基于三维荧光光谱结合二维线性判别分析的油类识别方法的研究

孔德明[1]，董　瑞[1]，崔耀耀[2*]，王书涛[1]，史慧超[3]

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. 北京化工大学信息科学与技术学院，北京　100029

摘　要　油类污染严重威胁到自然环境及人类健康。因此，识别和处理油类污染非常重要。由于三维荧光光谱能够表征石油的荧光特征，故一般利用三维荧光光谱法检测溶液中存在的油类污染物。但油类的三维荧光光谱数据维度较高且直接分析的难度较大，因此可以利用数据降维方法提取原始油类样本的光谱特征，并利用所得到的光谱特征对样本进行识别。基于此，利用二维线性判别分析(2D-LDA)对油类样本进行特征提取，研究提取的不同样本光谱特征的差别，将得到的光谱特征作为K最近邻(KNN)分类的输入，得到相应的分类结果。首先，分别配制四种不同的油类(柴油、汽油、航空煤油、润滑油)样本各20个，共计得到80个油类样本；然后，利用FS920光谱仪采集所有油类样本的三维荧光光谱数据；其次，对采集到的光谱数据进行预处理，去除光谱中散射的干扰并标准化；最后，利用2D-LDA算法对样本进行特征提取，利用KNN算法进行分类，并将其分类结果与经主成分分析(PCA)进行特征提取后的分类结果比较。研究结果表明，2D-LDA提取特征的分类效果优于PCA。利用2D-LDA分别提取发射和激发特征得到测试集识别的准确率相同且都为95%，而将发射和激发光谱特征的分类距离相结合并重新进行分类的准确率为100%。表明两类光谱相对于三维荧光光谱具有互补性，将发射和激发光谱特征相结合能够更好地对样本进行分类。而利用PCA对测试集识别的准确率仅为85%，表明2D-LDA对三维荧光光谱数据的特征提取效果更好。与PCA相比，2D-LDA通过类内散度和类间散度最大化投影向量来提取样本的特征，使得同类样本尽可能接近，不同样本尽可能分离。因此，2D-LDA具有使降维后的数据更容易被区分的特点，故其鲁棒性好。该研究为油类的降维识别提供了一种参考。

关键词　三维荧光光谱；二维线性判别分析；主成分分析；K最近邻

中图分类号：O433.4　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)08-2505-06

引　言

石油是人们生产生活中的重要能源之一，其具有不可替代的重要作用。但从开采到应用的每一个环节中都有大量的石油及其产品(汽油、煤油、柴油等)以各种方式泄露到自然环境中。这不仅严重污染了自然环境，还致使大量生物死亡，甚至威胁到人类的生命健康[1]。因此，只有及时对油类造成的污染进行处理，才能有效保护生态环境和人类身体健康。而对造成污染的油类进行准确地定性是处理油类污染的前提基础，具有十分重要的意义。

石油的检测方法主要有红外光谱法[2]、气相色谱法[3]和荧光光谱法[4]等。其中，三维荧光光谱法具有分析速度快、灵敏度高、可操作性强等优点，因此被广泛应用到油类识别领域[4]。鉴别石油的方法通常分为两种：一是采用多维分解算法(PARAFAC[5]、AWRCQLD[6]等)对油类的三维荧光光谱进行解析，以得到具有定性信息的相对发射光谱矩阵和相对激发光谱矩阵，并基于此对油类样本进行识别；二是先对样本的光谱数据进行降维，将其平均值、标准差、重心等[7]作为三维荧光光谱数据的特征，依此实现石油种类的识别。两种方法都是先提取能够定性光谱数据的信息，但前者易受

收稿日期：2020-07-09，**修订日期**：2020-11-13

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，F2017203220)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

*通讯作者　　e-mail：cuiyaoyao@stumail.ysu.edu.cn

算法迭代次数的影响且计算量大，部分二阶分析方法还有对组分数不敏感，易受环境影响等缺点；而后者所采用的方法不能够完全体现样本数据的特征。所以寻找能够直接、快速地提取不同油类光谱特征的方法对石油的准确分类具有重要意义。

