基于重构三维荧光光谱结合偏最小二乘判别分析的油类识别方法研究

崔耀耀[1]，孔德明[2,3*]，孔令富[1]，王书涛[2]，史慧超[4]

1. 燕山大学信息科学与工程学院，河北 秦皇岛 066004
2. 燕山大学电气工程学院，河北 秦皇岛 066004
3. Department of Telecommunications and Information Processing, Ghent University, B-9000 Ghent, Belgium
4. 北京化工大学信息科学与技术学院，北京 100029

摘 要 油类污染日渐频繁，给人类健康及生态环境造成了严重的威胁。因此，研究有效的油类识别方法对保护生态环境具有重要意义。三维荧光光谱技术是识别油类最有效的分析手段之一，利用二阶校正方法对三维荧光光谱数据进行解析，然后利用模式识别对二阶校正方法解析结果中的浓度得分矩阵进行分类，可以实现对未知样本的定性识别。然而，此类方法在对未知样本进行分类识别的过程中，只应用了浓度得分矩阵，其本质上只是利用样本所含化学成分的相对含量差异对未知样本进行了分类。并没有利用具有定性意义的载荷矩阵，即没有从样本所含化学成分本身实现对样本的定性。基于此，将重构的三维荧光光谱和偏最小二乘判别分析(PLS-DA)相结合，提出了一种针对油类样本的辨识方法。首先，利用四种油类(汽油、柴油、航空煤油和润滑油)在不同的背景环境下(纯净水、自来水、河水及海水配制的十二烷基硫酸钠溶剂)配制了80个油类样本；然后，利用FS920荧光光谱仪采集样本的三维荧光光谱数据，并对该数据进行去散射及标准化预处理；其次，利用Leverage值识别并删除其中的异常光谱，并利用平行因子分析算法(PARAFAC)对剩余的光谱进行重构；最后，通过PLS-DA建立重构三维荧光光谱的分类模型；并将重构与未重构的三维荧光光谱分别建立的分类模型进行了对比。分析结果表明，三维荧光光谱经过重构后，可以将四种油类的正确分类率分别从原来的100%，50%，60%和20%提高到100%，100%，100%和100%，表明重构的三维荧光光谱具有更加明显的类内特征。重构三维荧光光谱所建立的分类模型的灵敏度(SENS)、特异性(SPEC)及F分数分别为100%，100%和100%，表明所建立的模型具有稳健及可靠的分析结果。该研究中，重构三维荧光光谱利用了PARAFAC解析结果中的浓度得分矩阵及载荷矩阵，所建立的PLS-DA分类模型不仅从化学成分相对含量的差异而且从化学成分本身对样本进行了定性识别，所得结果更加具有说服力。该研究为油类识别提供了一种可靠的方法。

关键词 重构三维荧光光谱；PARAFAC；PLS-DA；油类识别

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)12-3789-06

引 言

石油产品作为最重要的能源及化工原料在现代社会中发挥着举足轻重的作用[1]。石油产品在开采、使用、运输及储存等过程中不可避免会发生泄漏，从而导致严重的生态环境污染[2]，对人类健康以及社会经济造成不可估量的影响[3]。因此，研究有效的油类识别方法对于相关部门进行应急处理以及保护生态环境具有重要的实用价值。

目前，三维荧光光谱法是鉴别复杂污染背景环境中油类最有效的方法之一[4-5]。通常使用平行因子分析(PARAFAC)[6]、交替三线性分解(ATLD)[7]等二阶校正方法解析三维荧光光谱数据(EEM)，从而获得具有化学意义的得分矩阵(代表被解析样本中所含化学成分的相对含量)以及载荷矩阵(代表被解析样本中所含化学成分本身的光谱特性)。然后使用判别分析(DA)、支持向量机(SVM)等[8]模式识别方法对二阶校正方法获得的浓度得分矩阵进行分类。从而实现对未知样本识别的目的。

