三维荧光光谱结合 GA-SVM 对多环芳烃的分类鉴别

王书涛*，刘　娜，程　琪，车先阁，李明珊，崔　凯，王玉田

燕山大学河北省测试计量技术与仪器重点实验室，河北 秦皇岛　066004

摘　要　多环芳烃(PAHs)作为一种芳香族化合物，普遍存在于人们的生产生活中，它具有强烈的致癌性，威胁着人们的生命和健康。所以，对多环芳烃实施简洁、高效、精确的检测方法很有必要。根据常见的多环芳烃类型，选取多环芳烃萘(NAP)、芴(FLU)、蒽(ANA)的固体粉末状物质作为实验样本。取 NAP，FLU 和 ANA 粉末各 1 g 溶于少量的甲醇(光谱级)溶液，然后转移到 100 mL 的去离子水溶液中，配置 PAHs 标准溶液。采用 FS920 荧光光谱仪，实验中为避免荧光光谱仪本身产生的瑞利散射影响，设置起始的发射波长滞后激发波长 10 nm。以标准溶液为基准，获取 ANA，NAP 和 FLU 单质的水溶液的荧光光谱图。在标准溶液的基础上，配置 0.1 mg·mL^{-1}的单质水溶液，然后将 ANA 与 NAP，FLU 分别取不同的体积相互混合形成两种混合溶液，各自形成 16 种不同浓度比例的混合溶液，再取不同体积的三种溶液相互混合，摇匀震荡，最后一共形成 48 种不同体积比例的混合溶液。最后将实验数据输入 Matlab 中得到蒽萘、蒽芴、蒽芴萘混合溶液的荧光光谱，发现混合溶液的激发波长在 260～320 nm、发射波长 300～380 nm 波长范围内，最佳发射波长的位置相似，荧光峰对应的激发波长有大部分重叠。针对荧光光谱不能直接辨别混合物的种类的不足，将基于遗传算法(GA)优化的支持向量机(SVM)应用于多环芳烃混合物种类的检测中，将数据随机打乱，并且将遗传算法的终止进化代数设为 200、训练数据和预测数据分别为 36 个和 12 个，得到训练结果的准确率为 95.42%。将实验结果对比分析普通支持向量机和 BP 神经网络，结果表明，基于遗传算法优化的支持向量机分类误差较小，能比较准确的分辨混合物的种类。

关键词　三维荧光光谱；遗传算法；支持向量机；多环芳烃

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2020)04-1149-07

引　言

随着现在生活水平质量的提高，人们现在更加关注自身身体的健康和生存环境的好坏，因为这与人类的未来息息相关。多环芳烃(polycyclic aromatic hydrocarbon，PAHs)作为一类长时间难以降解的污染物，大致上都具有一个以上的苯环[1]，广泛存在于大气、水中，长期积累会对人的呼吸系统或肝脏系统造成一定的损伤，更严重的可能会致癌。研究表明，有机物的不完全燃烧是导致 PAHs 的根本原因，一般可以分为两类[2]：一类为天然源，主要为一些像森林失火、火山爆发等自然灾害。而另一类为人们在生产生活中制造的一些烟气废料，产生的原因比较广泛，可以在石油燃烧和交通运输过程中大量产生，并通过干湿沉降、污泥农用和污水灌溉等方式在土壤或是水中不断积累，影响人们的生产和生活环境[3]。由于多环芳烃在环境中难以被分解，且持久性强[4]，因此国内外的学者对如何检测多环芳烃做了很多努力，以便找到能扼制多环芳烃对环境污染的方法。目前我国主要河流中都不同程度的受到 PAHs 的污染，刘小雪[5]通过对松花江干流沉积物中重金属和多环芳烃污染特征的检测，从空间和时间尺度上考察松花江沉积物中多环芳烃含量及其分布特征分析，了解到多环芳烃的组成多以 NAP，FLU 和 ANA 等比较常见的低环芳烃为主，而这三种物质荧光光谱会产生重叠[6]，容易融合，所以采用一种针对这三种物质混合物的快速鉴别的方法具有现实意义。

