三维荧光光谱结合稀疏主成分分析和支持向量机的油类识别方法研究

孔德明[1]，陈红杰[1]，陈晓玉[2*]，董 瑞[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

摘 要 石油污染的出现，导致生态环境遭到破坏。因此，油类识别方法的研究对于环境的保护具有重要意义。采用荧光光谱法获得石油光谱数据，并对其进行预处理，再通过降维方法来提取特征信息，最后利用模式识别算法进行分类，从而可以实现对油类的定性分析，因此研究一种更高效的数据降维方法以及识别分类算法极其重要。基于三维荧光光谱技术，利用稀疏主成分分析(SPCA)对FS920光谱仪测得的荧光光谱数据进行特征提取，再利用支持向量机(SVM)算法对提取的特征数据进行分类识别，获得了一种更加高效的油类识别方法。首先，利用海水和十二烷基硫酸钠(SDS)配制成浓度为0.1 mol·L^{-1}的胶束溶液，将其作为溶剂配制柴油、航空煤油、汽油以及润滑油各20种不同浓度的溶液；然后，利用FS920光谱仪测得样本溶液的三维荧光光谱数据，对得到的光谱数据进行预处理；最后，对预处理后的数据分别利用SPCA和主成分分析(PCA)进行特征提取，再利用SVM和K最近邻(KNN)两种模式识别算法对特征向量进行分类，最终得到四种模型PCA-KNN，SPCA-KNN，PCA-SVM以及SPCA-SVM的分类结果。研究结果表明，由四种模型得到的分类准确率分别为85%，90%，90%和95%，其中，在同种分类算法中，利用SPCA进行特征提取得到的分类准确率均比PCA的准确率高5%，因此可知，SPCA的稀疏性具有突出主要成分的作用，在提取光谱特征时能够减小非必要成分的影响，并且载荷矩阵的稀疏化可以去除变量之间的冗余信息，优化降维特征信息，为后续分类提供更有效的数据特征信息；在同种特征提取算法下，利用SVM算法进行分类得到的分类准确率均比KNN算法得到的准确率高5%，表明SVM算法在分类中更具有优势。因此，本文利用三维荧光光谱技术结合SPCA和SVM算法，实现了对石油的准确识别与分类，为今后对石油污染物的高效检测提供了新思路。

关键词 三维荧光光谱；特征提取；稀疏主成分分析；支持向量机

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)11-3474-06

引 言

随着海上石油运输业的迅速发展，海运船只不断增加，导致海上溢油事故频繁发生，海洋石油污染也日益严重，溢油的出现影响着海洋环境，并间接危害着人类的健康[1]。因此，研究一种更高效的油类识别算法，对海洋环境保护以及石油的污染防治具有重要的实用价值。

目前，溢油检测方法有色谱法、质谱法、光谱法等[2]，其中荧光光谱法以其测量灵敏度高、方法简单、环保无污染等优点，成为石油污染物识别的主要方法[3]。三维荧光光谱描述的是物质的荧光强度随激发和发射波长变化的关系，其中包含了丰富的被测物信息，但是在解析光谱时，因物质的荧光性质相似而出现光谱重叠现象，会给后续的分类识别造成影响，所以对光谱数据的预处理以及特征参量的提取是尤为重要的[4]。

特征提取是三维荧光光谱数据分类识别的关键，其主要由参量化方法以及数据降维方法来实现。参量化方法[5]主要通过提取三维荧光光谱的平均值、原点矩、中心距等特征参数代表原始光谱数据，并将其作为分类识别的输入数据，该方法虽明确了光谱的几何特征，但是却忽略了光谱之间的内在联系，只是特征间的简单组合；数据降维方法主要是将数据投影到低维可视空间，以便看清数据的分布，其中最具代表性的是主成分分析(PCA)算法[6]，李杰等[7]提出了基于

收稿日期：2020-10-14，**修订日期**：2021-02-15

基金项目：国家自然科学基金项目(61601399，61501394，61771419)和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：chjmessage@163.com

