Journal of Advances in Physical Chemistry 物理化学进展, 2023, 12(1), 1-12
Published Online February 2023 in Hans. https://www.hanspub.org/journal/japc
https://doi.org/10.12677/japc.2023.121001

Hans汉斯

# 基于太赫兹超材料和机器学习的挥发性有机物分类

付文凤，曹红燕，马　毅，陈　麟*

上海理工大学，上海

收稿日期：2022年12月30日；录用日期：2023年2月17日；发布日期：2023年2月27日

## 摘　要

**挥发性有机物(Volatile organic compounds, VOCs)作为高速发展的科技时代的副产品，存在来源广、范围大、危害多和处理难的问题。VOCs检测时存在采样方法操作复杂、检测时间长、成本高等问题。挥发性有机物的快速精准检测，对及时采取措施，减少其对危害环境空气健康，提高人类生活质量方面有很大帮助。在这项工作中，选用异丙醇、乙苯、乙酸乙酯三种VOCs的研究对象；通过腔体结构营造密闭环境减少有机物挥发对实验的影响；采用具有Fano共振的超表面芯片用于测量了三种挥发性有机物在土壤中不同微含量时太赫兹查表面相互作用后的透射光谱的频移和强度。此外，根据变化趋势选择不同的拟合函数建立单变量回归模型，表明三种挥发性有机物具有明显的不同特性。研究表明，支持向量机(SVM)对三种VOCs的分辨准确率达到96.7%，而基于主成分分析的高斯混合模型(PCA-GMM)分类可视化算法，对于微量的检测物质，PCA-GMM在分类可视化可实现在95%置信区间内实现有效分离。**

## 关键词

**太赫兹，超表面芯片，有机物，SVM，PCA-GMM**

# Classification of Volatile Organic Compounds Based on Terahertz Metamaterials and Machine Learning

**Wenfeng Fu, Hongyan Cao, Yi Ma, Lin Chen***

University of Shanghai for Science and Technology, Shanghai

Received: Dec. 30th, 2022; accepted: Feb. 17th, 2023; published: Feb. 27th, 2023

*通讯作者。

文章引用: 付文凤, 曹红燕, 马毅, 陈麟. 基于太赫兹超材料和机器学习的挥发性有机物分类[J]. 物理化学进展, 2023, 12(1): 1-12. DOI: 10.12677/japc.2023.121001

## Abstract

**Volatile organic compounds (VOCs), as by-products of the rapidly developing scientific and technological era, have problems of wide source, wide range, harmful to health and difficult to treatment. VOCs detection has the problems of complex sampling method, long detection time and high cost. Rapid and accurate detection of volatile organic compounds is of great help to take timely measures to reduce its harm to environmental air health and improve the quality of human life. In this work, three VOCs of isopropanol (IPA), ethyl benzene (EB) and ethyl acetate (EA) were selected. The cavity structure creates a closed environment to reduce the influence of organic volatilization on the experiment. A metasurface chip with Fano resonance was used to measure the frequency shift and intensity of transmission spectrum after terahertz surface interaction of three volatile organic compounds with different micro-contents in soil. In addition, the univariate regression model was established by selecting different fitting functions according to the variation trend, which showed that the three volatile organic compounds had obviously different characteristics. The results show that support vector machine (SVM) can distinguish the three VOCs with 96.7% accuracy, and the Gaussian mixture model (PCA-GMM) classification visualization algorithm based on principal component analysis can realize effective separation within 95% confidence interval for trace detected substances in classification visualization.**

## Keywords

**Terahertz, Meta-Surface Chips, Organics, SVM, PCA-GMM**



## 1. 引言

挥发性有机物(Volatile organic compounds, VOCs)作为工业迅速发展的副产品，危害环境空气健康，降低人类生活质量[1]。VOCs 是指常态下易挥发的有机化合物，主要包括非甲烷烃类，含氧有机物，含氮有机物，含硫有机物等，具体有烷烃、烯烃、芳香烃、卤代烃、醇、酮、醛、醚、酯、硫化物及杂环等[2]。VOCs 涉及行业众多，如机动车制造于维修、涂料生产、家具、家用电器、金属制品加工、彩钢板、集装箱、造船、电器设备印刷等[3]。VOCs 检测存在易挥发、难定量、不易监督的问题，众多工厂为降低成本，违规排放。土壤中 VOCs 的主要来源是人类活动，如石化、化工、交通、燃烧、居民生活等[4]。土壤复杂的三相共存体系和 VOCs 的吸附性导致 VOCs 更易隐藏，不能被及时发现[5]。目前，VOCs 治理难问题主要体现在以下几方面：VOCs 来源广，大至工厂排放，小至办公用品等都是元凶；范围大，仅仅化工企业关闭搬迁造成的废弃地面积远超 50 万公顷；危害多，生态环境危害和人体健康损伤。VOCs 常用的采样方法有袋采样，罐采样，吸附剂采样和固相萃取技术等[6]。常规的挥发性有机物检测方法有：气相色谱法，液相色谱法等[7]。

太赫兹(Terahertz, THz)指位于 0.1~10 THz 范围内的电磁波，介于微波和红外之间，具有光子能量低、波长短、穿透性好等特点，同时依靠氢键和范德华力等分子内和分子间振动可以激发分子对太赫兹波的吸收，传递丰富的信息[8]。太赫兹检测成为检测的新型“武器”，引起了众多领域研究者的关注。然而，如果样品的厚度远小于太赫兹的波长时，光与物质相互作用将会很弱，导致不同 VOC 液体之间的太赫兹

光谱特性差别不大，这也意味着三种挥发性有机物(含氧、含氮及含硫)溶液之间的太赫兹吸收差异不再显著。此外，由于极性液体吸收较强，挥发性有机物之间的吸收系数差异并不显著。增强光－物质相互作用需要增强光与物质的耦合作用。目前，增强光－物质的耦合作用依靠超材料传感器，如吸收器，石墨烯，谐振环等方法[9] [10] [11]。也就是说，通过对电磁场的强限制来增强太赫兹区域内的光－物质相互作用的技术是非常可取的。