收稿日期：2019-06-05，**修订日期**：2019-10-10

基金项目：国家自然科学基金项目(61501394，61771419)，河北省自然科学基金项目(F2016203155)资助

作者简介：崔耀耀，1992年生，燕山大学信息科学与工程学院博士研究生 e-mail：cuiyaoyao@stumail.ysu.edu.cn

*通讯联系人 e-mail：demingkong@ysu.edu.cn

然而，上述方法在建立分类模型的过程中，只是应用得分矩阵从样本所含化学成分的相对含量上对其进行识别，并没有利用具有定性信息的载荷矩阵从样本的化学成分本身对其进行定性。基于此，本文采集了四种油类在不同背景环境下配制的 80 个油类样本的三维荧光光谱数据。然后利用 PARAFAC 对三维荧光光谱数据进行了重构，以消除仪器误差、噪声等所带来的干扰。最后通过偏最小二乘判别分析(PLS-DA)建立样本的分类模型，从而建立了一种识别未知油类的新方法。

1 实验部分

1.1 材料与仪器

取汽油(Q)、柴油(C)、航空煤油(H)和润滑油(R)四种油类，按照表 1 中的浓度配制实验样本。具体步骤如下：(1)用纯净水溶解适量的十二烷基硫酸钠(SDS)得到浓度为 0.1 mol·L^{-1}的 SDS 溶剂，置于棕色玻璃瓶中避光保存；(2)利用精密电子秤分别称取上述油类各 0.1 g，用 SDS 溶剂分别定容于四个 10 mL 的容量瓶中，得到浓度为 10 mg·mL^{-1}的一级储备溶液；(3)用移液器吸取适量的一级储备液，经 SDS 溶剂稀释后，配制表 1 中的 20 个实验样本；(4)分别利用自来水、河水以及海水配制另外 3 种浓度为 0.1 mol·L^{-1}的 SDS 溶剂，并利用该溶剂重复步骤(2)和步骤(3)，最终得到不同溶剂背景下的 80 个油类实验样本。

表 1 油类样本浓度

Table 1 Oil samples concentration

Sample	Concentration/(mg·mL^{-1})				
Q	0.1	0.2	0.5	1	2
C	0.1	0.2	0.5	1	2
H	0.1	0.2	0.5	1	2
R	0.1	0.2	0.5	1	2

使用英国 Edinburgh Instruments 公司生产的 FS920 稳态荧光光谱仪采集实验样本的荧光光谱。设置激发和发射端的狭缝宽度为 0.44 mm；设置激发波长范围为 260∶10∶500 nm，发射波长范围为 280∶5∶520 nm。

1.2 光谱数据预处理

实验样本所获得的原始荧光光谱如图 1 所示(汽油样本)。原始荧光光谱中通常含有 Rayleigh 和 Raman 散射光谱，这些散射光谱不包含样本中荧光团的任何信息。而且所有样本中的散射光谱所处区域及其光谱形状一致，这会对后期正确分类实验样本带来极大干扰。因此，必须将散射光谱去除，去除散射后的光谱如图 2 所示。对去除散射后的荧光光谱数据进行标准化处理，结果如图 3 所示。

1.3 数据处理

1.3.1 三维荧光光谱重构

本文利用平行因子分析(PARAFAC)对三维荧光光谱数据进行重构，以消除仪器误差、噪声等所带来的干扰。PARAFAC 可以将三维数据($I\times J\times K$)分解为一个得分矩阵 $\boldsymbol{A}(I\times N)$和两个载荷矩阵 $\boldsymbol{B}(J\times N)$，$\boldsymbol{C}(K\times N)$以及一个残差矩阵 $\boldsymbol{E}(I\times J\times K)$