*通讯作者 e-mail：chenxiaoyu@ysu.edu.cn

PCA 的时间分辨油类荧光分类算法。但是上述方法中，用 PCA 进行特征提取时每一个主成分只是其原始变量的线性组合，并且在这些变量中还存在大量的冗余信息[8]，会降低分类识别的准确率。

因此在 PCA 的基础上提出了稀疏主成分分析(SPCA)算法来优化特征提取算法，得到更有效的特征信息。本文将三维荧光光谱技术结合 SPCA 特征提取方法与支持向量机(SVM)算法对油类进行识别，结果表明 SPCA 的稀疏特性使其在特征提取时更加突出主要成分，并且能够更好地去除荧光光谱间的冗余信息，实现降维的最优化，从而为分类提供更有效的光谱数据，得到更加准确的分类结果。

1　实验部分

1.1　仪器与材料

实验中石油的三维荧光光谱数据由 FS920 荧光光谱仪采集，其中激发波长的范围设置为 260～500 nm，步长为 10 nm，发射波长的范围设置为 280～520 nm，步长为 5 nm。

配制所需溶液的具体步骤如下：首先取适量海水，使其与十二烷基硫酸钠(SDS)混合配制成浓度为 0.1 mol·L^{-1}的 SDS 溶剂；然后用精密电子秤分别称取柴油(Diesel oil)、航空煤油(Jet fuel)、汽油(Gasoline)、润滑油(Lubricating oil)各 0.1 g；其次，用 SDS 溶剂分别与四种油类配制成浓度为 1 mg·mL^{-1}纯油储备液并进行避光保存；最后，以 0.1 mg·mL^{-1}作为浓度间隔，量取四种油类储备液并将浓度稀释至 0.1～2.0 mg·mL^{-1}，每种油得到 20 个样本，共得到 80 个样本。

1.2　数据预处理

本实验得到 80 个 49×25 的二维数据矩阵，记 $X=x_{kij}$ 即 $X=x_{80\times49\times25}$，其中 x_{kij} 表示第 k 个样本，发射波长为 i，激发波长为 j 时所对应的荧光强度，利用 Delaunay 三角形内插值法消除荧光光谱中的散射，再将荧光光谱曲线进行平滑处理，得到如图 1 所示的三维荧光光谱图。