在本文中，我们通过研究三种在土壤中的挥发性有机物：异丙醇(isopropanol, IPA)、乙酸乙酯(ethyl acetate, EA)和乙苯(ethyl benzene, EB)，探讨了太赫兹超表面芯片在腔体中的应用。超表面结构由带缺口的圆环构成，电场主要集中在开口间隙，Fano 共振处的大强度表面电流有助于增强所提超表面中开口间隙处的电场限制，从而增强光与物质之间的相互作用，同时提高了超表面的灵敏度。腔体结构达到密封条件，超材料进行环境检测，二者相互辅助实现对土壤中三种挥发性物质异丙醇、乙酸乙酯、乙苯含量的检测。分析三种物质不同含量时太赫兹透射光谱的变化，再结合数据分析拟合与支持向量机(Support Vector Machine, SVM)分类算法 96.7%的分类效果。基于主成分分析的高斯混合模型(Gaussian Mixture Model based on Principal Component Analysis, PCA-GMM)算法对三种物质达到 95%的置信度，而且可由等高线判断各样点在区域内出现的概率。本工作对扩大太赫兹检测器件的在污染物检测应用领域，有效鉴别挥发性有机物种类和排放含量，对工业区排放标准不达标的工厂及时进行溯源及责任追究，减少有机物排放对土壤的损伤有一定帮助。

## 2. 实验方法与进程

### 2.1. 化学试剂选购及样品制备

实验所用的化学试剂乙醇(货号 E111977)、异丙醇(货号 I292350)、乙酸乙酯(货号 E116138)、乙苯(货号 E117365)购置于阿拉丁公司，按照药品使用说明妥善保存。具体的实验样本制备方式如下：从校园教学楼前的花坛获取实验中的土壤；在烘干箱(博迅 BG2-30)以 100℃烘两小时，干燥土壤，并排除土壤中微生物对实验的影响；将烘干后的土壤放入玛瑙岩体中研磨成小粉末，筛出小于 60 目的土壤颗粒；制作的分析样品分为三大组，每组有 60 个，每组分为 6 小组，每个小组有 10 个，共 180 个分析样品。每个分析样品含有 10 g 土壤颗粒，每小组中的分析样品用滴定管分别加入 1、2、3、4、5、6 μL 的异丙醇、乙酸乙酯、乙苯，用凡士林和聚四氟乙烯膜封口，减少挥发，放置 24 小时。根据相似相溶原理，加入乙醇提取液，加热水浴振荡 6 小时，将土壤中的三种有机物溶解在乙醇中。处理完毕后，通过离心机，在 30,000 r/min 的离心速度下离心 10 min 实现固液分离。静置后取出上清液放置在试管中，依旧用凡士林和聚四氟乙烯膜封口，防止挥发。

### 2.2. 实验装置与操作

如图 1(a)所示，实验检测装置为日本爱德万(ADVANTEST)测试公司的 TAS7500SU 系统(Tera-hertz Spectroscopic System 7500SU)。该装置是一台桌面太赫兹脉冲波光谱测量与分析系统，采用 Cherenkov 太赫兹源，有效光谱范围为 0.5~7 THz。它不仅可以测量固体和液体样本，而且可以更换模块选择透射、反射、衰减全反射三种测量模式，空气净化和去除水蒸气干扰能力强，动态范围约为 50 dB。

实验中所用样品均为易挥发物质，选用腔体结构作为容器，有效避免有机物挥发对实验结果造成影响。该腔体结构由介电常数 1.88 的聚四氟乙烯制成，设置高 50 μm 的中空结构放置超表面芯片和添加待测液体，如图 1(b)和图 1(c)所示。

在采样时间为 8 ms，分辨率为 7.9 GHz 的情况下，对 1024 个采样点累计平均得到测量结果。实验操

作过程对每个分析样品进行测试，用乙醇清洗超表面芯片和腔体结构后，去离子水超声波震荡，取出后用无尘布擦干，避免上次测量残留物影响实验效果。实验在室温下进行，为避免水汽干扰，环境湿度控超表面芯片表面不添加任何物质，测得的时域信号经过傅里叶变换，可表示为公式(1)

$$\begin{aligned} E_r(\omega) &= \int E_r(t)\exp(-i\omega t)\mathrm{d}t \\ &= M_r(\omega)\left|\exp(-i\varphi(\omega))\right| \end{aligned} \tag{1}$$

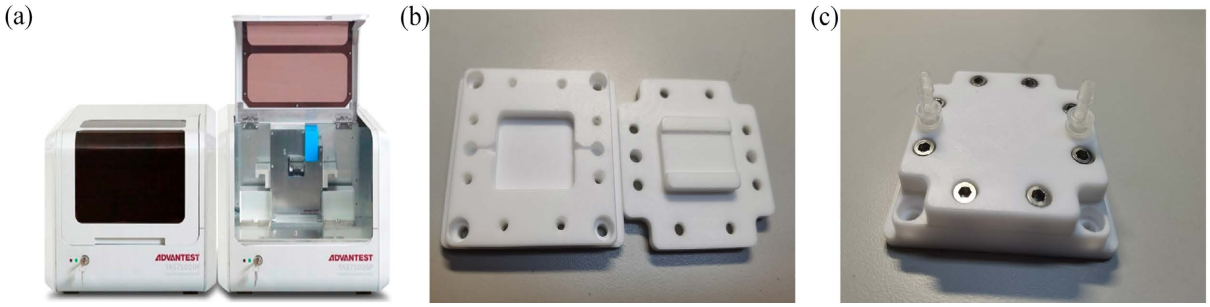


**Figure 1.** Main apparatus in the experiment. (a) TAS7500SU; (b) (c) Cavity structure
**图 1.** 主要实验装置图。(a) TAS7500SU；(b) (c) 谐振腔

当超表面芯片表面添加待测物质时，测得的时域信号经过傅里叶变换，可表示为公式(2)

$$E_s(\omega) = \int E_s(t)\exp(-i\omega t)\mathrm{d}t = M_s(\omega)\left|\exp(-i\varphi(\omega))\right| \tag{2}$$

这里，$E_r$ 是参考信号电场，$M_r$ 是参考信号幅值强度，$E_s(\omega)$ 是样本信号电场，$M_s$ 是参考信号幅值，$t$ 是时间，$\varphi$ 是相位，$\omega$ 是频率。

