# 基于太赫兹衰减全反射光谱的水质分析

曹秋红，林红梅，周　薇，李照鑫，张同军，黄海青，李学敏，李德华*

山东科技大学电子信息工程学院，青岛市太赫兹重点实验室，山东 青岛　266590

**摘　要**　随着人口的增长和社会的迅速发展，水资源短缺和水污染问题日益严重。水质分类作为水质污染评估工作中的一项重要环节，其意义和作用也更加突出。基于太赫兹衰减全反射(THz-ATR)光谱和模式识别技术，提出了一种水质分析模型。利用太赫兹时域光谱系统和衰减全反射模块测量了纯净水、自来水、河水、海水A和海水B五种水样的太赫兹衰减全反射光谱，通过光学参数提取模型获得0.2～1.0 THz频率范围内五种水样的折射率、吸收系数、介电常数实部和介电常数虚部。利用主成分分析(PCA)对折射率进行降维和特征提取，分别作出样品在第一、二主成分上的二维得分图和前三个主成分上的三维得分图，结果显示，基于折射率的主成分得分图可以明显的区分不同的水样。为了进一步对不同水样进行准确分类，将降维之后的数据输入到支持向量机(SVM)中构建水样分类模型，每种水样随机选取其中的五分之三作为训练集，剩余的数据作为测试集，同时引入网格搜索(GridSearch)、遗传算法(GA)和粒子群(PSO)三种优化算法对支持向量机参数进行优化。结果显示，基于网格搜索算法的支持向量机最优参数 $c$ 和 $g$ 分别为1.414 2和2.0，准确率为99.0%；基于遗传算法的支持向量机最优参数 $c$ 和 $g$ 分别为1.675 4和5.966 5，准确率为99.5%；基于粒子群算法的支持向量机最优参数 $c$ 和 $g$ 分别为3.154 9和12.589，准确率为100%。可以看出，使用不同的优化算法得到的最优参数不同，所构建的支持向量机分类模型都可实现正确的分类，且分类准确率均高达99.0%以上。研究结果表明，利用粒子群优化算法基于折射率构建的PCA-SVM分类模型效果最优，可以准确识别不同水样，为水质分类奠定了基础。

**关键词**　太赫兹；衰减全反射；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(2022)01-0031-07

## 引　言

随着人口的增长和社会的迅速发展，水资源短缺和水污染问题日益严重。水质分类作为水质污染评估工作中的一项重要环节，其意义和作用也更加突出。随着太赫兹技术日趋成熟，太赫兹光谱技术在安全监控[1]、食品添加剂检测[2]等领域都表现出巨大应用价值。由于水对太赫兹波有很强的吸收，利用太赫兹透射谱测量水样时需将样品厚度控制在100 μm以内[3]，对样品池精度要求较高，而太赫兹衰减全反射技术操作简单，无需对样品进行预处理，因此利用太赫兹衰减全反射(Terahertz attenuated total reflection，THz-ATR)技术对水溶液和液体样品进行检测、分析近年来逐渐成为了研究热点。2004年Hirori等[4]利用THz-ATR技术测定了蒸馏水的衰减全反射光谱，并推导了它的介电常数，结果表明ATR测得水的复介电常数与通过THz时域反射光谱法获得的结果有很好的一致性。2006年Nagai等[5]利用THz-ATR技术准确测定蒸馏水和蔗糖溶液的介电常数。2008年Newnham等[6]使用太赫兹脉冲光谱仪和硅ATR模块，测量了固体材料和液体的ATR光谱(0.3～3.6 THz)，所测得太赫兹衰减全反射谱与测得的透射谱特征一致。2013年Shiraga等[7]提出了一种用太赫兹时域衰减全反射结合双界面模型来确定单层细胞复介电常数的方法，这一方法使我们能够在皮秒尺度上估计细胞内的水分子动力学，表明细胞单层内存在弱水合水分子。以上研究表明THz-ATR技术不需要对液体样品进行处理，可直接用于水溶液的测定，具有检测方便、灵敏度高、精确度高和无损检测等特点。

利用THz-ATR技术对不同水质的研究分析，至今未见到相关报道。本文利用THz-ATR技术对海水等水样品进行了测量，并提取0.2～1.0 THz频段的折射率、吸收系数、介

**收稿日期**：2020-12-29，**修订日期**：2021-04-06

**基金项目**：国家重点研发计划重点专项(2017YFA0701000)资助

**作者简介**：曹秋红，1994年生，山东科技大学电子信息工程学院硕士研究生　　e-mail：2224646549@qq.com

*通讯作者　　e-mail：jcbwl@sdust.edu.cn

电常数等光学参数，结合主成分分析和支持向量机等模式识别方法对所提取的光学参数建立分类模型，实现对不同水样的分类识别，为水质评估提供一种新的模式。

## 1 实验部分

### 1.1 装置

实验中使用的测量仪器是德国 BATOP 公司生产的 TDS1008 太赫兹时域光谱系统。实验中无需对样品进行预处理[7]，用滴管直接吸取 2 mL 的水样，滴至 ATR 模块硅棱镜表面。如图 1 为 ATR 测量结构示意图，太赫兹波以 $\theta$ 角入射到 ATR 棱镜中，太赫兹波在棱镜-样品界面发生全反射，倏逝波渗透到样品中，其渗透深度取决于样品和 ATR 晶体的折射率、太赫兹波的入射角、偏振态和频率[5]。本文采用 S 偏振 THz 波，选取纯净水、自来水、河水、海水 A 和海水 B（海水 A 和海水 B 取自黄海海域不同水域）五种水样品，测得 0.2～1.0 THz 频率范围内样品的 ATR 光谱，每种样品测量十次，共获得 50 组数据。