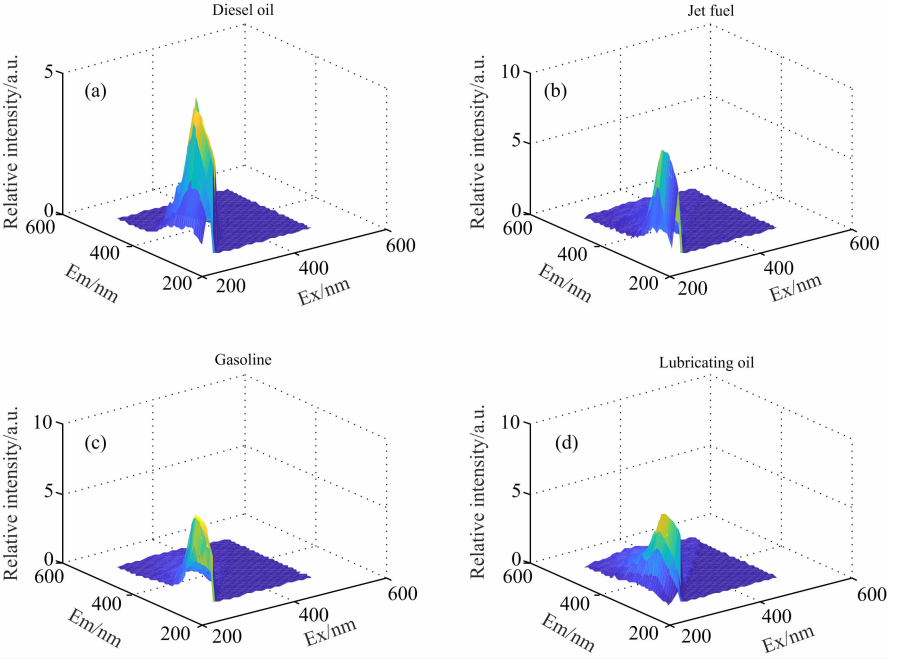

图 1　去除散射后的荧光光谱图

Fig. 1　Fluorescence spectrum after removing scattering

将 X 经过数据重组得到二维数据矩阵 $X=x_{80\times1\,225}$，利用 Kennard-Stone 算法按 3∶1 的比例将 80 个样本分为 60 个训练样本和 20 个测试样本，即得到的训练样本数据与测试样本数据分别为 $X_1=x_{60\times1\,225}$ 和 $X_2=x_{20\times1\,225}$。对得到的训练集和测试集样本数据同时进行归一化处理，避免了因维度量纲的差距导致较低的数量级属性变为 0 的问题，同时保留了更多的原始数据信息，为后续的分类提供了更有效的数据信息。

2　检测原理

2.1　稀疏主成分分析

稀疏主成分分析(sparse principal component analysis，SPCA)是利用弹性网产生具有稀疏载荷的修正主成分，其稀疏主成分由 PCA 算法的线性组合问题转化为回归型优化问题，其通过对回归稀疏施加 Lasso 约束，得到稀疏载荷[9-10]。

SPCA 算法求解过程：

(1)使 A 从 $V[:,1:k]$开始，即前 k 个普通主分量的载荷；

(2)给定一个固定的 $A=[\alpha_1,\cdots,\alpha_k]$，求解以下弹性网问题：$j=1,2,\cdots,k$；

$$\beta_j=\operatorname{argmin}(\alpha_j-\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\alpha_j-\beta)+\lambda\|\beta\|^2+\lambda_{1,j}\|\beta\|_1$$

(3)给定一个固定的 $A=[\alpha_1,\cdots,\alpha_k]$，计算 $X^{\mathrm{T}}XB=$

UDV^{T} 的奇异值，然后更新 $A = UV^{T}$；

(4)重复步骤(2)—(3)，直至收敛；

(5)标准化：$\hat{V}_j = \frac{\beta_j}{\| \beta_j \|}$，$j = 1, 2, \cdots, k$。

SPCA 在 PCA 的基础上，添加了惩罚函数，并且以略微牺牲贡献率为代价获得载荷稀疏化，从而明确了变量之间的关系，体现出了 SPCA 的优越性，并且 SPCA 的稀疏性可突出主要成分，同时可以极好地去除变量间的冗余信息，为分类提供更加有效的数据基础，从而使评价结果更加可靠且确切。

2.2 SVM 支持向量机

支持向量机(support vector machine，SVM)是基于结构风险最小原理的监督学习算法，其适用于小样本、非线性及高维的数据模型，在收敛、最优解及泛化能力等方面有着一定的优势[11]。

SVM 既可以解决线性问题又可以解决非线性问题[12]。对于线性问题，根据分类间隔最大化的原则，来定义最大超平面和支持向量机，将线性问题转化成凸约束条件下的凸规划问题进行解决。针对非线性问题，基本思想是通过变换输入变量的运算空间，将非线性内积映射到高维的特征空间，来使非线性问题转化成容易解决的线性问题，即将求解最优分类面的问题转化成求解约束优化问题，引入松弛变量 ξ 和惩罚因子 c，转化成线性不可分的 SVM 问题，因其满足 Slater 条件，通过拉格朗日对偶将其转化为对偶问题进行求解，最后通过引入核函数将对偶问题转化约束优化问题，并求其最优解。

本文所采用的核函数是径向基(RBF)核函数，它对复杂模型数据有着较好的处理能力。同时又通过网格搜索法选取核参数 g 和惩罚参数 c[13]，即先选定惩罚参数 c 和 RBF 核参数 g 的变化范围，在此范围内寻找最佳的参数值，并对其进行 K 折交叉验证(K-fold cross validation)，此方法可遍布网格内所有的参数，得到全局最优解，即交叉验证意义下最高的分类准确率。