由于异丙醇、乙酸乙酯、乙苯在太赫兹波段没有明显的吸收峰，超表面芯片在太赫兹光谱下形成的两个吸收峰是本文的主要着眼点，将添加两个待测物峰值点的频率偏移量记为 $\Delta f_1$，和 $\Delta f_2$，强度变化量表示为 $\Delta M_1$ 和 $\Delta M_2$。

## 3. 实验结果分析

利用 Matlab2016 编写程序对 180 组数据的两个频点峰值和强度进行批量提取，并求出异丙醇，乙酸乙酯，乙苯三种物质同一物质的量测量的十组数据频率偏移量 $\Delta f_1$ 和 $\Delta f_2$，强度变化量 $\Delta M_1$ 和 $\Delta M_2$，分别平均后得到的平均值作为这一物质的量下的有效值，减少实验过程中可能存在的误差对实验结果的影响。最后，三种物质的频率偏移量 $\Delta f_1$，和 $\Delta f_2$ 和强度变化量 $\Delta M_1$ 和 $\Delta M_2$ 与其物质的量分别采用适宜的拟合方式拟合，并标出误差条。此后，利用 Classification Learner 计算出多种常用分类算法的准确率，混淆矩阵以及 ROC 和 AUC 值，评估最佳算法后进行进一步优化，实现物质的定性分析，并结合前期的拟合结果，实现对物质的定量求解。

### 3.1. 超材料芯片

采用 CST STUDIODU SUITE (2017 版)电磁仿真软件设计了一种基于狄拉克半金属薄膜的 Fano 响应的超表面设计，单元结构示意图如图 2 所示。超表面结构由两部分组成：1) 聚酰亚胺衬底，介电常数为 3.5 + 0.0027i，厚度 $h$ = 25 μm；2)表面不对称开口谐振环，材质为铝，电导率为 $3.56 \times 10^7$ S/m，厚度为 $t$ = 200 nm。具体的结构参数:单元结构长 $P_x/2$ = 70 μm，宽 $P_y/2$ = 70 μm，圆环外环半径 $R$ = 30 μm，内环半径 $r$ = 24 μm，圆环宽度 $G$ = 3 μm，开口中心距圆心距离，即不对称参数 $d$ = 13 μm。

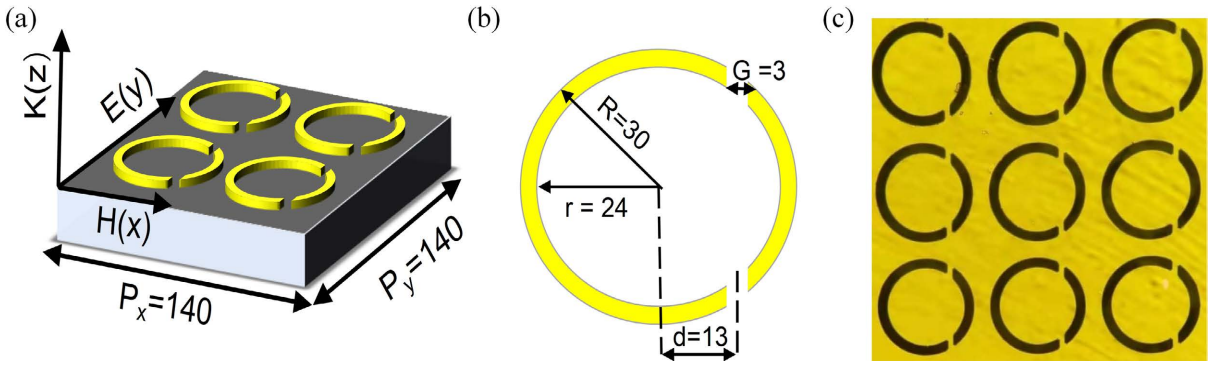


**Figure 2.** Schematic of Fano super-surface structure. (a) Stereogram; (b) Plan; (c) Optical microscopy 500×
**图 2.** 超表面芯片结构。(a) 示意图；(b) 平面图；(c) 光学显微镜 500×放大图

### 3.2. 单变量分析结果

为定量测定 18 个样品(1~6 μL/10g 土壤)中的微量有机物(IPA、EA、EB)，分析了不同含量有机物对 Fano 超表面芯片传感器的光谱频移Δf 和吸收强度变化(Δ*M*)，结果如图 3(a)~(c)所示。随着 VOC 含量的增加，频率发生红移，幅值也明显变化，建立光谱的单变量回归模型。IPA 的频移采用 ExpGro2 Fit 方式拟合，拟合情况如图 3(d)所示：在 0.885 THz，$R^2$ = 0.963，Δ*f* 位于 0.05~0.10 THz；在 1.778 THz，$R^2$ = 0.963，Δ*f* 位于 0.07~0.18 THz。EA 的频移采用 Asymptotic1 Fit 方式拟合，拟合情况如图 3(e)所示：在 0.885 THz，$R^2$ = 0.927，Δ*f* 位于 0.04~0.10 THz，在 1.778 THz，$R^2$ = 0.974，Δ*f* 位于 0.06~0.16 THz。EB 的频移采用 Polynomial Fit 方式拟合，拟合情况如图 3(f)所示：在 0.885 THz，$R^2$ = 0.920，Δ*f* 位于 0.08~0.10 THz；在 1.778 THz，$R^2$ = 0.865，Δ*f* 位于 0.13~0.18 THz。IPA 的幅值变化采用 ExpGro2 Fit 方式拟合，拟合情况如图 3(g)所示：在 0.885THz，$R^2$ = 0.982，Δ*M* 位于 0.10~0.18 (a.u.)；在 1.778 THz，$R^2$ = 0.983，Δ*M* 位于 0.08~0.16 (a.u.)。EA 的幅值变化采用 Asymptotic1 Fit 方式拟合，拟合情况如图 3(h)所示：在 0.885 THz 时，$R^2$ = 0.950，Δ*M* 位于 0.12~0.18 (a.u.)；在 1.778 THz 时，$R^2$ = 0.944，Δ*M* 位于 0.08~0.14 THz。EA 的幅值变化采用 Linear Fit 方式拟合，拟合情况如图 3(h)所示：在 0.885 THz 时，$R^2$ = 0.973，Δ*M* 位于 0.06~0.14 (a.u.)；在 1.778 THz 时，$R^2$ = 0.869，Δ*M* 位于 0.04~0.10 (a.u.)。

