因子分析算法(PARAFAC)是一种基于交替最小二乘原理实现多维数据矩阵分解的算法[9]。实验样本测得的荧光光谱数据组成一个 $I\times J\times K$ 型的三维响应数阵 $\boldsymbol{X}$，其中 K 为样本个数，I 和 J 分别为激发波长和发射波长扫描个数。该算法对三维响应数阵 $\boldsymbol{X}$ 进行分解的过程可由三线性成分模型表示

$$x_{ijk}=\sum_{n=1}^{N}a_{in}b_{jn}c_{kn}+e_{ijk} \tag{1}$$

式中，$i=1,2,\cdots,I$；$j=1,2,\cdots,J$；$k=1,2,\cdots,K$；x_{ijk} 为三维响应数阵 $\boldsymbol{X}$ 中的元素；a_{in} 为相对激发矩阵 $\boldsymbol{A}_{I\times N}$ 中的元素；b_{jn} 为相对发射矩阵 $\boldsymbol{B}_{J\times N}$ 中的元素；c_{kn} 为相对浓度矩阵 $\boldsymbol{C}_{K\times N}$ 中的元素；e_{ijk} 为三维残差矩阵 $\boldsymbol{E}_{I\times J\times K}$ 中的元素；N 为矩阵 $\boldsymbol{A}_{I\times N}$，$\boldsymbol{B}_{J\times N}$ 和 $\boldsymbol{C}_{\mathrm{K}\times\mathrm{N}}$ 的列数，代表所有响应的组分数，包括目标分析物、未知和未校正的干扰物以及变化的背景等。

1.2.2　偏最小二乘判别分析算法(PLS-DA)

偏最小二乘判别分析(partial least square discriminant analysis，PLS-DA)是一种基于偏最小二乘原理的数据分类算法[10]。在实验数据集合中，每个样品有 m 个预测变量 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_m$ 和一个分类变量 $\boldsymbol{Y}$；需将 $\boldsymbol{Y}$ 转换为 q 个潜在变量，即

$$\begin{cases}\boldsymbol{Y}_k=1,\boldsymbol{Y}=k\\ \boldsymbol{Y}_k=0,\boldsymbol{Y}\neq\mathrm{k}\end{cases},\ \mathrm{k}=1,2,\cdots,\mathrm{q} \tag{2}$$

由矩阵 $\boldsymbol{X}_{n\times m}$，$\boldsymbol{Y}_{n\times q}$ 分别代表预测变量和分类变量矩阵。利用 PLSDA 算法对变量矩阵 $\boldsymbol{X}_{n\times m}$ 和 $\boldsymbol{Y}_{n\times q}$ 进行分解，得到正交得分矩阵和载荷矩阵，其实现分解过程的计算公式为

$$\begin{cases}\boldsymbol{X}_{n\times m}=\boldsymbol{T}_{n\times a}\boldsymbol{P}^{\mathrm{T}}_{a\times m}+\boldsymbol{E}_{n\times m}\\ \boldsymbol{Y}_{n\times q}=\boldsymbol{U}_{n\times a}\boldsymbol{Q}^{\mathrm{T}}_{a\times q}+\boldsymbol{F}_{n\times q}\end{cases} \tag{3}$$

式中，$\boldsymbol{T}_{n\times a}$ 和 $\boldsymbol{U}_{n\times q}$ 为隐变量得分矩阵；$\boldsymbol{P}_{a\times m}$ 和 $\boldsymbol{Q}_{a\times q}$ 为载荷矩阵；$\boldsymbol{E}_{n\times m}$ 和 $\boldsymbol{F}_{n\times q}$ 为残差矩阵；a 为特征提取的数目。

2　结果与讨论

2.1　光谱预处理分析

经光谱仪扫描后得到的荧光光谱会存在 Raman 散射和 Rayleigh 散射，如图 1(a)和(b)所示(以汽油样本为例)。散射的存在会导致利用 PARAFAC 算法建立的三线性成分模型带有偏差，严重影响油品的荧光特征分析。从(a)和(b)可以看出，散射的荧光峰过高，掩盖了汽油本身的荧光峰，所以在分析前需要去除散射的干扰。通过 Delaunay 三角形内插值法可以有效地消除散射的干扰。由图 1(c)和(d)可知：经三维荧光光谱数据预处理后，油品的散射得到了有效去除，本身的荧光特征峰得到凸显。