3 结果与讨论

3.1 不同类型石油的三维荧光光谱

实验测得 80 个荧光光谱样本，可得到如表 1 所示的石油荧光特征参数，四种石油的荧光光谱特征参数之间既存在相似性又存在差异性，根据四种石油的激发峰值波长范围、发射峰值波长范围以及荧光强度可知，不同类型的石油光谱间出现了光谱重叠现象；由荧光峰个数可知，只有柴油光谱有两个峰。出现异同的原因可能是四种油所含荧光物质的种类以及各类荧光物质的比例不同[14]。

表 1 四种类型石油的荧光特征参数

Table 1 Fluorescence characteristic parameters of four types of petroleum

Classification	Em peak range/nm	Ex peak range/nm	Number	Relative intensity of main peak/(a. u)
Diesel oil	295～360	280～310	2	3.302～4.874
	390～405	330～340		2.907～4.713
Jet fuel	280～340	270～290	1	4.377～7.028
Gasoline	280～340	270～280	1	5.844～5.952
Lubricating oil	300～365	290～310	1	3.964～5.571

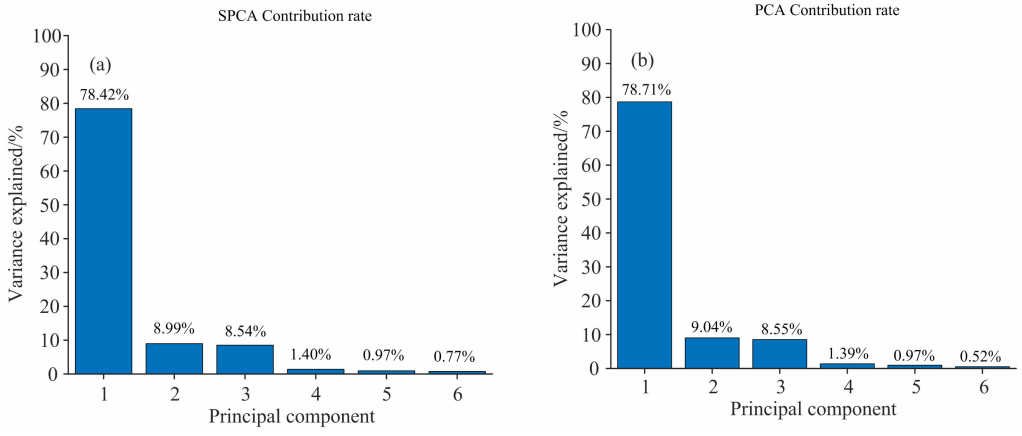

图 2 前六个成分的贡献率

Fig. 2 Contribution rate of the first six principal components

3.2 特征提取

直接通过四种石油的荧光光谱特征参数很难对其进行鉴别分析，因此本文分别利用 PCA 和 SPCA 算法对预处理后的数据进行特征提取，两种算法提取得到的贡献率如图 2 所

示，由 SPCA 算法得到的前六个主成分的累计贡献率为 99.09%，贡献率分别为 78.42%，8.99%，8.54%，1.40%，0.97%和 0.77%；由 PCA 算法得到的前六个主成分的累计贡献率为 99.18%，贡献率分别为 78.71%，9.04%，8.55%，1.39%，0.97%和 0.52%。由于载荷稀疏是以损失方差贡献率为代价，故 SPCA 第一主成分的贡献率低于 PCA，但两者的累计贡献率均超过了 99%，能够表示样本的数据信息，为后续分类提供了强有力的数据特征信息。

3.3　模型的建立

3.3.1　PCA-SVM 与 SPCA-SVM 模型

对预处理的数据分别利用 PCA 与 SPCA 算法进行提取特征后，前 6 个主成分的累计贡献率分别达到 99.18%和 99.09%。将得到的降维数据分别输入到 SVM 中进行分类，采用网格搜寻方法，得到最佳参数 c 与核参数 g 如表 2 所示，再利用最佳参数对模型进行训练，得到 CV 意义下的分类准确率，其中 SVM 对训练集分类的准确率为 100%，从而为测试集的预测提供了很好的模型基础。如图 3 为两种算法得到的参数选择结果图。