目前用于 PAHs 种类检测的方法主要有气相色谱法、高效液相色谱法、气相色谱-串联质谱法[1,7]等，但这些方法存在前期处理比较繁琐、价格昂贵、不能实时检测等缺陷。而 FS920 荧光光谱仪具有单次检测的成本低，操作简单，可以实时检测物质变化，具有丰富的信息含量等优点[8]，因此适

收稿日期：2019-02-28，**修订日期**：2019-06-04

基金项目：国家自然科学基金项目(61771419)和河北省自然科学基金项目(F2017203220)资助

作者简介：刘　娜，女，1995 年生，燕山大学河北省测试计量技术及仪器重点实验室硕士研究生　　e-mail：1124865942@qq.com

*通讯联系人　　e-mail：wangshutao@ysu.edu.cn

用于各种可以发出荧光物质的液体检测。Yang[9]等实现了用荧光光谱法中荧光激发发射矩阵检测水体质量的好坏，刘婷婷[10]等实现了用三维荧光光谱结合小波压缩与 APTLD 对水中多环芳烃的测定，吴兴[11]等实现了用平行因子结合支持向量机实现了对水中多环芳烃单质物质的检测，多环芳烃在自然界中多以混合物的形式存在，支持向量机作为一种比较好的分类模型，本文采用荧光光谱结合优化的支持向量机实现对多环芳烃混合物的分类与鉴别。

支持向量机(support vector machine，SVM)由 Vapnik 首先提出，最早可以用于模式分类，后来 Vapnik 将支持向量机加以改进并且适用于非线性回归上。在模式分类问题中，为避免因数据训练过程过于完美而造成的过拟合，支持向量机能提供较好的泛化性能。将遗传算法优化的支持向量机和三维荧光光谱技术相结合，可以快速准确的辨别混合多环芳烃物质中的种类。

1 基本原理

1.1 SVM 分类

支持向量机(SVM)作为单层感知机的一种延续和发展，区别在于感知机学习算法时会因采用的初值不同而得到不同的超平面，而 SVM 试图寻找一个最佳的超平面来划分数据。SVM 的主要思想是建立一个分类超平面作为决策面，使得需要被分离的物体之间的隔离边缘被极限的拉开；支持向量机的种类比较多：C-SVC，H-SVMs，DAG-SVMs(有向无环图支持向量机)等，其中 C-SVC 是比较常见的二分类支持向量机模型，其具体形式如下：

1)设已知训练集

$$T=\{(x_1, y_1), \cdots, (x_l, y_l)\} \in (X\times Y)^l \quad (1)$$

其中，$x_i \in X=R^n$，$y_i \in Y\{1, -1\}(i=1, 2, \cdots, l)$；$x_i$ 为特征向量。

2)选取适当的核函数 $K(x, x')$和适当的惩罚参数 C，构造并求解最优化问题

$$\min_a \frac{1}{2}\sum_{i=1}^{j}\sum_{j=1}^{l} y_i y_j a_i a_j K(x_i, x_j) - \sum_{j=1}^{l} a_j \quad (2)$$

$$\text{s. t.} \quad \sum_{i=1}^{l} y_i a_i = 0,\ 0 \leqslant a_i \leqslant C,\ i=1, \cdots, l \quad (3)$$

$$K(x_i, x_j) = \exp\left[-\frac{|x_i - x_j|^2}{g^2}\right] \quad (4)$$

得到最优解：$a^* = (a_1^*, \cdots, a_l^*)^{\mathrm{T}}$，其中 x 为空间中的点，g 为径向基半径。

3)选取 a^* 的一个正分量 $0<a_j^*<C$，并据此计算阈值：

$$b^* = y_j - \sum_{i=1}^{l} y_i a_i^* K(x_i - x_j) \quad (5)$$

4)构造决策函数

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} a_i^* y_i K(x, x_i) + b^*) \quad (6)$$

由于惩罚参数 C 与 g 的选择决定着 SVM 分类的准确率与精度，传统支持向量机需要靠经验寻找最佳参数，因此，使用 GA 参数优化方法来优化参数对比传统支持向机。

1.2 GA 优化 SVM 对参数的选择

遗传算法(genetic algorithm，GA)作为一种优化算法，是对达尔文生物进化论的自然选择和遗传学机理生物进化过程进行的模拟，在模拟自然进化过程搜索从而寻找最优解的方法，它最初由美国 Michign 大学 J. Holland 教授于 1975 年提出，根据自然界中优胜劣汰的选择规律，可以应用于很多领域，本文依据遗传算法来寻找最佳的支持向量机的参数，使训练和测试的结果达到最优。