本文将三维荧光光谱技术与 2D-LDA 算法相结合，并利用 K 最近邻算法对目标油类进行分类。结果表明利用 2D-LDA 算法提取的二维特征能够比较全面的表征原始数据，将其用于石油分类能够获得更优的识别效果。

1 实验部分

1.1 材料与方法

实验采集样本三维荧光光谱数据的仪器为 FS920 荧光光谱仪。设置其发射波长范围为 280～520 nm，步长为 5 nm，激发波长范围为 260～500 nm，步长为 10 nm。实验分别配制了航空煤油(J)、润滑油(L)、柴油(D)、汽油(G)四种不同类型的油类溶液。

实验配制油类溶液的步骤如下：(1)分别取用适量纯净水及十二烷基硫酸钠(SDS)配制成溶解石油所用的样本溶剂；(2)用精密电子秤分别称取相同质量的航空煤油、润滑油、柴油、汽油于四个烧杯中，加入适量的样本溶剂，并用玻璃棒进行搅拌使其充分溶解，分别将溶液转移至四个 100 mL 的容量瓶中并定容，此为四种石油溶液的一级储备液；(3)利用移液枪分别移取 20 个不同体积的航空煤油的一级储备液于 10 mL 的容量瓶中并定容，此为航空煤油的二级储备液；(4)取适量航空煤油的二级储备液于比色皿中，并将比色皿放入 FS920 光谱仪中采集光谱数据；(5)按照步骤(3)—(4)的方法分别对润滑油、柴油、汽油进行配制，得到浓度范围为 0.1～2.0 mg·mL^{-1}且梯度为 0.1 mg·mL^{-1}的四种油类样本。

实验结束后每种石油得到 20 个样本，四种石油共计采集得到 80 个样本。利用 Kennard-Stone 算法将样本分成两组，其中一组作为训练集，另一组作为测试集，训练集中含有 60 个样本，测试集含有 20 个样本。

1.2 去除散射光谱

利用光谱仪采集得到每个样本的光谱数据维度大小为 49×25，其中 49 为发射波长数，25 为激发波长数。一般地，由于光的散射效应，使得所采集溶液的三维荧光光谱中存在瑞利散射和拉曼散射，如图 1(a)所示。图 1(a)中凸起的峰为