表 2　SVM 模型的最佳参数及准确率

Table 2　The best parameters and accuracy of SVM model

	模型参数		训练集准确率/%	预测集准确率/%
	惩罚参数 c	核参数 g		
PCA-SVM	1.741 1	0.006 8	100	90
SPCA-SVM	0.189 5	9.189 6	100	95

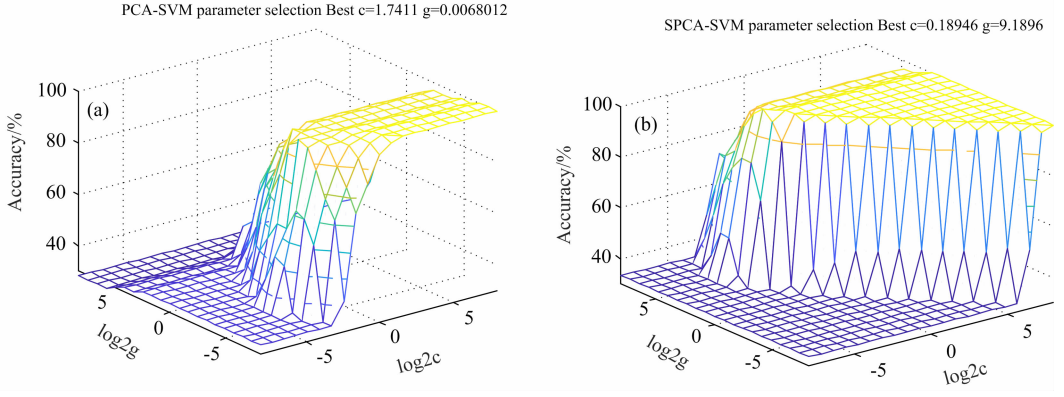

图 3　SVM 参数选择结果图

Fig. 3　SVM parameter selection result

通过 SVM 进行分类预测后，得到的预测结果如图 4 所示，PCA-SVM 算法得到的预测结果中有 2 个样本被错误分类。其中，柴油错分为润滑油，汽油错分为润滑油，分类准确率为 90%。而经过 SPCA-SVM 算法得到的预测结果中仅有一个样本被错误分类，即汽油错分为润滑油，减少了柴油的错分率，使分类准确率从PCA-SVM算法的90%提高到95%。对比两种算法可知，SPCA-SVM 算法提高了分类准确率，因为 SPCA 算法是将其稀疏特性与主成分结合，将主成分的系数即构成主成分时每个变量前面的系数变得稀疏，可以在特征提取时将其非必要成分的系数进行稀疏，从而更加突出主成分的作用，同时这样也可以减少数据间的冗余，从而使 SVM 在分类时更加准确，因此 SPCA 提取到的特征信息比 PCA 得到的更具有代表性。

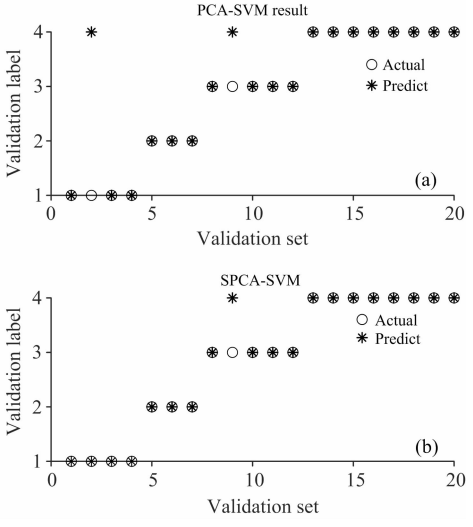

图 4　SVM 预测得到的分类结果图

Fig. 4　Classification results predicted by SVM

3.3.2　PCA-KNN 与 SPCA-KNN 模型

将数据利用 KNN 算法进行分类，得到的预测结果如图 5 所示，对于 PCA-KNN 而言，20 个样本中错分了三个样本，柴油错分为航空煤油，汽油错分为润滑油，润滑油错分为航空煤油，PCA-KNN 算法得到的分类准确率为 85%。而对于 SPCA-KNN 而言，有两个样本被错误分类，其中，汽油错分为润滑油，润滑油错分为航空煤油，由分类结果可知，SPCA-KNN 算法分类准确率高于 PCA-KNN，主要是因为 SPCA 特征提取方法优于 PCA，同样体现出了 SPCA 在降维中的优势，可为分类提供更可靠的特征信息。