遗传算法主要特点是直接对结构对象进行操作，操作简单，搜索范围大，应用范围广泛。运用 GA 来寻找最佳的参数 C 和 g，可以不必像网格划分那样遍历网格内的所有参数点，也能找到最优的答案[12]。遗传算法作为一种适用性很强的优化技术，近几年的发展极为迅速，掀起了一股遗传算法研究的热潮。利用 GA 算法对 SVM 参数选择优化的建模流程如图 1 所示。

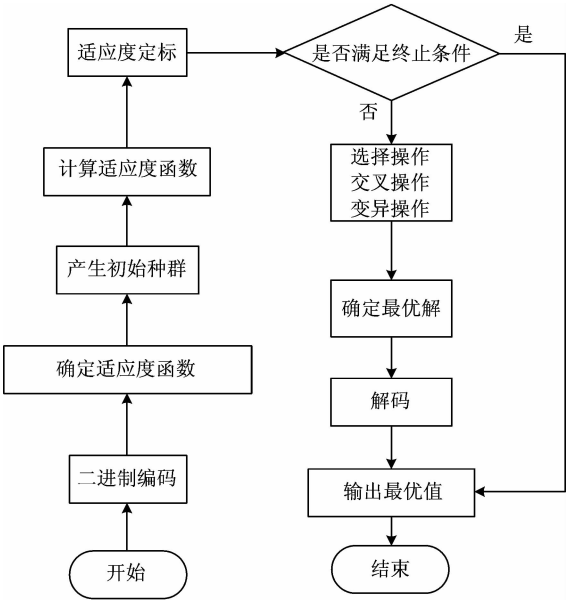

图 1 利用 GA 优化 SVM 参数的算法流程图

Fig. 1 GA optimization SVM parameters algorithm flow chart

2 实验部分

2.1 样品光谱的采集

多环芳烃的种类多种多样，例如苊烯(ANY)、芘(ANT)、荧蒽(FLT)等 27 种。本实验根据常见的多环芳烃类型，选取 3 种多环芳烃作为实验样本：萘(NAP)、芴(FLU)、苊(ANA)固体粉末状物质，购买自上海阿拉丁生化科技。称量仪器为天津天马横基仪器有限公司生产的 FA1004 型，取 NAP，FLU 和 ANA 粉末各 1 g 溶于少量的甲醇(光谱级)溶液，然后转移到 100 mL 的去离子水溶液中，配置 PAHs 标准溶液。实验过程中保证使用甲醇的浓度为 99%，体积分数低于 1%，避免在实验中对多环芳烃的测量造成影响。

在测量过程中，采用的检测仪器为英国 Edinburgh Instruments 公司生产的 FS920 荧光光谱仪，扫描范围为 200～

900 nm，比色皿为石英材质，光程 10 nm；实验中设置激发波长 200～370 nm，步长为 10 nm，发射波长为 240～390 nm，步长为 2 nm，狭缝宽度为 2.8 nm；为避免荧光光谱仪本身产生的瑞利散射影响，设置起始的发射波长滞后激发波长 10 nm。