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

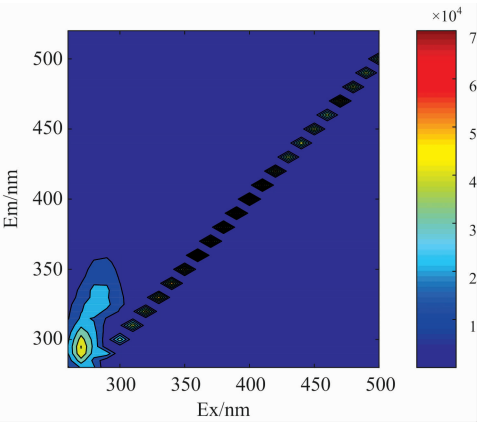

图 1 汽油样本的原始荧光光谱图

Fig. 1 Original fluorescence spectrum of a gasoline sample

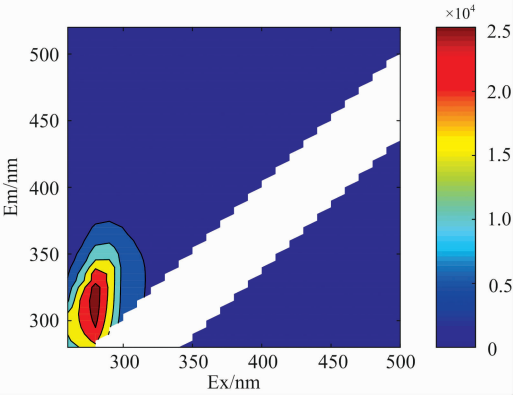

图 2 去除散射后的汽油荧光光谱图

Fig. 2 Fluorescence spectrum of gasoline removal scattering

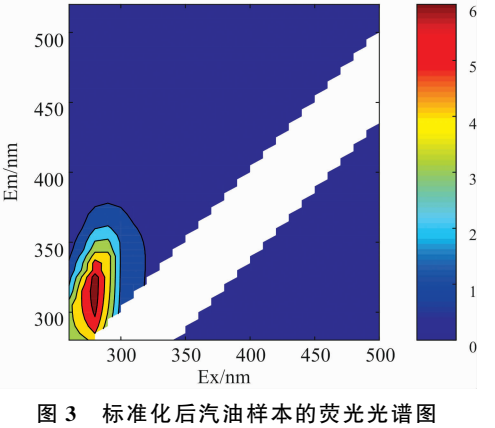

图 3 标准化后汽油样本的荧光光谱图

Fig. 3 Fluorescence spectrum of normalized gasoline samples

式(1)中，$i=1,2,3,\cdots,I$，I 为样本数量；$j=1,2,3,\cdots,J$，J 为发射波长数量；$k=1,2,3,\cdots,K$，K 为激发波长数量；$n=1,2,3,\cdots,N$，N 为 *PARAFAC* 建模时的组件数量；x_{ijk} 表示第 i 个样本在激发波长为 k、发射波长为 j 时的荧光强度值；a_{in} 是得分矩阵 $\boldsymbol{A}(I\times N)$ 中的元素；b_{jn} 是发射矩阵 $\boldsymbol{B}(J\times N)$ 中的元素；c_{kn} 是激发矩阵 $\boldsymbol{C}(K\times N)$ 中的元素；e_{ijk} 是三维残差矩阵 $\boldsymbol{E}(I\times J\times K)$ 中的元素。

其中，每一个 n 值都对应一个 PARAFAC 组件。这些组件在有效模型中具有直接的化学成分解释，e_{ijk} 表示模型未考虑的可变性残差，主要代表了荧光光谱中不可解释的成分(如仪器误差、噪声等)。在光谱重构过程中，若去除残差项 e_{ijk}，则可以得到直接反映样本化学成分的稳健性三维荧光光谱。光谱重构公式如式(2)

$$x_{ijk}^{\text{new}}=\sum_{n=1}^{N}a_{in}b_{jn}c_{kn} \tag{2}$$

式(2)中，x_{ijk}^{new} 表示第 i 个样本在激发波长为 k、发射波长为 j 时重构三维荧光光谱的强度值。

1.3.2　偏最小二乘判别分析

偏最小二乘判别分析(PLS-DA)是一种基于 PLS2 的分类方法[9]。它将 PLS 的回归结果转换为一组可用于预测因变量的中间线性潜在变量(组件)。因变量即是给定的类标签，它用于指示给定样本是否属于给定类。利用上述原理构建的模型可用于预测新样本所属的类[10]。

1.3.3　评价指标

使用的评价指标包括：正确分类率($CC\%$)、准确度($AC\%$)、灵敏度($SENS\%$)、特异性($SPEC\%$)和 F 分数[11]。$CC\%$表示正确分类为正数的样本数；AC 代表考虑到真假阴性的正确分类的样本总数；$SENS$ 衡量正确识别的阳性比例；$SPEC$ 衡量正确识别的阴性比例；F 分数衡量模型的性能[12]。计算公式如式(3)—式(7)所示

$$CC(\%)=100-\frac{(\varepsilon_1+\varepsilon_2)}{N}\times 100 \tag{3}$$

$$AC(\%)=\left(\frac{TP+TN}{TP+FP+TN+FN}\right)\times 100 \tag{4}$$

$$SENS(\%)=\left(\frac{TP}{TP+FP}\right)\times 100 \tag{5}$$

$$SPEC(\%)=\left(\frac{TN}{TN+FP}\right)\times 100 \tag{6}$$

$$F\text{-score}=\frac{2\times SEND\times SPEC}{SEND+SPEC} \tag{7}$$

其中，TP 代表真阳性，TN 代表真阴性，FP 代表假阳性，FN 代表假阴性；N 是测试集中的样本数；ε_1 和 ε_2 表示第 1 类和第 2 类测试集中的错误分类的样本数量。