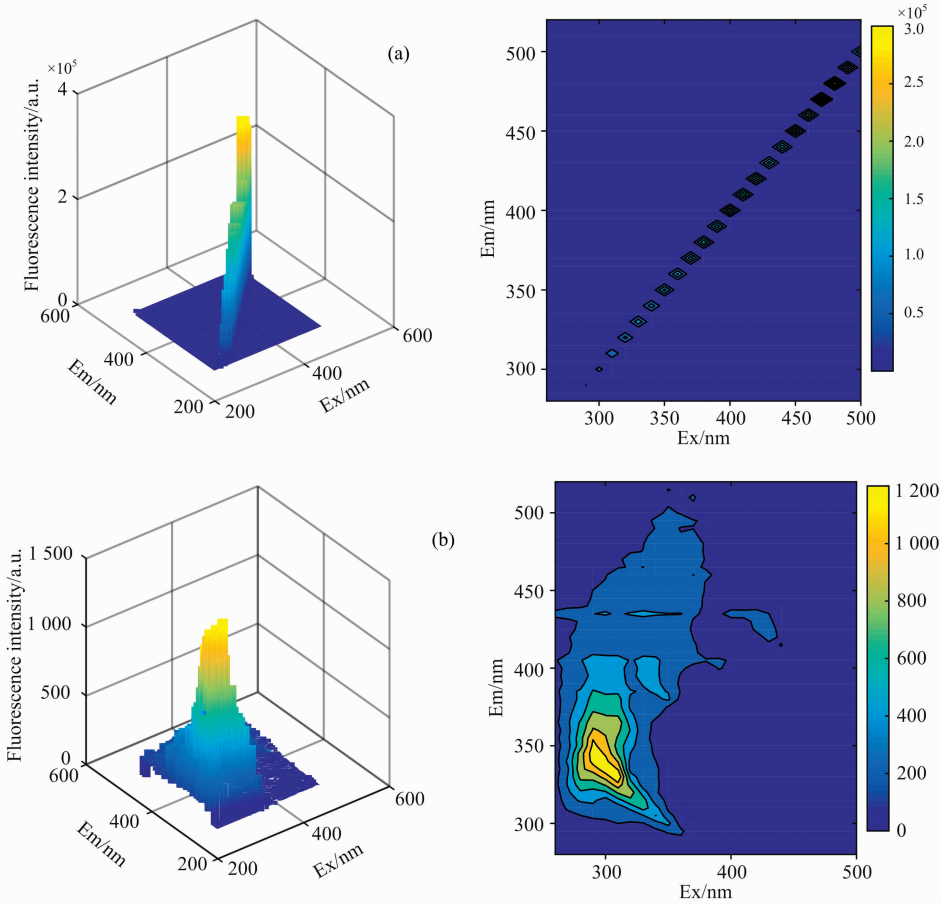


图 1 润滑油三维荧光光谱图

(a)：去散射前；(b)：去散射后

Fig. 1 Three-dimensional fluorescence spectrum of lubricating oil

(a)：Before removing scattering；(b)：After removing scattering

瑞利散射，瑞利散射的强度严重掩盖了润滑油本身的光谱，为了避免散射对实验产生的干扰，必须对光谱进行去散射处理。图 1(b)为利用 Delaunay 三角形内插值法去除散射后润滑油的三维荧光光谱图和等高线图，能够清晰的发现散射光谱被去除，润滑油的光谱得到凸显。

1.3　二维线性判别分析

2D-LDA 利用类内散度和类间散度优化投影矢量，通过原始矩阵在投影矩阵上投影，得到相应的特征矢量。因此，2D-LDA 能够直接通过矩阵提取特征，而不需要先将二维矩阵展开为一维向量再提取特征。所以，2D-LDA 能够在保留原始结构信息基础上有效提取用于分类的特征信息。

设 $\boldsymbol{X}_{m\times n}$ 为一个三维荧光光谱数据矩阵，其中 m 为发射波长数，n 为激发波长数，将 $\boldsymbol{X}$ 乘以一个投影向量，会得到关于矩阵 $\boldsymbol{X}$ 的特征向量。通过最大化线性投影准则，得到最佳投影向量为 $b_{\text{opt}}=\arg\max_b \dfrac{b^{\mathrm{T}}S_B b}{b^{\mathrm{T}}S_W b}$，其中 $S_B(n\times n)$ 和 $S_W(n\times n)$ 分别为类间散度和类内散度，它们是通过训练数据集计算的[8-10]

$$S_B=\sum_{p=1}^{L}N_p(\overline{X}_p-\overline{X})^{\mathrm{T}}(\overline{X}_p-\overline{X}) \tag{1}$$

$$S_W=\sum_{p=1}^{L}\sum_{k\in I_p}N_p(\overline{X}_k-\overline{X}_p)^{\mathrm{T}}(\overline{X}_k-\overline{X}_p) \tag{2}$$

其中，$\overline{X}_p=\dfrac{1}{N_p}\sum_{k\in I_p}X_k$，$\overline{X}=\dfrac{1}{N}\sum_{k=1}^{N}X_k$，$\overline{X}_p$ 表示为在所有样本中属于 p 类样本的平均值，$\overline{X}$ 表示所有样本的平均值。如果 S_W 是非奇异的，则 b_{opt} 可以通过广义特征值的特征向量获得，即 b_{opt} 必须满足以下条件：$S_W^{-1}S_B b_{\text{opt}}=\lambda b_{\text{opt}}$。同样，这个过程可以被扩展为 M 个投影向量，假设 M 为 $S_W^{-1}S_B$ 的秩，则其第 q 个投影向量可以通过 $S_W^{-1}S_B b_q=\lambda_q b_q$ 获得，其中 $q=1,2,\cdots,M$，如果 $r(r\leqslant M)$ 是通过以上方式获得的投影向量的数量，则将 $b_1,b_2,\cdots,b_r$ 作为投影矩阵 $\boldsymbol{B}_{n\times r}$ 的列，最后可以由 $\boldsymbol{Y}=\boldsymbol{XB}$ 得到二维特征矩阵 $\boldsymbol{Y}_{n\times r}$。