为得到多环芳烃的原始光谱，将配置的 PAHs 标准溶液放入比色皿中进行测试，图 2 为实验中 10 g·L^{-1}多环芳烃单质水溶液的原始荧光光谱图。

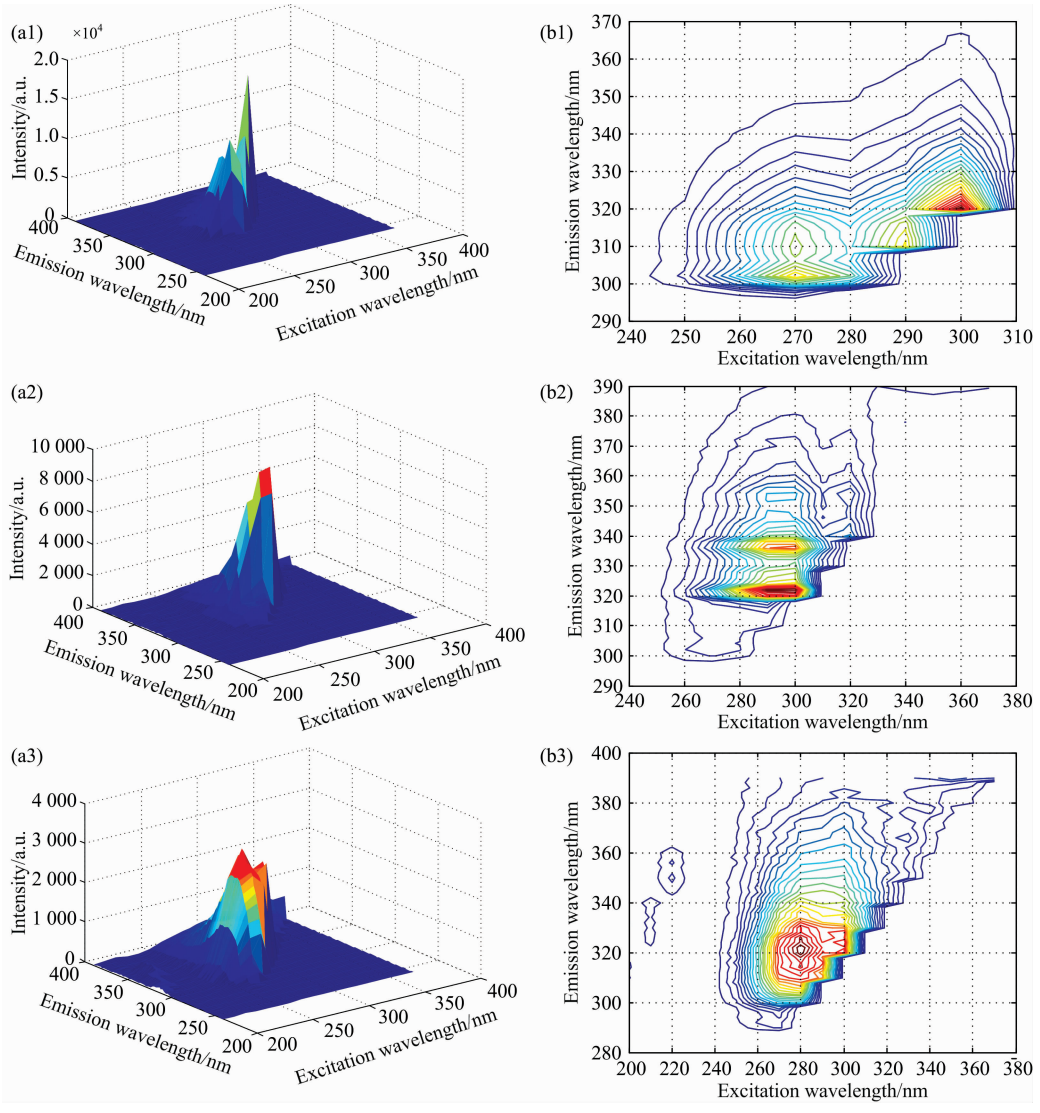

图 2　芴(FLU)、苊(ANA)、萘(NAP)的水溶液荧光光谱图

(a)：三维荧光光谱；(b)：溶液的等高线图

Fig. 2　Fluorescence spectra of FLU, ANA, NAP aqueous solution

(a): Three dimensional fluorescence; (b): Contour map of solution

由图 2 中三种多环芳烃的荧光光谱图可知，FLU 的荧光峰值位置在激发波长 300 nm，发射波长 322 nm，ANA 较强的荧光光谱范围是激发波长为 285～310 nm，发射波长为 320～340 nm，NAP 的荧光范围在激发波长 260～290 nm，发射波长在 310～330 nm 之间，而 NAP 的光谱范围较为广泛，涵盖了 ANA 的荧光光谱和 NAP 的荧光光谱，考虑到在自然界中，多环芳烃多为痕量物质不易被检测，而且以混合物的形式存在，以标准溶液为基准，配置了浓度为 0.1 mg·mL^{-1}的单质水溶液。实验将 ANA 与 NAP，FLU 分别取不同的体积相互混合形成两种混合溶液，各自形成 16 种不同浓度比例的混合溶液，然后再取不同体积的三种溶液相互混合，摇匀震荡，共形成 48 种不同比例的混合溶液。图 3 为不同体积分数混合溶液的部分荧光光谱图。