2　结果与讨论

2.1　三维荧光光谱重构

在光谱测量过程中，由于受到环境因素以及人为误差的影响，导致所获得的光谱数据中可能存在不能真实反映油类荧光团信息的异常光谱。这些可能存在的异常光谱会使重构的光谱出现位置的偏移甚至形状的改变。因此，在三维荧光光谱重构之前首先需要检测可能存在的异常光谱并将其删除。

通过实验样本的 Leverage 值识别异常光谱，Leverage 值越大则其为异常光谱的可能性就越大。20 个汽油样本的 Leverage 值如图 4 所示。图中 19 个样本的 Leverage 值基本一致，而第 13 个样本的 Leverage 值远大于其他样本，因此可将其判断为光谱存在异常的样品。用同样的方法检测出柴油中的第 1 和第 3 个样本、航空煤油中的第 1 个样本以及润滑油中的第 1 和第 17 个样本为光谱存在异常的样本。

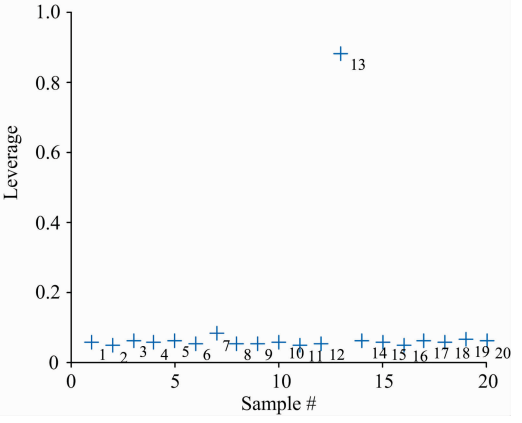

图 4　异常样品的识别

Fig. 4　Identification of abnormal samples

然后，利用激发和发射光谱的残差来确定平行因子建模时所需的组件数量。汽油样本组件残差图如图 5 所示。其中，组件数为 2 时的激发和发射光谱的残差最大，随着组件数量增加，激发和发射光谱残差显著降低，当组件数为 5，6 和 7 时，残差基本一致，变化不再明显。为加快建模速度，本文选用 5 组件对三维荧光光谱进行重构。汽油样本三维荧光光谱、重构三维荧光光谱及残差分布如图 6 所示。

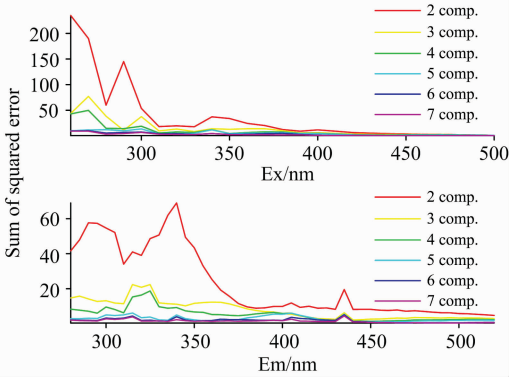

图 5　组分残差图

Fig. 5　Residual figure of components

2.2　基于 PLS-DA 的油类样本识别

首先，利用 Kennard-Stone 采样选择算法将剩余的 74 个样本划分为校正样本($n=60$)和预测样本($n=14$)，然后利用 PLS-DA 对校正样本进行建模。在建立 PLS-DA 校准模型之前，利用交叉验证选择潜在变量(LVs)的数量，交叉验证将