1.4　KNN 分类算法

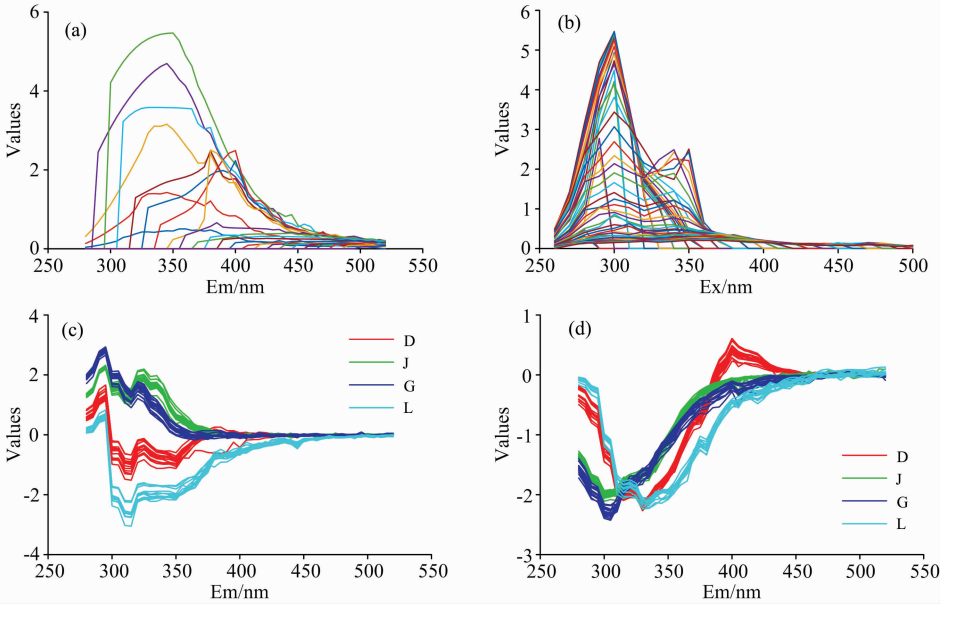
K 最近邻分类器是通过比较特定的测试元组和一组与它类似的训练元组来进行学习，最后基于最近邻居的类别进行分类的分类方法。KNN 通常应用欧几里德距离或者训练组与测试组之间的余弦相似度[11-12]。一般地，两个元组例如 $E=(e_1,e_2,\cdots,e_n)$ 和 $F=(f_1,f_2,\cdots,f_n)$ 之间的欧几里德距离为

$$\text{dist}(E,F)=\sqrt{\sum_{i=1}^{n}(e_i-f_i)^2}$$

2　结果与讨论

2.1　利用 2D-LDA 提取特征

图 2(a)和(b)分别为单个润滑油样本原始的发射光谱和激发光谱。图 2(c)和(d)分别为提取的所有训练样本水平方向上第一、第二投影向量的特征信息，也即发射光谱的特征信息；图 2(e)和(f)分别为提取的所有训练样本垂直方向上第一、第二投影向量的特征信息，即激发光谱的特征信息；其中 D，J，G 和 L 分别为柴油、航空煤油、汽油、润滑油。由图 2 可知，2D-LDA 算法提取的油类样本的特征光谱信息降低了原来样本数据的维度，通过前两个主要投影向量对样本的三维荧光光谱投影得到的光谱信息具有明显区分不同类型石油样本的作用。图 2(c)和(d)所示的发射光谱特征中，不同类型石油的差别集中在 280～450 nm；图 2(e)和(f)所示的激发光谱特征中，不同类型石油的区别集中在 260～350 nm。产生这种现象原因是发射光谱中 450 nm 之后的石油的荧光强度极低且接近于 0，同样在激发光谱中 350 nm 之后的石油的荧光强度极低且接近于 0。因此，在发射波长为 280～520 nm，激发波长为 260～500 nm 的范围内，柴油、航空煤油、汽油、润滑油这四种油的有效发射光谱波长和激发光谱波长范围分别为 280～450 和 260～350 nm。