2.2　基于平行因子分析算法的分析结果

采用 PARAFAC 算法分析预处理后得到的 80×49×25 三维数据矩阵 $\boldsymbol{X}$。利用核一致诊断法和残差分析法确定分析时应选取的组分数，结果如图 2(a)和(b)所示。当组分数超过 7 时，核一致值显著降低，残差平方和基本趋于稳定，故选取组分数为 7。运用 7 因子 PARAFAC 模型对 X 进行分析，得到的结果如图 2(c)，(d)和(e)所示。由图 2(c)和(d)可知：因子 1 的激发/发射荧光峰位置为 280/325 nm；因子 2 的激发/发射荧光峰位置为 290/305 nm；因子 3 的激发/发射荧光峰位置为 310/330 nm；因子 4 的激发/发射荧光峰位置为 300/305 nm；因子 5 的激发/发射荧光峰位置为 340/395 nm；因子 6 的激发/发射荧光峰位置为 350/435 nm；因子 7 的激发/发射荧光峰位置为 270/305 nm。由图 2(e)可知：在三维得分图中，几种样品之间出现不同程度的重叠，这说明了仅用 PARAFAC 算法难以将不同石油类油品明显区分开。

2.3　基于模式识别方法的分析结果

为了建立稳健的分类模型，先利用 Kennard-Stone 算法将实验样本划分为训练集和测试集。其中训练集包含 60 个实验样本，测试集包含20个实验样本。为了提高样本利用

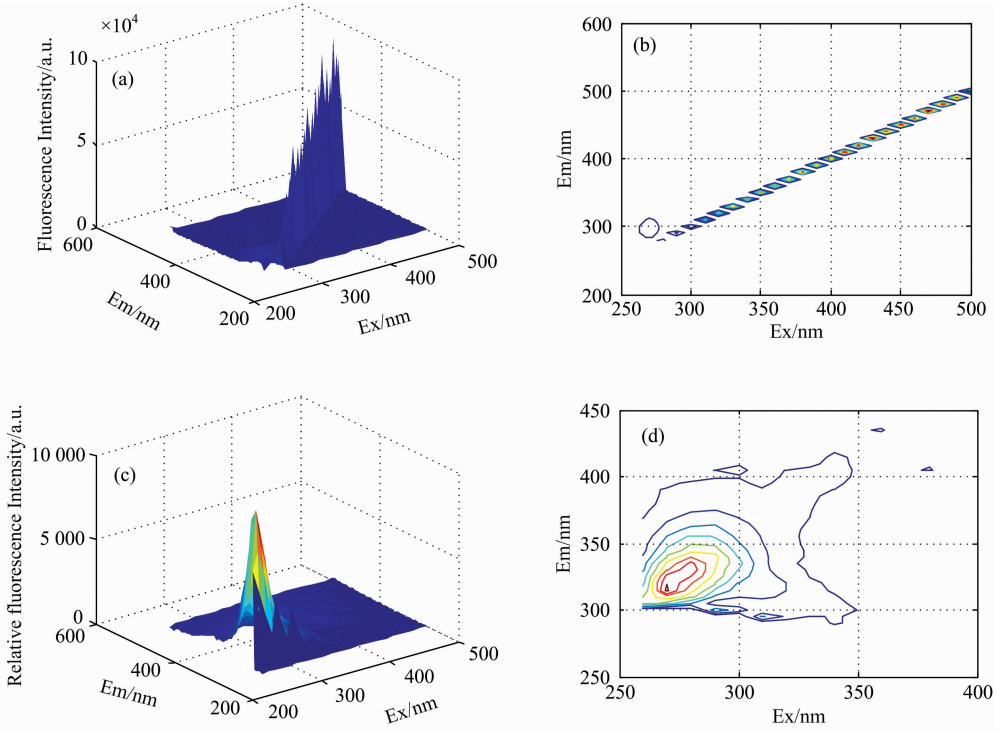
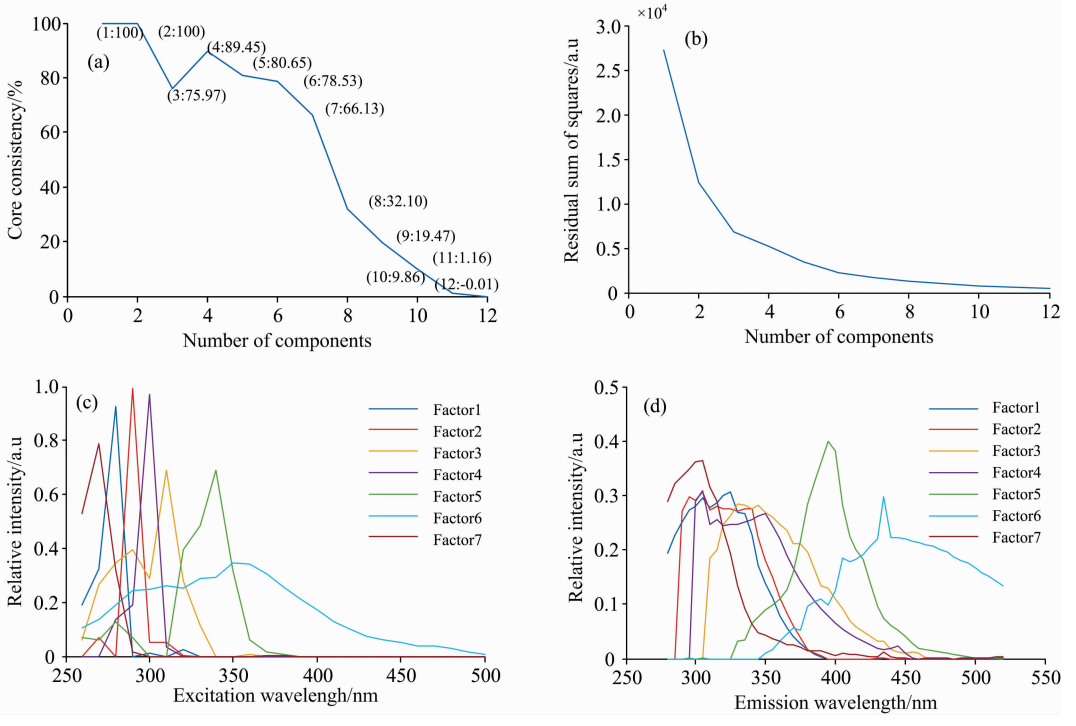