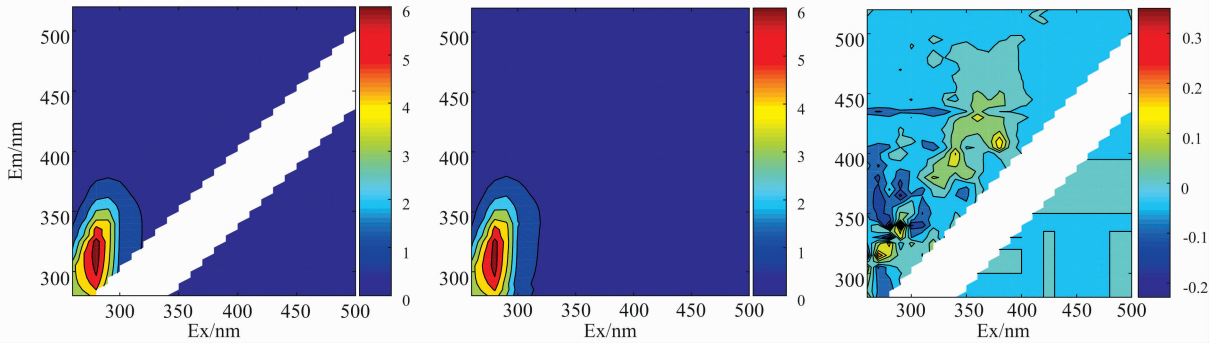

图 6 汽油样本的三维荧光光谱、重构后的三维荧光光谱以及残差分布图

Fig. 6 3D fluorescence spectrum, reconstruction 3D fluorescence spectrum and residual distribution of gasoline samples

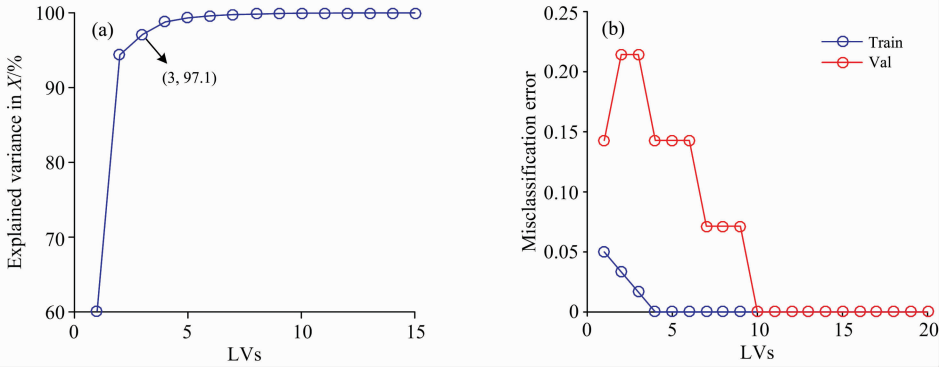

图 7 潜在变量的选择

Fig. 7 Selection of LVs

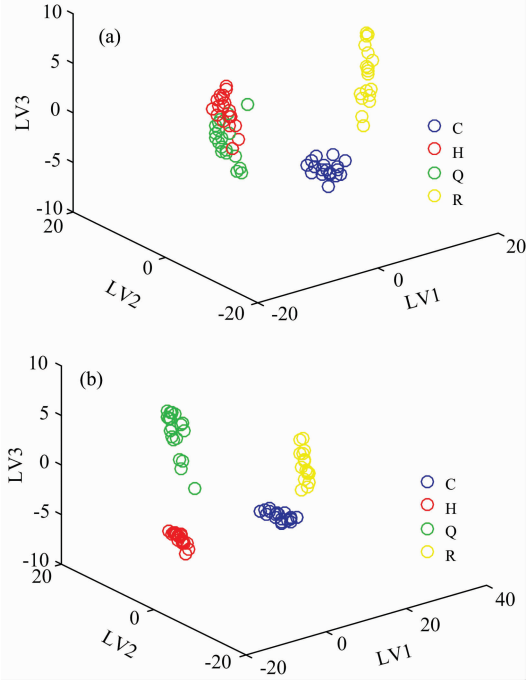

图 8 油类样本的前 3 个 LVs 得分图

Fig. 8 The first 3 LVs scores of oil samples

校正样本数据划分为训练组和测试组，并根据训练组和测试组的解释方差及错误分类率选取 LVs 的数量，如图 7 所示。由图可知，当选取 LVs＝10 时，解释方差[图 7(a)]为 100%，错误分类率[图 7(b)]为 0。其中，数据 97.1%的变化可由前 3 个 LVs 解释[图 7(a)]，观察油类样本的前 3 个 LVs 得分图(图 8)，图 8(a)为未经重构的三维荧光光谱前 3 个 LVs 的得分，图中的航空煤油和汽油相互重叠，难以区分两种油类。而经过重构的三维荧光光谱前 3 个 LVs 的得分[图 8(b)]则将航空煤油和汽油完全分离，并且与图 8(a)中四种油类 LVs 得分相比，经过重构的三维荧光光谱得分可以更加密集的将同种油类聚集在一起。表明经过重构的三维荧光光谱能够更加准确的反映同种油类间的特征。