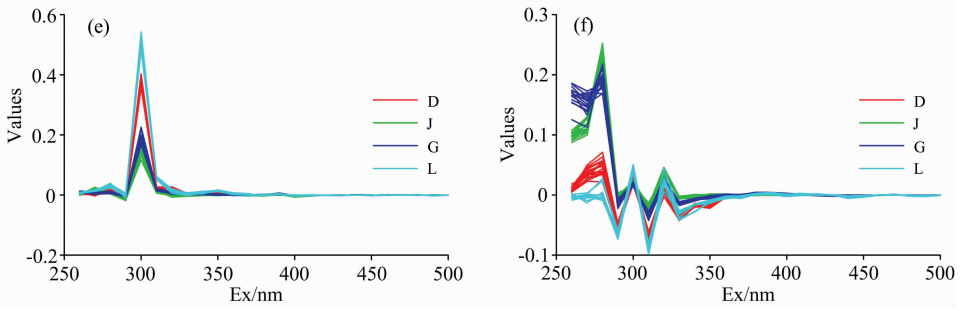


图 2 润滑油原始发射、激发光谱图及通过投影向量获取的训练集样本光谱特征

(a)：发射光谱图；(b)：激发光谱图；(c)：第一投影向量的发射特征；(d)：第二投影向量的发射特征；(e)：第一投影向量的激发特征；(f)：第二投影向量的激发特征

Fig. 2 Original emission and excitation spectrum of lubricating oil and characteristics of training sample obtained by projection vector

(a): Emission spectrum; (b): Excitation spectrum; (c): First emission characteristic; (d): Second emission characteristic; (e): First excitation characteristic; (f): Second excitation characteristic

2.2 利用 PCA 提取特征

利用 PCA 提取原始光谱数据的特征信息，得到相应的主成分的特征值及对应的贡献率。根据每个主成分对应的贡献率和累积贡献率，选取合适的主成分数建立分类模型。前十个主成分所对应的贡献率如表 1 所示。由表 1 可知，前四个主成分的贡献率分别为 66.52%，19.63%，4.61% 和 3.12%，累积贡献率为 93.88%。在主成分分析中选取的主成分数需要包含原始数据的大部分信息，因此选取前四个主成分作为后续分析的主成分数。

表 1 主成分的贡献率

Table 1 Contribution rate of principal component

Principal component	Contribution rate/%
PC1	66.52
PC2	19.63
PC3	4.61
PC4	3.12
PC5	2.61
PC6	0.92
PC7	0.34
PC8	0.27
PC9	0.26
PC10	0.23

绘制训练集前三个主成分的得分散点图，如图 3 所示。由图 3 可知，图中同种类型的样本聚集在一起，而不同类型的样本彼此分离，具有明显的区别。并且图中不同类型的样本没有重叠的情况发生，表明 PCA 能够较好的提取光谱的特征信息，但存在少数样本会偏离同类型大部分样本的聚集位置。将测试集的样本在由训练集建立的模型上投影，得到测试集中各样本的得分，并以此作为分类的信息。

2.3 样本分类

2.3.1 2D-LDA 提取特征后分类

分别将 2D-LDA 提取后的发射和激发光谱特征作为 KNN 分类模型的输入，并通过计算样本之间的距离对所有样本分类，分类结果如表 2 所示。由表 2 可知，发射光谱特征作为输入时，测试集中的柴油、航空煤油和润滑油分类的准确率为 100%，存在一个汽油样本被错误分类为航空煤油，故汽油分类的准确率为 80%，但在整个测试集中，存在 20 个样本，只有一个样本被错误分类，因此测试集中样本分类的准确率为 95%；激发光谱特征作为输入时，存在一个柴油样本被错误分类为润滑油，但其整个测试集分类的准确率也为 95%。三维荧光光谱包含发射光谱和激发光谱，且两类光