图 1 汽油的荧光光谱

(a)：去散射前的三维荧光光谱；(b)：去散射前的指纹图；(c)：去散射后的三维荧光光谱；(d)：去散射后的指纹图

Fig. 1 Fluorescence spectra of gasoline

(a): Three-dimensional fluorescence spectrum before scattering removal; (b): Fingerprint map before scattering removal; (c): Three-dimensional fluorescence spectrum after scattering removal; (d): Fingerprint map after scattering removal



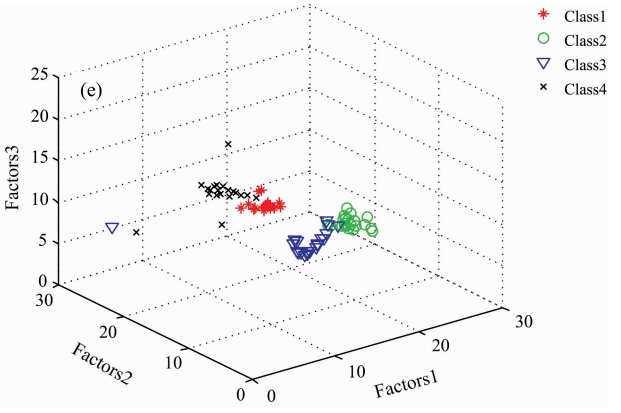


图 2　PARAFAC 算法的分析结果

(a)：核一致值曲线；(b)：残差平方和曲线；(c)：相对激发光谱图；(d)：相对发射光谱图；(e)：因子得分图

Fig. 2　Analysis results by PARAFAC algorithm

(a)：Curves of core consistency value；(b)：Residual sum of squares；(c)：relative excitation spectra；(d)：Relative emission spectra；(e)：Factor score plot

率，得到可靠稳定的模型，在建模前，采用留一法进行交叉验证，并按照使四类油品的校正误差最小的标准选取潜在变量数。灵敏度、特异性和准确率这三个参数能够评估所建立的分类模型的分类效果，KNN，PCA-LDA 和 PLS-DA 分类模型对训练集的识别准确率都可达到 100%，验证了三种模型的稳健性。

利用经验证的分类模型来预测测试集中的 20 个独立样本。分类模型常用混淆矩阵来表示分类结果，由测试集获得的混淆矩阵如表 1 所示。其中黑体数字代表正确预测每类油品的样本个数。根据表 1 中混淆矩阵得出分类模型的灵敏度、特异性和准确率如表 2 所示。由表 1 和表 2 可以看出：这三种分类方法的灵敏度、特异性以及准确率都比较高，说明采用模式识别方法可以很好的对不同种类油品样本进行分类研究。对于 KNN 和 PCA-LDA 模型，识别准确率分别为 85%和 90%，相比而言，采用 PLS-DA 模型取得了更好的分类结果，测试集识别准确率达到了 94%。