利用训练好的校正模型对预测样本进行预测，得到最终结果如图 9 所示。图 9(a)是油类样本未经重构的三维荧光光谱的 PLS-DA 建模及分类结果。其中，四种油类都出现分类错误的情况，分类效果较差。图 9(b)是油类样本重构三维荧光光谱的 PLS-DA 建模及分类结果，四种油类建模及分类均完全正确，分类效果理想。表 2 列出了模型的具体评价结果，从表中可以看出，重构三维荧光光谱获得各项评价指标值均优于未重构的三维荧光光谱。该结果表明油类样本的三维荧光光谱经重构后再用于分类，可以获得更好的分类性能。

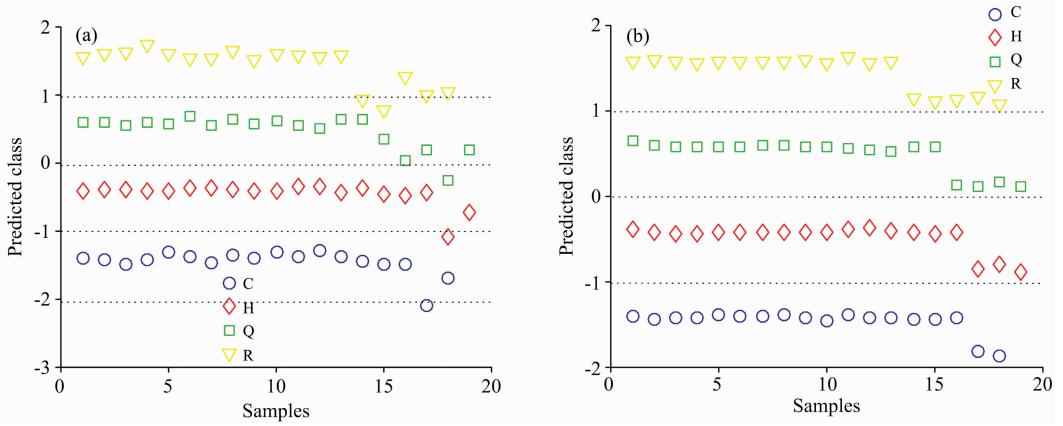

图 9　PLS-DA 建模及分类结果

Fig. 9　PLS-DA modeling and classification results

表 2　三维荧光光谱的 PLS-DA 建模及分类评价结果

Table 2　PLS-DA modeling and classification evaluation results of 3D fluorescence spectra

Evaluation index	EEM[a]-PLS-DA				ConstEEM[b]-PLS-DA			
	C	H	Q	R	C	H	Q	R
CC training/%	100	100	100	100	100	100	100	100
CC Test/%	100	50	60	20	100	100	100	100
Accuracy/%	50	41.67	30	−16.67	100	100	100	100
Sensitivity/%	41.67	40	0	−200	100	100	100	100
Specificity/%	100	50	60	20	100	100	100	100
F-score/%	58.82	44.44	0	44.44	100	100	100	100

注：[a] 样本的三维荧光光谱；[b] 样本的重构三维荧光光谱

3　结　论

对未知油类进行有效识别是解决油类污染问题的前提。本文采集了四种油类在不同背景环境下配制的 80 个油类样本的三维荧光光谱数据，然后利用 PARAFAC 对三维光谱数据进行了重构，并通过 PLS-DA 建立了油类样本的分类模型。该模型能够对四种不同的油类进行准确分类，识别准确率均为 100%。本文为油类污染识别提供了一种实用的新方法。

References

[1] Danovaro R, Carugati L, Berzano M. Front. Mar. Sci., 2016, 3: 213.

[2] Laffon B, Eduardo Pásaro, Valdiglesias V. Journal of Toxicology & Environmental Health Part B, 2016, 19(3-4): 105.

[3] Langangen O, Olsen E J, Stige L C, et al. Marine Pollution Bulletin, 2017, 119(1): 102.