综合 IPA，EA，EB 的Δ*f* 拟合曲线和Δ*M* 拟合曲线发现，当土壤中添加较少 VOC 时，土壤中残留的 VOC 占添加总量的比例较大，更易对实验结果造成影响，导致误差条偏大。但是Δ*f* 和Δ*M* 的拟合度极高，而 IPA、EA、EB 的拟合方式不同，所以有希望通过Δ*f* 和Δ*M* 拟合关系构建 IPA、EA 和 EB 特定的“后天指纹”，并通过“后天指纹”进行不同物质的定性鉴别。

### 3.3. 算法筛选

算法性能好坏需要量化的数据进行评估，本节选用混淆矩阵，准确率，ROC 和 AUC，混淆矩阵是将分类预测模型的预测结果集中记录并展现的表格，因此常用作机器学习的辅助工具。多分类任务的混淆矩阵可用公式(3)描述

$$\text{Confusion Matrix} = \begin{bmatrix} N_{1,1} & N_{1,2} & \cdots & N_{1,n} \\ N_{2,1} & N_{2,2} & \cdots & N_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ N_{n,1} & N_{n,2} & \cdots & N_{n,n} \end{bmatrix} \tag{3}$$

即假设有 $n$ 类物质，可分别用 $a_1, a_2, \cdots, a_n$，$N_{i,i}$ 表示实际与预测相符的样本量，而 $N(i \neq j)$ 表示预

测时将实际属于 $a_i$ 的归为 $a_j$ 的样本量。

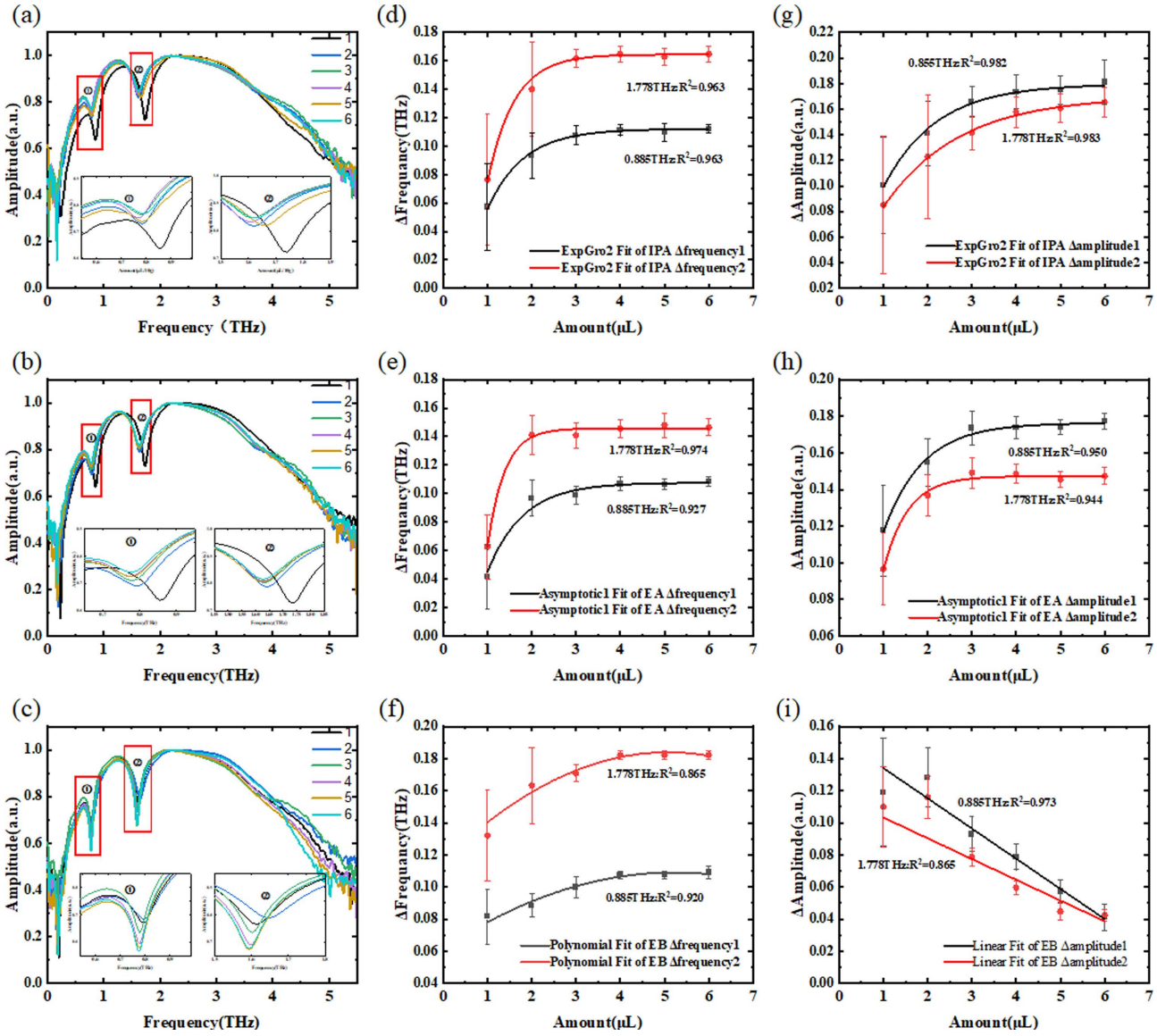


**Figure 3.** Univariate fit results. (a)~(c) The THz Spectra, (d)~(f) frequency shift response and (g)~(i) amplitude shift response corresponding to 6 VOC concentration used for the detection of IPA, EA, EB, respectively