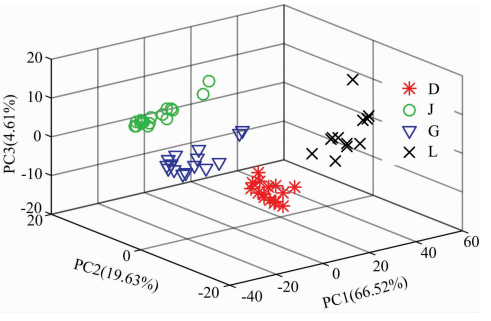


图 3 前三主成分得分图

Fig. 3 First three principal component score

表 2 利用 2D-LDA 特征提取后的分类结果

Table 2 Classification results after 2D-LDA characteristics extraction

	emission				excitation			
	D	J	G	L	D	J	G	L
D	4	0	0	0	3	0	0	1
J	0	3	0	0	0	3	0	0
G	0	1	4	0	0	0	5	0
L	0	0	0	8	0	0	0	8
SENS/%	100	100	80	100	75	100	100	100
SPEC/%	100	94	100	100	100	100	100	92
ACC/%	100	100	80	100	75	100	100	100

注：Acc(Accuracy)为样本识别的准确率。

谱表征三维荧光光谱的不同方向，在一定程度上两类光谱相对三维荧光光谱具有互补性，所以利用两类光谱特征对样本分类结果具有差异。而将两类特征的 KNN 分类距离叠加并重新作为训练集和测试集样本分类的标准，得到识别的准确率为 100%，表明融合发射和激发光谱特征能够对油类样本实现更好地识别。

2.3.2　PCA 提取特征后分类

将 PCA 提取后样本的前四个主成分的得分作为 KNN 分类模型的输入，计算样本之间的距离并分类，结果如表 3 所示。由表 3 可知，只有航空煤油分类准确率为 100%，而柴油、汽油和润滑油中都存在一个样本被错误分类。因此，在整个测试集中，存在 3 个样本被错误分类，故分类的准确率为 85%。由图 3 可知，利用 PCA 特征提取的结果中不存在不同类型石油重叠的情况，但存在少数的样本偏离大部分同类样本。因此偏离的样本可能会存在与其他类型样本的距离小于同类样本距离的情况，从而导致错误分类。

对比两种特征提取方法的分类结果，表明 2D-LDA 提取的光谱数据特征具有提高分类准确率的作用，经 2D-LDA 提取特征后的分类的准确率更高。尽管 2D-LDA 算法提取的不同类型样本的光谱特征曲线存在部分重叠，但其类内差别远小于类间差别，并且识别率高，表明该算法的鲁棒性好。而 PCA 提取的前四个主成分的贡献率虽然已经达到 93.88%，但仍可能会丢失一些重要的辨别样本种类的特征信息，导致分类结果出现错误。因此，2D-LDA 提取石油光谱特征的性能优于 PCA。

表 3　主成分特征提取后的分类结果

Table 3　Classification result after principal component characteristic extraction

	D	J	G	L	Acc/%
D	3	0	0	1	75
J	0	3	0	0	100
G	0	0	4	1	80
L	0	1	0	7	87.5
SENS/%	75	100	80	88	
SPEC/%	100	94	100	83	

3　结　论

采集了航空煤油、润滑油、柴油和汽油的三维荧光光谱，通过 2D-LDA 对其进行二维特征提取，并利用 KNN 算法对样本分类，得到样本的分类结果。实验结果表明，利用 2D-LDA 提取特征后的分类准确率较高，达到 95%，且结合两类光谱特征分类得到的准确率为 100%，而 PCA 特征提取后的分类准确率为 85%。因此，利用二维线性判别分析直接提取三维荧光光谱的二维光谱特征并将其用于定性分析，能够获得更优的油类识别效果。

References

[1] Chen H, Liu S, Xu X R, et al. Marine Pollution Bulletin, 2015, 90(1-2): 181.