[4] KONG De-ming, ZHANG Chun-xiang, CUI Yao-yao, et al(孔德明, 张春祥, 崔耀耀, 等). Acta Optica Sinica(光学学报), 2018, 38(11): 1130005.

[5] ZHOU Yan-lei, ZHOU Fei-fei, JIANG Cong-cong, et al(周艳蕾, 周飞飞, 姜聪聪, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(2): 475.

[6] Silvana M Azcarate, Adriano de Araújo Gomes, Mirta R Adriano, et al. Food Chemistry, 2015, 184: 214.

[7] Gu Huiwen, Wu Hailong, Yin Xiaoli, et al. Analytica Chimica Acta, 2014, 848: 10.

[8] Camilo L M Morais, Kássio M G Lima, Francis L Martin. Analytica Chimica Acta, 2019, 1063: 40.

[9] Callejóna R M, Amigo J M, Pairo E, et al. Talanta, 2012, 88: 456.

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

[11] Camilo L M Morais, Kássio M G Lima, et al. Chemometrics and Intelligent Laboratory Systems, 2019, 188: 46.

[12] Morais C L M, Lima K M G. Chemometr. Intell. Lab. Syst., 2017, 170: 1.

An Oil Identification Method Based on Reconstructed 3D Fluorescence Spectra Combined With Partial Least Squares Discriminant Analysis

CUI Yao-yao[1], KONG De-ming[2,3*], KONG Ling-fu[1], WANG Shu-tao[2], SHI Hui-chao[4]

1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
2. School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
3. Department of Telecommunications and Information Processing, Ghent University, B-9000 Ghent, Belgium
4. School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract Oil pollution is becoming more and more frequent, which brings a serious threat to human health and the ecological environment. Therefore, it is of great significance to study effective oil identification methods to protect the ecological environment. Three-dimensional (3D) fluorescence spectra are one of the most effective analytical methods for oils identification. 3D fluorescence spectra data are analyzed by using second-order calibration method. And then the concentration score matrix in the analysis results of the second-order correction method is classified by using pattern recognition, which can realize the qualitative identification of unknown samples. However, in the process of classifying and identifying unknown samples, the above methods only apply the concentration score matrix, which is essential to classify the unknown samples by using the relative content difference of the chemical components contained in the samples. The qualitative load matrix is not used, that is, the qualitative analysis of the sample is not achieved from the chemical components contained in the sample. Thus, a new identification method for oil samples was proposed by combining the reconstructed 3D fluorescence spectra with partial least squares discriminant analysis (PLS-DA). First, 80 oil samples were prepared by using four oils (gasoline, diesel, jet fuel and lubricating oil) in different backgrounds (sodium lauryl sulfate solvent prepared from purified water, tap water, river water and sea water); The 3D fluorescence spectra data of the sample was collected by using FS920 fluorescence spectrometer, and the data were preprocessed by de-scattering and standardized; Then, the abnormal spectra data was identified and deleted by using the Leverage value, and the remaining spectral data was reconstructed by using parallel factor analysis algorithm (PARAFAC); Finally, a classification model of reconstructed 3D fluorescence spectra was established by PLS-DA. The classification model established by reconstructing 3D fluorescence spectra was compared with the classification model established by unreconstructed 3D fluorescence spectra. The results show that, after the reconstruction of the 3D fluorescence spectrum, the correct classification rates of the four oils can be increased from 100%, 50%, 60% and 20% to 100%, 100%, 100% and 100%, respectively. It indicates that the reconstructed 3D fluorescence spectra have obvious intra-class characteristics. The sensitivity (SENS), specificity (SPEC) and F-scores of the classification model established by reconstructing the 3D fluorescence spectrum were 100%, 100%, and 100%, respectively. It indicates that the model established has robust and reliable analysis results. In this paper, 3D fluorescence spectra were reconstructed by using concentration score matrix and load matrix in the PARAFAC analysis results. Therefore, the PLS-DA classification model established by reconstructing 3D fluorescence spectra qualitatively identified samples not only from the difference in the relative content of chemical components, but also from the chemical components itself. Its results were convincing. This study provides a reliable method for oil identification.

Keywords Reconstructed 3D fluorescence spectrum; PARAFAC; PLS-DA; Oil identification

(Received Jun. 5, 2019; accepted Oct. 10, 2019)

* Corresponding author