**图 3.** 单变量拟合的结果：(a)~(c)分别用于检测 IPA、EA、EB 的 6 个 VOC 浓度对应的 THz 光谱；(d)~(f) 频移响应和(g)~(i)振幅位移响应

混淆矩阵中，实际数值分为 positive 和 negative，预测结果也分为 positive 和 negative，得到四个基础指标:True Positive (TP)、False Negative (FN)、False Positive (FP)、True Negative (TN)。TP 表示正确检测的样品数；FP 表示错误但检测正确的样品数；FN 表示错误检测为错误的样品数；TN 表示错误检测为正确的样品数。

准确率被正确分类的结果占比概率，可以用公式(4)描述：

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{4}$$

检测率(True Positive Rate)表示可识别的 TP 与正实例的比例，可以用公式(5)描述：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

误检率(False Positive Rate)表示识别的 FP 与负实例的比例，可以用公式(6)描述：

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

Receiver operating characteristic curve (ROC)是验证概率模型质量的重要手段，ROC 曲线上的点代表这不同阈值下的分类效果，横坐标表示 FPR，纵坐标表示 TPR，而 the area under the curve (AUC)是指 ROC 曲线与 x 轴所围的面积，在分类器的性能评估中起主导作用，AUC 的值位于 0~1 之间，且 AUC 越接近于 1，分类器的性能越好。首先，选择频域信号中的共振峰 1 的频点和强度值，共振峰 2 的频点的和强度值，土壤中 EB、IPA、EA 的含量五个变量作为原始自变量，物质分类作为因变量，构成一个 180 × 6 的矩阵 X。

**Table 1.** ACC comparison of kinds of classification
**表 1.** 各种分类器的比较

| 分类器 | | 顺序 | 准确度 |
|---|---|---|---|
| Tree | Complex | 1.1 | 86.7% |
| | Medium | 1.2 | 86.7% |
| | Simple | 1.3 | 80.6% |
| Discriminant | Linear | 1.4 | 60.0% |
| | Quadratic | 1.5 | 65.0% |
| SVM | Linear | 1.6 | 60.0% |
| | Quadratic | 1.7 | 88.9% |
| | Cubic | 1.8 | 91.7% |
| | Fine Gaussian | 1.9 | 91.7% |
| | Medium Gaussian | 1.10 | 84.4% |
| | Coarse Gaussian | 1.11 | 60.6% |
| KNN | Fine | 1.12 | 91.7% |
| | Medium | 1.13 | 77.2% |
| | Coarse | 1.14 | 58.9% |
| | Cosine | 1.15 | 72.2% |
| | Cubic | 1.16 | 72.8% |
| | Weighted | 1.17 | 89.4% |
| Ensemble | Boosted Trees | 1.18 | 66.7% |
| | Bagged Trees | 1.19 | 90.0% |
| | Subspace Discriminant | 1.20 | 62.2% |
| | Subspace KNN | 1.21 | 70.0% |
| | RUS Boosted Trees | 1.22 | 66.1% |

分类预测算法分为非监督学习和监督学习，本文中先采用 Matlab2016 中的 Classification Learner App 对 180 组数据进行监督学习预测，初步筛选效果较好的算法进行优化。监督学习的过程中给 180 组数据按照 5:1 的比例分配到测试集和验证集，即测试集为 150 组，验证集 30 组，采用五折交叉验证，具体涉及的监督学习方法有决策树(Decision Tree, DC)、判别分析(Discriminant Analysis, DA)、支持向量机(Support Vector Machines, SVM)、K 最邻近算法(K-Nearest Neighbor, KNN)、集成算法(Ensemble)五大类，共计 22 个小分类。将 ACC 罗列在表 1，预测准确率最高的分类算法为 Cubic SVM (Model 1.8)、Fine Gaussian SVM (Model 1.9)和 Fine KNN (Model 1.12)，最高预测准确率为 91.7%，预测准确率最低的分类算法为 Linear Discriminant (Model 1.4)和 Linear SVM (Model 1.6)，最低预测准确率为 60%。

混淆矩阵是评估分类器优劣的较可信的指标，选择准确度较高的六种分类模型绘制出混淆矩阵进行综合评估，如图 4 所示：混淆矩阵表中，1 代表乙苯，2 代表异丙醇，3 代表乙酸乙酯。图 4(a)中 Complex Tree (Model 1.1)第一行中出现准确度和为 101%的情况，是由于再计算准确率的时候四舍五入造成的，每一个三维混淆矩阵中对角线位置处代表某一类物质的预测正确率，对比六个模型的混淆矩阵发现，只有图 4(c)中 Cubic SVM (Model 1.8)对三种物质的预测正确率都在 90%以上，而图 4(e)中 Fine KNN(Model 1.12)虽然对乙苯的预测正确率高达 100%，但是对异丙醇的预测正确率只有 83%。图 4(b)中 Quadratic Discriminant (Model 1.5)可信度极低，Cubic SVM (Model 1.8)较高且稳定的正确率在六个模型中崭露头角。