表 1　测试集获得的混淆矩阵

Table 1　Confusion matrix from testing set

Models	Predicted											
	KNN				PCA-LDA				PLS-DA			
Actual	Class1	Class2	Class3	Class4	Class1	Class2	Class3	Class4	Class1	Class2	Class3	Class4
Class1	**4**	0	0	0	**4**	0	0	0	**4**	0	0	0
Class2	0	**3**	0	0	0	**3**	0	0	0	**3**	0	0
Class3	1	1	**3**	0	1	1	**3**	0	1	0	**3**	0
Class4	1	0	0	**7**	0	0	0	**8**	0	0	0	**7**

表 2　测试集得到的灵敏度、特异性和准确率

Table 2　Sensitivity, specificity and accuracy obtained from testing set

Models	Sensitivity				Specificity				Accuracy
	Diesel	Jet fuel	Gasoline	Lube	Diesel	Jet fuel	Gasoline	Lube	
KNN	1.00	1.00	0.60	0.88	0.88	0.94	1.00	1.00	0.85
PCALDA	1.00	1.00	0.60	1.00	0.94	0.94	1.00	1.00	0.90
PLSDA	1.00	1.00	0.75	1.00	0.93	1.00	1.00	1.00	0.94

3　结　论

利用三维荧光光谱技术结合平行因子分析算法和模式识别方法对多种石油类污染物进行了组成成分的荧光特性表征和油品种类的分类。研究结果表明，在利用 Delaunay 三角形内插值法去除实验样本中散射的基础上，利用 PARAFAC 算法分解得到的三线性组分模型所构建的 PLS-DA 分类模型较 KNN 和 PCA-LDA 分类模型具有最佳的分类效果，识别准确率最高，达到 94%。本研究提供了一种三维荧光光谱技术与平行因子分析算法和模式识别方法相结合的油品检测方法，可为石油类污染物的快速检测提供一种新的思路和重要参考。

References

[1] LIU Bao-zhan, WEI Wen-pu, DUAN Meng-lan, et al(刘保占, 魏文普, 段梦兰, 等). Marine Environmental Science(海洋环境科学), 2017, 36(1): 15.

[2] LI Yin, LI Guan-nan, CUI Can(李 颖, 李冠男, 崔 璨). Marine Science Bulletin(海洋通报), 2017, 36(3): 241.

[3] YIN Hui-min, DONG Liang, LI Ling-ling, et al(殷惠民, 董 亮, 李玲玲, 等). Environmental Monitoring in China(中国环境监测), 2018, 34(2): 83.

[4] AN Le(安 乐). Marine Environment Science(海洋环境科学), 2017, 36(2): 303.

[5] Yang R J, Dong G M, Sun X S, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 190: 342.

[6] SHEN Hai-dong, BAI Yu-hong, ZHENG Hua(沈海东, 白玉洪, 郑 华). Offshore Oil(海洋石油), 2017, 37(2): 61.

[7] CHENG Peng-fei, WANG Yu-tian, CHEN Zhi-kun, et al(程朋飞, 王玉田, 陈至坤, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(7): 2162.

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

[9] Zhou Z, Guo L, Shiller A M, et al. Marine Chemistry, 2013, 148: 10.

[10] Lenhardt L, Bro R, Zekovic I, et al. Food Chemistry, 2015, 175: 284.

Three-Dimensional Fluorescence Spectroscopy Coupled With Parallel Factor and Pattern Recognition Algorithm for Characterization and Classification of Petroleum Pollutants

KONG De-ming[1,3], SONG le-le[1], CUI Yao-yao[2*], ZHANG Chun-xiang[1], WANG Shu-tao[1]