通过分析混合溶液的荧光光谱图可知，ANA、NAP 的混合溶液和 ANA、FLU 的混合溶液最佳发射波长的位置相同，荧光峰对应的激发波长也有大部分重叠，ANA，FLU 和 NAP 混合溶液的荧光峰也包括 320 nm，激发波长范围也和前两类混合物相近。荧光光谱范围集中在激发波长 260～300 nm，发射波长 320～360 nm 之间，仅从光谱图特性上并不能及时准确的辨别是哪种物质的混合物，因此，采用 GA 优化

的 SVM 算法来进行辨别，为提升水流域中多环芳烃类混合物种类的测量效果提供一种简单有效的方法。

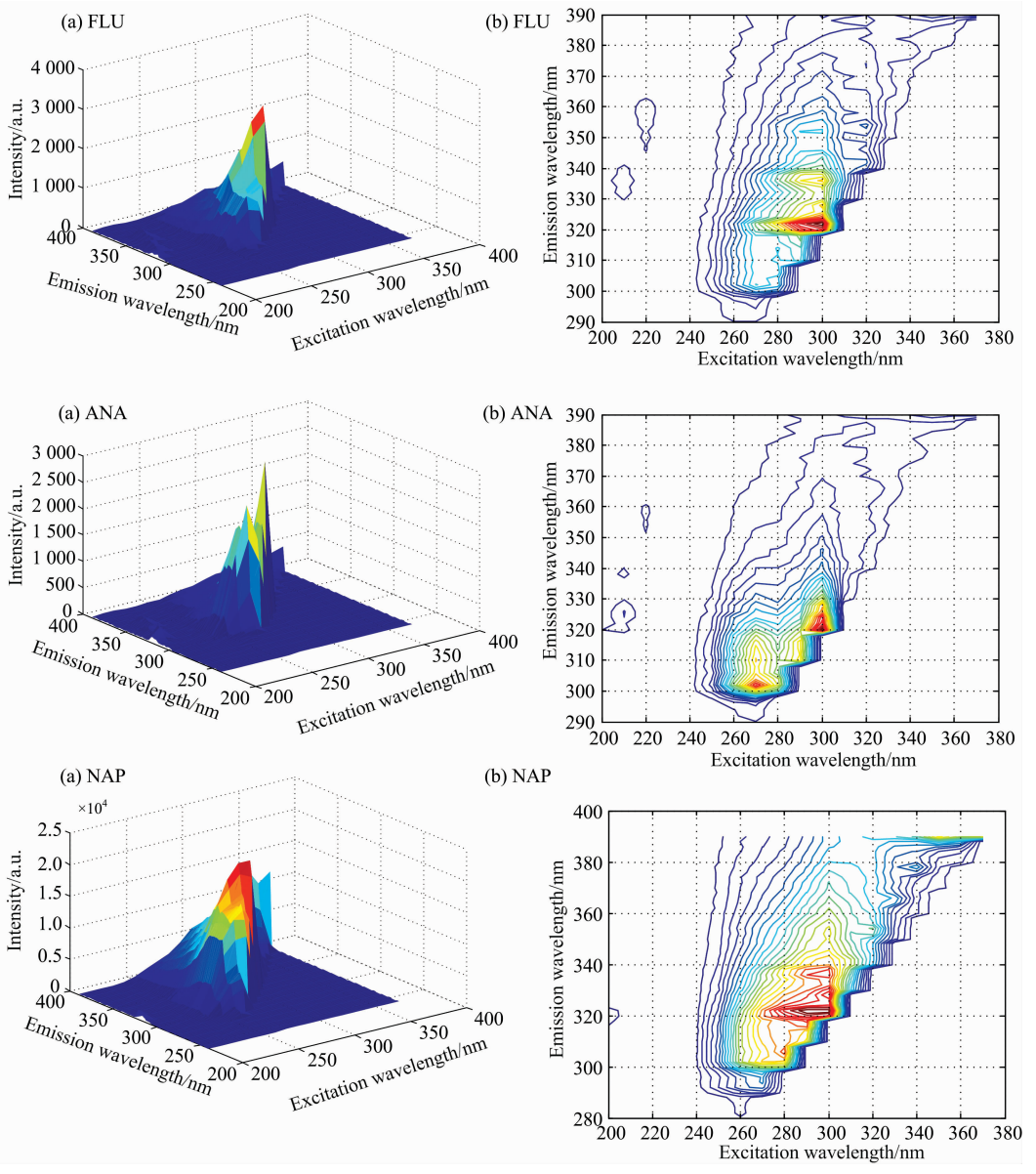

图 3　ANA∶NAP 体积比为 1∶9、ANA∶FLU 为 1∶9、ANA∶FLU∶NAP 为 2∶1∶3 的混合溶液

(a)：三维荧光光谱；(b)：溶液的等高线图

Fig. 3　A mixture with a volume ratio of 1∶9 for ANA∶NAP, A mixture with a volume ratio of 1∶9 for ANA∶FLU, A mixture with a volume ratio of 2∶1∶3 for ANA∶FLU∶NAP

(a): Three dimensional fluorescence; (b): Contour map of solution

2.2　GA-SVM 模型设计

GA-SVM 模型处理过程就是对数据进行的选择和重构矩阵，为了增加实验的准确性，避免人为因素造成的不确定因素，因此增大了荧光光谱的取值范围，取荧光光谱范围差别较大的波段：激发波长为 260～320 nm，发射波长为 300～380 nm，3 种蒽萘、芘芴和芘芴萘的混合溶液分别标定 1，2 和 3。每组 16 个样本，共 48 组，在 GA-SVM 模型中，将数据随机打乱，然后被分为训练组和预测组，训练组设为 36 组，预测组设为 12 组。其中，遗传算法需要提前设定的参数群体大小为 20、交叉概率为 0.9、变异概率为 0.01、遗传算法的终止进化代数为 200。