表 3 为四种分类模型预测结果。对比可知，在同类型的分类算法中，利用 SPCA 进行特征提取得到的分类准确率高于 PCA；在相同的特征提取算法中，SVM 算法的分类结果优于 KNN。结果表明：SPCA-SVM 分类算法是更加高效的油类识别算法。

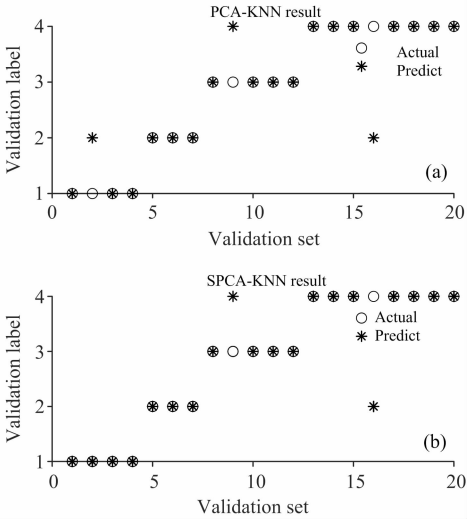

图 5 KNN 预测得到的分类结果图

Fig. 5 Classification results predicted by KNN

表 3 不同算法得到的分类结果(准确率/%)

Table 3 Classification results obtained by different algorithms (accuracy rate/%)

Classification	Feature extraction	
	PCA	SPCA
KNN	85%	90%
SVM	90%	95%

4 结 论

采集了四种类型石油的荧光光谱，首先利用 SPCA 特征提取方法进行光谱特征提取，然后利用 SVM 算法进行分类。研究结果表明，SPCA 算法可以更好地去除油类光谱数据变量间的冗余信息，实现对数据的降维，为分类提供更加可靠且有效的数据，对比参量化法和 PCA 等降维方法，SPCA 得到的提取信息更具代表性，再结合高效的 SVM 分类算法，从而使预测结果变的更加准确，因此本文得出了一种高效且准确检测石油的算法即 SPCA-SVM 算法，为后续石油污染的检测研究提供了一个新思路。

References

[1] Loh Andrew, Ha Sung Yong, An Joon Geon, et al. Marine Pollution Bulletin, 2019, 138(Jan.): 328.

[2] LIU Hui, YUAN Xing-qi, WANG Juan, et al(刘 慧, 袁兴启, 王 娟, 等). Physical Testing and Chemical Analysis Part B(Chemical Analysis)(理化检验·化学分册), 2018, 54(2): 241.

[3] DU Yun, ZHENG Ya-nan, WANG Shu-tao(杜 云, 郑亚南, 王书涛). Optics and Precision Engineering(光学精密工程), 2018, 26(9): 2213.

[4] XU Li, ZHANG Zhi-rong, DONG Feng-zhong, et al(许 丽, 张志荣, 董凤忠, 等). Laser & Optoelectronics Progress(激光与光电子学进展), 2019, 56(19): 193003.

[5] CHEN Zhi-kun, GUO Rui, CHENG Peng-fei(陈至坤, 郭 蕊, 程朋飞). Laser & Optoelectronics Progress(激光与光电子学进展), 2020, 57(13): 133002.

[6] Jana Sádecká, Michaela Jakubíková, Pavel Májek. Food Control, 2018, 88: 75.

[7] LI Jie, LI Xiao-long, TANG Qiu-hua, et al(李 杰, 李晓龙, 唐秋华, 等). Optics and Precision Engineering(光学精密工程), 2017, 25(4): 352.