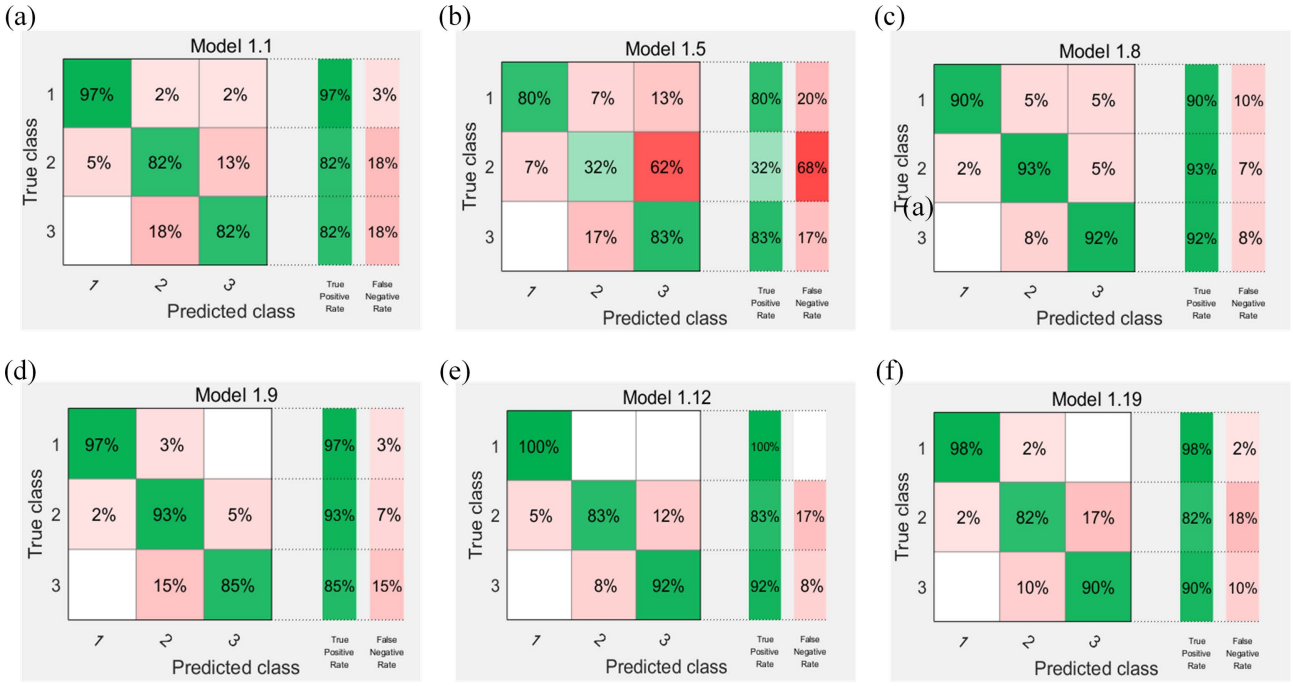


**Figure 4.** Confusion matrix of better classifiers
**图 4.** 较优分类器的混淆算法

六种模型的 AUC 值均在 0.90 以上，图 5(b)中 Quadratic Discriminant (Model 1.5)的表现最差，仅有 0.90，图 5(d) Fine Gaussian SVM (Model 1.9)和图 5(f) Bagged Trees Ensemble (Model 1.19)表现最好，高达 1.00，接近完美，其余三种模型均为 0.99，除了 Quadratic Discriminant (Model 1.5)外，其他模型的 AUC 值不相上下，且均呈现出较高的可靠性。

综合模型的准确率、混淆矩阵和 ROC 曲线，可以明显地比较出 Cubic SVM (Model 1.8)在多方面的表

现都很优异，准确率 91.7%，对 IPA、EA 和 EB 的预测正确率分别为 90%、93%和 92%，非线性 SVM 模型确定为针对本分类问题的最优算法，加以改进，以得到更优越的表现。

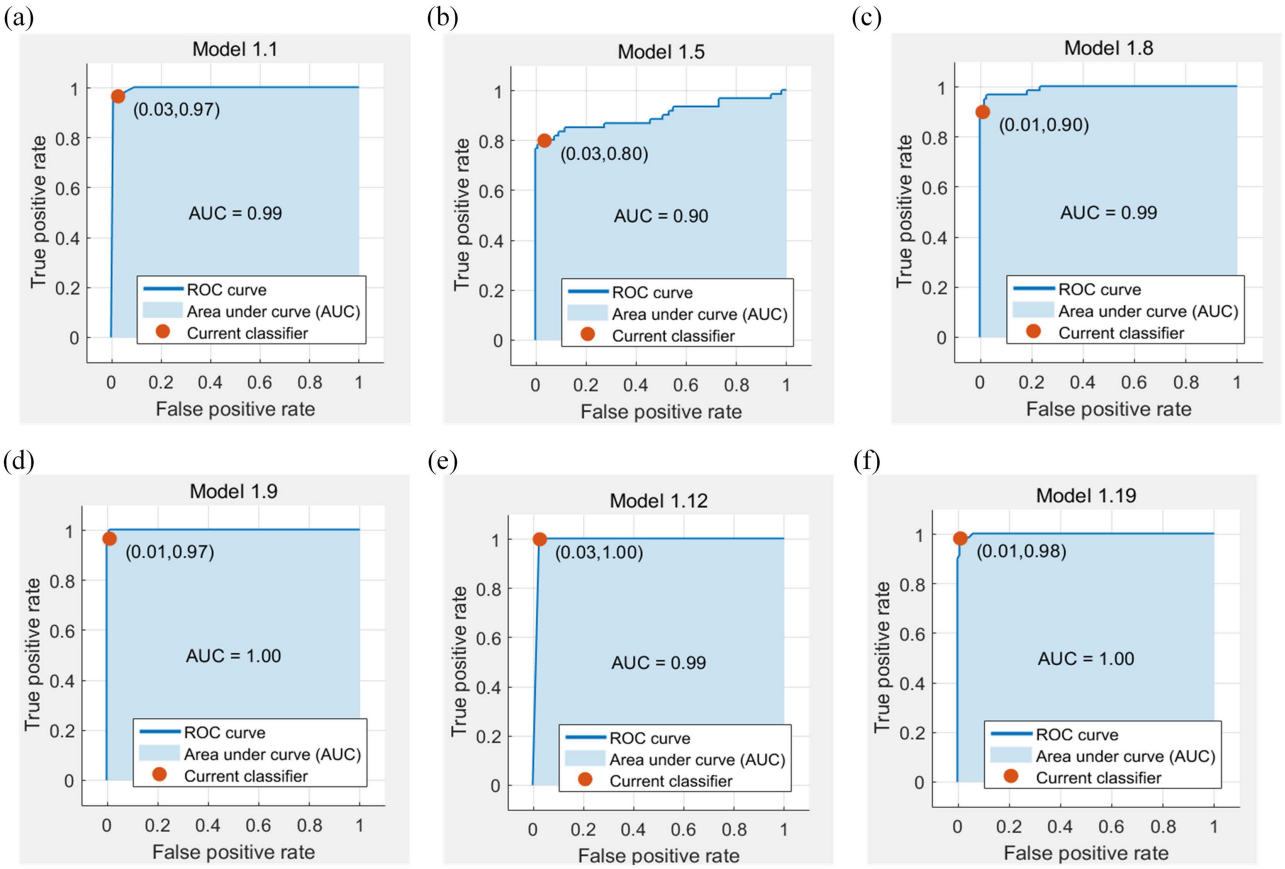


**Figure 5.** ROC and AUC of better classifiers
**图 5.** 较优分类器的 ROC 和 AUC 值

SVM 是在统计方法的基础上寻找分离两类的最优超平面[12]。假设初始因素变量为 $X=\left(X_1,X_2,\cdots,X_n\right)$，分类变量(EB, IPA, EA)为 $Y_j=\left(Y_1,Y_2,Y_3\right)$，可用公式(6)表示：

$$f\left(X\right)=sign\left[\sum_{i=1}^{n}\alpha_i Y_j K\left(X,X_i\right)+c\right] \tag{7}$$