3　结果与讨论

将重构好的数据输入 GA-SVM 模型进行训练，模型经过 200 次迭代后，较好的实现了多环芳烃光谱的模式识别，训练好的模型测试结果如表 1 所示，共测试 10 次，取平均值。

表 1　多环芳烃的分类测试结果

Table 1　Classification test tesult of PAHs

Iteration number	Accuracy/%
200	95.42

从表 1 可知，GA-SVM 对三种多环芳烃混合物的识别率为 95.42%。这表明，GA-SVM 模型能准确识别不同种类的多环芳烃混合物三维荧光光谱。

3.1　GA-SVM 模型与 BP 神经网络对比

实验中设计得 GA-SVM 模型输入的训练集与测试集分别为随机取的 36 个和 12 个，所以，BP 神经网络的输入层神经元个数为 259(数据构成的为 259×24 矩阵)，隐层神经元个数为 5，输出节点个数为 3。激活函数为 sigmoid 函数，学习率取 0.1，为保证实验的可靠性，采用 36 个作为训练集，12 个作为测试集，训练次数为 100。训练好的模型平均测试结果如表 2 所示。

表 2　PAHs 光谱的分类测试结果

Table 2　Results of classification test of PAHs

Algorithm	Accuracy/%
GA-SVM	95.42
BP neural network	91.62

由表 2 可知，在对多环芳烃荧光光谱的分类中，GA-SVM 模型的光谱分类精度更高。

3.2　GA-SVM 模型与普通 SVM 的对比

实验中设计得 GA-SVM 模型输入的训练集与测试集分别为随机取的 36 个和 12 个，所以传统支持向量机的输入也为 36 个，将训练集和测试集的数据进行归一化处理后输入模型中，经过几次试验后，发现在惩罚参数 C 为 2，g 为 2 时，准确率最高，得出结果如图 4 所示。