[2] Ng W, Malone B P, Minasny B. Geoderma, 2017, 289: 150.

[3] Damavandi H G, Gupta A S, Reddy C, et al. Conference on Signals, Systems & Computers. IEEE, 2015.

[4] Cui Y Y, Kong D M, Kong L F, et al. IEEE Access, 2020, 8: 17999.

[5] 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.

[6] KONG De-ming, ZHANG Chun-xiang, CUI Yao-yao, et al(孔德明，张春祥，崔耀耀，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(10): 3129.

[7] YUAN Yuan-yuan, WANG Shu-tao, KONG De-ming, et al(苑媛媛，王书涛，孔德明，等). Acta Photonica Sinica(光子学报), 2017, 46(11): 1130002.

[8] Silva A C, Soares S F, Insausti M, et al. Analytica Chimica Acta, 2016, 938: 53.

[9] Imani M, Ghassemian H. Photogrammetric Engineering & Remote Sensing, 2015, 81(10): 777.

[10] Li Q, You J. Multimedia Tools and Applications, 2019, 78: 30397.

[11] Chen Z, Zhou L J, Li X D, et al. Procedia Computer Science, 2020, 166: 523.

[12] Swetapadma A, Yadav A. Computers & Electrical Engineering, 2018, 69: 41.

Study on Oil Identification Method Based on Three-Dimensional Fluorescence Spectrum Combined With Two-Dimensional Linear Discriminant Analysis

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

1. School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
2. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
3. School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract Oil pollution seriously threatens the natural environment and human health. Therefore, it is very important to identify and deal with oil pollution. Therefore, three-dimensional fluorescence spectroscopy is generally used to detect the presence of oil contaminants in a certain solution. However, the three-dimensional fluorescence spectrum data of oils have high dimensions, and direct analysis is difficult. Therefore, the data dimensionality reduction method can be used to extract the spectral characteristics of the original oil samples. And the obtained spectral characteristics is used to identify and classify the samples. Based on this, the two-dimensional linear discriminant analysis (2D-LDA) is used to extract the characteristics of the oil samples. The differences in the spectral characteristics of the different samples extracted are studied. The obtained spectral characteristics are used as the input of the K nearest neighbor (KNN) classification to obtain the corresponding. Firstly, four different oils samples (diesel, gasoline, aviation kerosene, lubricating oil) was prepared, and each of the oils has 20 samples. So, 80 oils samples were prepared totally. Secondly, three-dimensional (3D) fluorescence spectrum data of all oil samples are collected by an FS920 spectrometer. Then, the spectral data is pre-processed to remove the scattering and to standardize it. Finally, the 2D-LDA algorithm is used to extract the characteristics of the samples, and the KNN algorithm is used to classify. The results were compared between principal component analysis (PCA) and 2D-LDA. 2D-LDA extracted the emission and excitation characteristics. Both accuracy is 95%. However, the accuracy of combining the classification distances of the emission and excitation spectrum characteristics and re-classifying is 100%. It shows that the two types of spectra are complementary to the three-dimensional fluorescence spectrum, and the combination of emission and excitation spectrum characteristics can better classify the sample. The results show that the classification effect of 2D-LDA characteristics extraction is superior to PCA. It shows that 2D-LDA is better for characteristics extraction of 3D-fluorescence spectrum data. Compared with PCA, 2D-LDA uses the intra-class matrix and the inter-class matrix to maximize the projection vector to extract the characteristics of samples. So, the same type of samples are closer, and the different type of samples are separated as much as possible. Therefore, the 2D-LDA can make it easier to identify data after reducing data dimensionality. Its robustness is good. This study provides a reference to identify oils.

Keywords Three-dimensional fluorescence spectra; Two-dimensional linear discriminant analysis; Principal component analysis; K nearest neighbor

(Received Jul. 9, 2020; accepted Nov. 13, 2020)

* Corresponding author