$c$ 为超平面原点的偏移量，$n$ 为影响分类结果的因素量，$\alpha_i$ 是正实数，$Y_j K\left(X,X_i\right)$ 为核函数，核函数可替换目标函数与分类决策函数中实例间的内积，得到非线性支持向量机，使 SVM 可应用高维特征空间，核函数类型如表 2 所示。

核函数使用最广泛的是多项式核和高斯核，其中的 $d$ 为多项式阶数，$\gamma$ 是高斯核宽度的预定义参数。经过探索将 SVM 的设置为多分类 C-SVC 模式，核函数采用 RBF，在五折交叉验证中，如图 6 所示，AUC 值达到 0.99，预测结果仅出现一例 IPA 预测为 EA 的情况，预测正确率达到 96.7%。

PCA 是数理统计中常用的降低数据维度的方法[13]，将降低维度后的结果用作协变量，可减少各原始变量的之间的关联造成的假阳性，同时可通过置信区间判断各种类之间的相似性。

如图 7 所示，通过 PCA 算法，将 5 个原始变量转化为两个主要影响变量 PC1 和 PC2，PC1 占比 54.5%，PC2 占比 28.0%，两者的和为 82.2%，可有效代替原始变量。图中圆圈内的区域代表不同种类在 95%的

置信区间时的分布，可见 EB 与 IPA 和 EA 的相似性相对较低，更易区分，而 IPA 和 EA 的分布高度重合，在实际分类时容易犯错，仅仅依靠 PCA 无法有效分类，可能原因是对于小分子有机物，碳链结构和苯环对吸收系数的影响大于官能团对吸收系数的影响。

**Table 2.** The common Kernel function
**表 2.** 常见的核函数

| 核函数类型 | 核函数公式 |
|---|---|
| Linear Kernel | $K(X,X_i)=(X^{\mathrm{T}}X_i)$ |
| Polynomial Kernel | $K(X,X_i)=((X^{\mathrm{T}}X_i)+1)^d$ |
| Radial based Kernel (RBF) | $K(X,X_i)=\exp(-\gamma\|x-x_i\|^2)$ |
| Sigmoid Kernel | $K(X,X_i)=\tanh(x^{\mathrm{T}}x_i)+b$ |

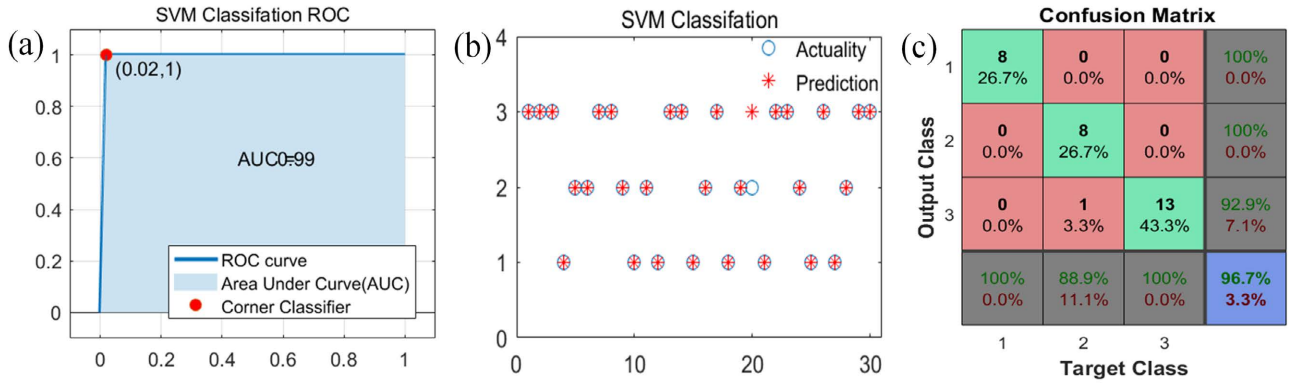


**Figure 6.** SVM predict results: (a) ROC; (b) Actual and predict difference; (c) Confusion matrix
**图 6.** SVM 预测结果：(a) ROC；(b) 实际与预测结果图；(c) 混淆矩阵