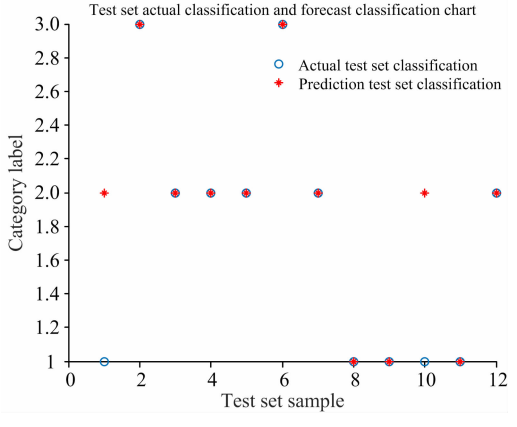

图 4　SVM 测试集的实际分类和预测分类图

Fig. 4　Test set actual classification and forecast classification chart for SVM

将传统 SVM 的测试结果与 GA-SVM 对比如图 5 所示。

由图 4 可知，SVM 的分类结果中，测试的 12 个分类样本中有 2 分类错误。而从图 5 中可以看到，GA-SVM 没有分类错误。说明 GA 的优化作用在寻找惩罚参数 C 和 g 中比人为寻找的更为准确、可靠。

GA-SVM 的适应度曲线如图 6 所示，从图中可以看到，当进化代数大于 8 时，最佳适应度值达到最大并在一定范围内震荡，保持稳定的波动并且与自家适应度之间的距离较小，总体收敛速度较快，适应度较好，能够快速检测多环芳烃种类。

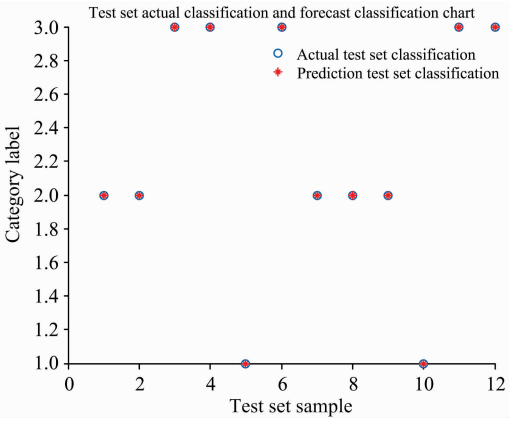

图 5　GA-SVM 测试集的实际分类和预测分类图

Fig. 5　Test set actual classification and forecast classification chart for GA-SVM

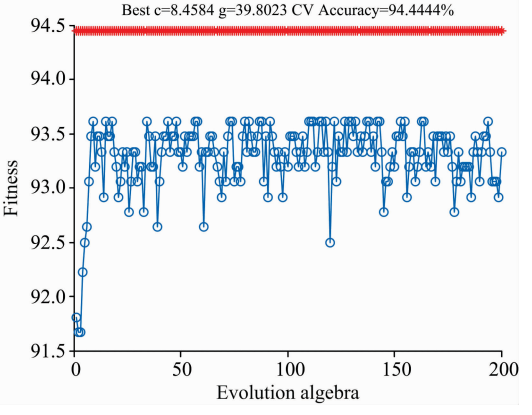

图 6　GA-SVM 参数的适应度曲线

Fig. 6　The fitness curve of particle swarm GA-SVM parameters

运行得到最优的惩罚因子 $C=8.46$，$g=39.80$。将得到的优化后的参数输入 GA-SVM 训练，运行模型训练，得到 ANA，FLU 和 NAP 的混合物的分类模型。

4　结　论

利用三维荧光光谱技术快速获取了 3 种多环芳烃混合物的荧光光谱，从光谱图特性中发现不同体积比例混合的多环芳烃单质物质，在激发波长在 260～320 nm、发射波长 300～380 nm 范围内发射波长位置相近，荧光峰对应的激发波长范围有大部分重叠，然后利用 GA-SVM 对不同种类的多环

芳烃进行分类，实验中，3 种混合物的平均分类正确率为 95.42%。实验结果表明，三维荧光光谱结合 GA-SVM 技术能准确识别不同种类的多环芳烃混合物，虽然这种方法在更多种类多环芳烃混合的情况下的运用有待进一步研究，但是为水流域中多环芳烃混合物的种类鉴别提供了一种新思路与新方法。

References

[1] SHI Xiao-feng, ZHANG Xin-min, YAN Xia, et al(史晓凤，张心敏，严 霞，等). Acta Optica Sinica(光学学报), 2018, 38(7): 0724001.

[2] Kang Yan, Xie Huijun, Li Bo, et al. Chemical Engineering Journal, 2019, 357: 280.