1. School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China

2. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China

3. Department of Telecommunications and Information Processing, Ghent University, B-9000 Ghent, Belgium

Abstract With the continuous development of petroleum resources in the ocean, more and more petroleum is leaking into the marine environment. It not only threatens the marine ecological environment but also seriously affects people's health. Therefore, the rapid and effective detection of petroleum pollutants in the marine environment is of great significance for the protection of the marine ecological environment and human health. Petroleum products contain a large number of polycyclic aromatic hydrocarbons, which have strong fluorescence characteristics. Therefore, fluorescence spectroscopy technology has become one of the important means to detect petroleum pollutants. In this paper, three-dimensional fluorescence spectroscopy combined with parallel factor analysis algorithm and pattern recognition method is used to characterize and classify petroleum pollutants. Firstly, the micelle solution prepared by seawater and sodium dodecyl sulfate (SDS) was used as a solvent to prepare different concentrations of diesel,jet fuel, gasolineand lube solutions, and 80 experimental samples were finally obtained. Then, three-dimensional fluorescence spectra of experimental samples were collected by FLS920 fluorescence spectrometer, and the effect of scattering was removed by using the Delaunay triangle interpolation method. Secondly, the paralleled factor analysis (PARAFAC) algorithm is used to decompose the three-dimensional fluorescence spectrum data after scattering, and the component number is estimated by using the nuclear consistency diagnosis method and residual analysis method. Finally, in order to establish a robust classification model, 80 experimental samples were divided into 60 training set samples, and 20 test set samples by Kennard-Stone algorithm. The K-nearest neighbor (KNN) algorithm, principal component discriminant analysis (PCA-LDA) algorithm and partial least squares discriminant analysis (PLS-DA) algorithm are used to establish the classification model respectively, and sensitivity, specificity and accuracy are used to evaluate the classification effect. The results show that the recognition accuracy of the three classification models is 85%, 90% and 94% respectively. The PLS-DA classification model has the highest recognition accuracy and the best classification effect. Therefore, based on extracting the fluorescence spectrum data of petroleum pollutants by using parallel factor analysis algorithm and combining with the pattern recognition method, the classifi-

cation of different kinds of oil products can be well studied. In this paper, three-dimensional fluorescence spectroscopy combined with parallel factor analysis algorithm and pattern recognition method is used to detect petroleum pollutants quickly and effectively, which provides a new research idea and an important reference for the rapid detection of petroleum pollutants.

Keywords Spectroscopy; Petroleum pollutants; Three-dimensional fluorescence spectrum; PARAFAC; Pattern recognition

* Corresponding author

(Received Aug. 6, 2019; accepted Dec. 16, 2019)

第 21 届全国分子光谱学学术会议暨 2020 年光谱年会
(第二轮通知)

由中国光学学会和中国化学会主办的“第 21 届全国分子光谱学学术会议”暨由中国光学会光谱专业委员会主办的“2020 年光谱年会”将于 2020 年 10 月 30—11 月 2 日在成都召开，会议由四川大学分析测试中心承办。本次大会将秉承前 20 届分子光谱学学术会议之宗旨，以期形成自由研讨的学术氛围，让光谱相关或相近的思想撞击出火花，期待颠覆性创新创造力泉涌。

一、会议简要日程安排

2020 年 10 月 30 日

全天注册报到

16:00—18:00　组织委员会和学术委员会会议；《光谱学与光谱分析》编委会会议

2020 年 10 月 31 日

08:30—12:00　开幕式、大会报告

14:00—18:00　大会报告

2020 年 11 月 1 日

08:30—12:00　分组邀请报告和口头报告

14:00—18:00　分组邀请报告和口头报告

2020 年 11 月 2 日

08:00—12:00 大会报告及闭幕式

二、学术报告

本次会议将采用邀请报告和申请口头报告相结合的形式，同时也将开设青年论坛和墙报展示。组委会对青年学者、博士和硕士研究生等设立优秀论文奖(包括优秀口头报告和墙报)，届时将组织专家进行评选。

2.1　邀请报告

已经确认参加会议并作大会报告的院士及国内外著名学者：

李　灿 院士　中国科学院大连化学物理研究所

陈洪渊 院士　南京大学

田中群 院士　厦门大学

孙世刚 院士　厦门大学

谭蔚泓 院士　湖南大学

张　锦 院士　北京大学

邀请报告信息将陆续更新，请大家关注会议主页浏览相关信息：

http://www.sinospectroscopy.org.cn/meeting/index.php? mid=24

2.2　口头报告

会议将开放一定数量的口头报告，大家可以自由申请，申请方式为在会议注册系统中提交口头报告题目(在口头报告栏目)，并在会议截稿日期前通过会议稿件提交系统按要求提交论文摘要，申请截止日期为 2020 年 6 月 30 日。

2.3　青年论坛

对于青年学者，博士和硕士研究生可以申请青年论坛报告，申请办法和截止日期与口头报告相同，组委会将组织专家进行优秀报告评选，并颁发优秀论文证书和奖金。

(下转 2808 页)