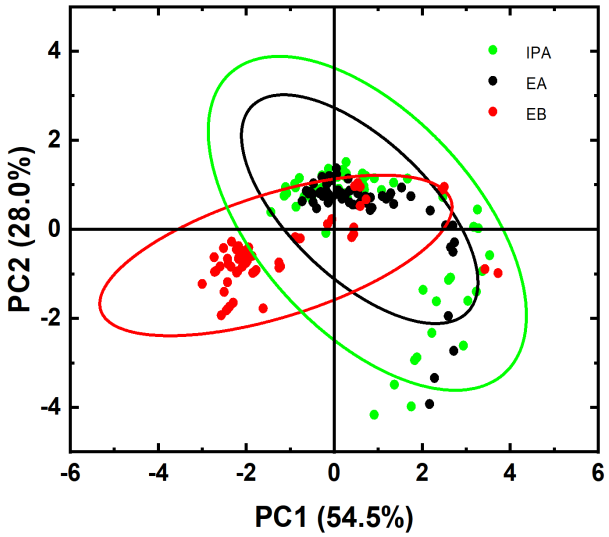


**Figure 7.** PCA score plot
**图 7.** PCA 得分图

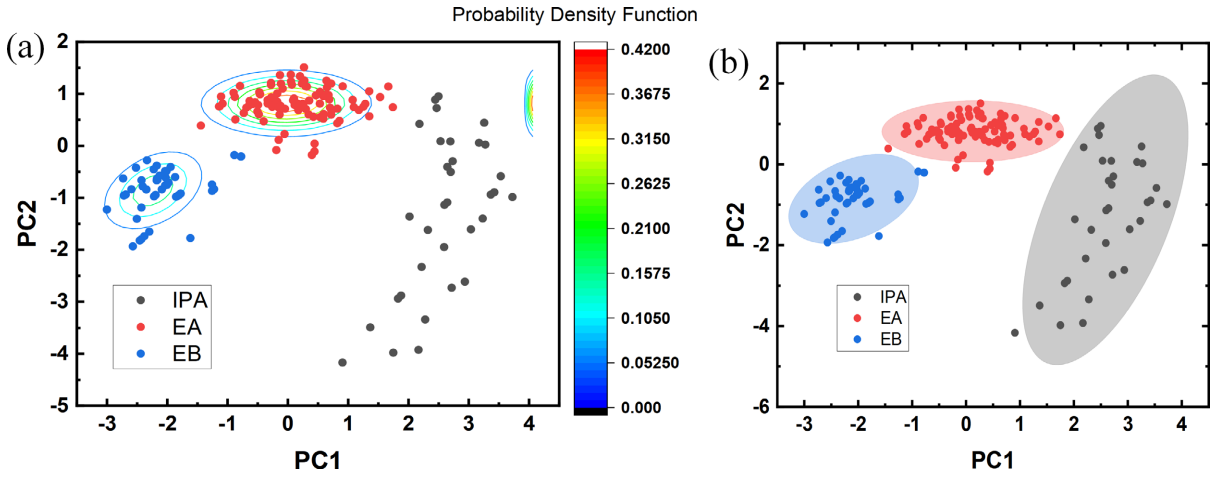


**Figure 8.** PCA-GMM (a) Contour map; (b) Confidence Interval
**图 8.** PCA-GMM (a) 等高线图；(b) 置信区间

高斯混合模型(Gaussian Mixture Model, GMM)通过迭代求出模型中高斯分布的似然函数收敛，判断某个点属于某个类的概率大小进行聚类[14]。EM 算法可有效求解存在隐含变量优化的问题。采用 PCA 降低数据维度后，结合 EM 方法拟合高斯混合模型，通过 EA 将似然函数取得最大值的参数值用作参数估计，代入 GMM 的极大似然函数，计算概率密度函数以及按后验概率对测试数据进行分类，并通过 BIC 找到最佳模型，绘制出等值线图和置信区间。分类结果如图 8(a)所示，EA 和 EB 实现了有效分离，在 95%的置信区间内绘制的概率密度函数等高线图中，置信区间由内而外可观察到各分类的出现概率逐步降低，而 IPA 等高线虽然并不明显，图 8(b)的置信区间图中也实现了有效分离，可视化效果更佳。

## 4. 结论

本文针对 VOC 物质在土壤中含量少、检测难的问题，采用 Fano 超表面设计的共振峰频移和强度变化进行拟合，构建了太赫兹波段 IPA、EA 和 EB 的“后天指纹”。在“后天指纹”的构造过程中，采用 classification Learner 对分类方法进行与选择，考虑到准确度、ROC 曲线和 AUC 值，以及混淆矩阵三方面的表现，从 22 个常用算法中选择非线性 SVM 作为分类方式，参数寻优后达到 96.7%的预测正确率。同时，还选用 PCA-GMM 的方式，对 IPA、EA 和 EB 的分类情况精准描绘，运用等高线辅助判断样点出现概率，为 VOCs 物质检测方式提供了新的思路。

## 参考文献

[1] 陆海杰, 姚乾秦, 屠秉坤, 等. 化工园区 VOCs 污染综合治理技术研究进展[J]. 中国资源综合用, 2022, 40(9): 90-92.

[2] 杨航. 典型化工园区 VOCs 排放扩散的预测溯源方法研究[D]: [博士学位论文]. 杭州: 浙江大学, 2020.

[3] 陈颖, 叶代启, 刘秀珍, 等. 我国工业源 VOCs 排放的源头追踪和行业特征研究[J]. 中国环境科学, 2012, 32(1): 48-55.

[4] 李守信, 宋剑飞, 李立清, 等. 挥发性有机化合物处理技术的研究进展[J]. 化工环保, 2008, 163(1): 1-7.

[5] 梅明, 郭兆云. 土壤挥发性有机物分析方法概述[J]. 武汉工程大学学报, 2013, 35(3): 18-24.

[6] 姜林, 钟茂生, 姚珏君, 等. 挥发性有机物污染土壤样品采样方法比较[J]. 中国环境监测, 2014, 30(1): 109-114.

[7] 殷甫祥. 气相抽提法(SVE)去除污染土壤中挥发性有机物(VOCs)的技术研究[D]: [硕士学位论文]. 扬州: 扬州大学, 2010.

[8] 牧凯军, 张振伟, 张存林. 太赫兹科学与技术[J]. 中国电子科学研究院学报, 2009, 4(3): 221-237.

[9] 赵碧辉, 文岐业, 谢云松, 等. 电磁超材料吸收器的研究进展[J]. 电子元件与材料, 2011, 30(11): 82-86.

[10] 刘元忠, 张玉萍, 曹妍妍, 等. 基于石墨烯超材料深度可调的调制器[J]. 光学学报, 2016, 36(10): 416-425.

[11] 付亚男, 张新群, 赵国忠, 等. 基于谐振环的太赫兹宽带偏振转换器件研究[J]. 物理学报, 2017, 66(18): 73-82.

[12] 冯晓瑜. 基于支持向量机的有机化合物红外光谱结构解析[D]: [硕士学位论文]. 成都: 四川大学, 2007.

[13] 曹萌萌, 杨圣舒, 丁胜男, 等. 基于土壤反射光谱聚类分析的有机质预测模型[J]. 中国农业信息, 2017, 205(10): 58-62.

[14] 刘佳斌, 郜允兵, 李永涛, 等. 基于高斯混合模型的土壤环境质量分区研究[J]. 农业环境科学学报, 2021, 40(8): 1746-1757.