[3] Li Ruifei, Hua Pei, Zhang Jin, et al. Science of the Total Environment, 2019, 633: 438.

[4] Lin Yuanchung, Li Yaching, Kassian T T Amesho, et al. Science of the Total Environment, 2019, 660: 188.

[5] LIU Xiao-xue(刘小雪). Jilin University(吉林大学), 2016

[6] 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): 2212.

[7] Zhang Yanhao, Chen Yanyan, Li Ruijin, et al. Talanta, 2019, 195: 757.

[8] WANG Shu-tao, ZHENG Ya-nan, WANG Zhi-fang, et al(王书涛，郑亚楠，王志芳，等). Acta Photonica Sinica(光子学报), 2017, 46(9): 930002.

[9] Yang Y Z, Peleato N M, Legge R L, et al. Environmental Scieance-Water Research& Technology, 2019, 5(2):315..

[10] WANG Yu-tian, LIU Ting-ting, LIU Ling-fei, et al(王玉田，刘婷婷，刘凌妃，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(8): 2441.

[11] 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.

[12] Moura, Heloise O M A, Camara Anne B F, Santos Marfran C D, et al. Analytical and Bioanalytical Chemistry, 2019, 411(11): 2301.

Classification and Identification of Polycyclic Aromatic Hydrocarbons by Three-Dimensional Fluorescence Spectroscopy Combined with GA-SVM

WANG Shu-tao*, LIU Na, CHENG Qi, CHE Xian-ge, LI Ming-shan, CUI Kai, WANG Yu-tian

Measurement Technology and Instrument Key Lab of Hebei Province, Yanshan University, Qinhuangdao 066004, China

Abstract As an aromatic compound, Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous in human production and life. They have strong carcinogenicity and threaten human lives and health. Therefore, it is necessary to implement a simple, efficient, universal and accurate method to detect polycyclic aromatic hydrocarbons. According to the common types of polycyclic aromatic hydrocarbons, solid powdery substances of polycyclic aromatic hydrocarbon naphthalene (NAP), fluorene (FLU) and acenaptene (ANA) were selected as experimental samples. Of all the samples NAP, FLU, ANA powder were carried out by 1 g and dissolved in a small amount of methanol (spectral grade) solution, then transferred them to 100 mL of deionized water solution, getting a configure PAHs standard solution. The experiment was carried out by the FS920 fluorescence spectrometer. In order to avoid the Rayleigh scattering effect generated by the fluorescence spectrometer itself, the initial emission wavelength was set to lag the excitation wavelength by 10 nm. It could obtain the fluorescence spectrum of the aqueous solution of ANA, NAP and FLU, on the basis of the standard solution, a 0.1 $\mu g \cdot mL^{-1}$ aqueous solution of a simple substance was placed. Then, different volumes of ANA, NAP and FLU were mixed to form two mixed solutions, each of them formed a mixed solution of 16 different concentration ratios, and then took different volumes. The three solutions were mixed with each other, they were shaken and finally a total of 48 mixed solutions of different volume ratios were formed. Finally, the experimental data were input into Matlab to obtain the fluorescence spectrum of the mixed solution of naphthalene, anthracene and anthracene naphthalene. It was found that the excitation wavelength of the mixed solution was in the wavelength range of 260～320 nm and the emission wavelength was 300～380 nm, and the position of the optimal emission wavelength was similar. Most of the excitation wavelengths corresponding to the fluorescence peaks overlap. Support vector machine (SVM) based on genetic algorithm (GA) optimization was applied to the species detection of PAHs mixture, because the shortage of species in which the fluorescence spectrum cannot directly react with the mixture solution. The data were randomly scrambled, and the genetic evolutionary algorithm havd a termination evolution algebra of 200. Training data and prediction data are 36 and 12, respectively. Under the optimal condi-

tions, the average accuracy of the training result was 95.42%. The experimental results were evaluated by comparing with traditional support vector machine and BP neural network. The results showed that SVM based on genetic algorithm optimization has potential for the smaller classification error and can distinguish the mixture more accurately.

Keywords Three-dimensional fluorescence spectroscopy; Genetic algorithnm; Support vector machine; Polycyclic aromatic hydrocarbons (PAHs)

(Received Feb. 28, 2019; accepted Jun. 4, 2019)

* Corresponding author