[8] WANG Lei, NIE Hui(王 磊, 聂 晖). Instrument Technique and Sensor(仪表技术与传感器), 2017, (9): 94.

[9] GUI Dong-dong, LU Qi, JIN Can-can, et al(桂冬冬, 鲁 齐, 金灿灿, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(S1): 133.

[10] FU Jian-hua, ZHOU Xin-qi, LIU Hui-jun, et al(付建华, 周新奇, 刘辉军, 等). Physical Testing and Chemical Analysis Part B(Chemical Analysis)(理化检验·化学分册), 2017, 53(2): 146.

[11] Wu Xijun, Zhao Zhilei, Tian Ruiling, et al. Food Chemistry, 2020, 311: 125882.

[12] WANG Shu-tao, WU Xing, ZHU Wen-hao, et al(王书涛, 吴 兴, 朱文浩, 等). Acta Optica Sinica(光学学报), 2019, 39(5): 0530002.

[13] CHEN Ying, ZHANG Can, XIAO Chun-yan, et al(陈 颖, 张 灿, 肖春艳, 等). Acta Optica Sinica(光学学报), 2020, 40(10): 1030002.

[14] Santos M C D, Azcarate S M, Lima K M G, et al. Microchemical Journal, 2020, 155: 104783.

Research on Oil Identification Method Based on Three-Dimensional Fluorescence Spectroscopy Combined With Sparse Principal Component Analysis and Support Vector Machine

KONG De-ming[1], CHEN Hong-jie[1], CHEN Xiao-yu[2*], DONG Rui[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

Abstract The emergence of oil pollution has destroyed the ecological environment. Therefore, the study of oil identification methods is of great significance to the protection of the environment. Petroleum spectrum data can be obtained by fluorescence spectroscopy. At the same time, the spectrum data is preprocessed, and feature information is extracted by dimensionality reduction. Then the pattern recognition algorithm is used for classification, it can realize the qualitative analysis of oil. However, it is vital to study a more efficient way of data dimensionality reduction and recognition algorithms. Based on the three-dimensional fluorescence spectroscopy technology, this paper uses sparse principal component analysis (SPCA) to extract the features of the fluorescence spectrum data measured by the FS920 spectrometer, and the support vector machine (SVM) algorithm applies for classification and recognition, thereby a more efficient oil identification method is obtained. First, seawater and sodium dodecyl sulfate (SDS) was prepared into a micelle solution with a concentration of 0.1 mol·L^{-1}. It was used as a solvent to prepare solutions of 20 different concentrations of 4 kinds of oil: Diesel oil, Jet fuel, Gasoline and Lubricating oil. Then, the three-dimensional fluorescence spectrum was measured by the FS920 spectrometer, and the data schould be preprocessed. Finally, the pre-processed data is extracted using SPCA, and principal component analysis (PCA), and the feature vectors are classified by SVM and K-nearest neighbor (KNN) two pattern recognition algorithms, the classification results of four models PCA-KNN, SPCA-KNN, PCA-SVM and SPCA-SVM are obtained. The research results show that the classification accuracy rates obtained by the four models are 85%, 90%, 90% and 95% respectively. In the same classification algorithm, the classification accuracy obtained by using SPCA is 5% higher than that of PCA. Therefore, SPCA can better highlight the main components in its sparsity, and the sparsity of the load matrix can remove redundant information between variables, achieve the optimization of dimensionality reduction, and provide a better classification for subsequent classification. Effective data feature information; Under the same feature extraction algorithm, the classification accuracy rate obtained by using the SVM algorithm for classification is 5% higher than the accuracy rate obtained by the KNN algorithm, it shows that the SVM algorithm has more advantages in classification. Therefore, this paper uses three-dimensional fluorescence spectroscopy technology combined with SPCA and SVM algorithms to accurately identify petroleum, which provides a new idea for the efficient detection of petroleum pollutants in the future.

Keywords Three-dimensional fluorescence spectrum; Feature extraction; Sparse principal component analysis; Support vector machine

(Received Oct. 14, 2020; accepted Feb. 15, 2021)

* Corresponding author