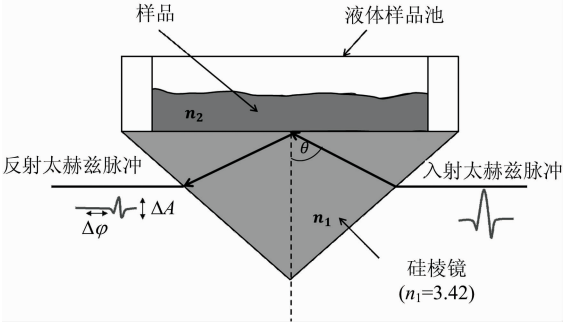



**图 1 ATR 的结构示意图，入射角 $\theta$ 为 51.6°，硅棱镜的折射率为 3.42**

**Fig. 1 A schematic diagram of the structure of the ATR, the incident angle $\theta$ is 51.6°, The refractive index of the silicon prism is 3.42**

### 1.2 光学参数提取

对所测得的时域信号进行快速傅里叶变换[5-8]，得到传输函数 $H(\omega)$，传输函数幅值 $\rho(\omega)$ 和相位 $\phi(\omega)$ 如式(1)—(3)所示

$$H(\omega)=\frac{E_{\text{sam}}(\omega)}{E_{\text{ref}}(\omega)}=\frac{E_{\text{in}}(\omega)r}{E_{\text{in}}(\omega)r'}=\frac{r}{r'} \tag{1}$$

$$\rho(\omega)=\left|\frac{r}{r'}\right| \tag{2}$$

$$\phi(\omega)=\text{Arg}\left(\frac{r}{r'}\right) \tag{3}$$

其中，$E_{\text{in}}(\omega)$ 为入射太赫兹波的电场强度，$E_{\text{sam}}(\omega)$ 和 $E_{\text{ref}}(\omega)$ 分别为样品信号和参考信号的电场强度，$r$ 和 $r'$ 分别为棱镜-样品界面和棱镜-空气界面的全反射系数。

输入信号 $E_{\text{in}}(\omega)$ 和输出信号 $E_{\text{out}}(\omega)$ 由全反射系数 $r$ 决定，即 $E_{\text{out}}(\omega)=E_{\text{in}}(\omega)r$。对于 S 偏振和 P 偏振的太赫兹波

$$r_{\text{S}}=\frac{\tilde{n}_1\cos\theta-\tilde{n}_2\sqrt{1-\left(\frac{\tilde{n}_1}{\tilde{n}_2}\sin\theta\right)^2}}{\tilde{n}_1\cos\theta+\tilde{n}_2\sqrt{1-\left(\frac{\tilde{n}_1}{\tilde{n}_2}\sin\theta\right)^2}} \tag{4}$$

$$r_{\text{P}}=\frac{\tilde{n}_2\cos\theta-\tilde{n}_1\sqrt{1-\left(\frac{\tilde{n}_1}{\tilde{n}_2}\sin\theta\right)^2}}{\tilde{n}_2\cos\theta+\tilde{n}_1\sqrt{1-\left(\frac{\tilde{n}_1}{\tilde{n}_2}\sin\theta\right)^2}} \tag{5}$$

其中，$r_{\text{S}}$ 和 $r_{\text{P}}$ 分别为 S 偏振和 P 偏振的全反射系数，$\tilde{n}_1$ 和 $\tilde{n}_2$ 分别为硅棱镜和空气的复折射率。联立式(1)—式(4)，可以得到样品信号的全反射系数 $r$

$$r=\rho(\omega)\text{e}^{-j[\phi(\omega)+1.73]} \tag{6}$$

样品介电常数为

$$\varepsilon_2=n_2^2=\frac{\cos^2(\theta)(1-r)^2+\sin^2(\theta)(1+r)^2}{(r+1)^2}\times n_1^2 \tag{7}$$

获得样品的折射率 $n$ 和吸收系数 $\alpha$ 如式(8)和式(9)

$$n=\sqrt{\frac{\sqrt{[\text{Re}(\varepsilon_2)]^2+[\text{Im}(\varepsilon_2)]^2}+\text{Re}(\varepsilon_2)}{2}} \tag{8}$$

$$\alpha=\frac{\sqrt{2}\omega\sqrt{\sqrt{[\text{Re}(\varepsilon_2)]^2+[\text{Im}(\varepsilon_2)]^2}-\text{Re}(\varepsilon_2)}}{c} \tag{9}$$

## 2 结果与讨论

### 2.1 光谱分析

利用 THz-ATR 测量了纯净水、自来水、河水、海水 A 和海水 B 五种水样品的时域光谱，通过光学参数提取模型得到样品在 0.2～1.0 THz 范围内的吸收系数、折射率和介电常数。图 2 为海水 A 样品在 0.2～1.0 THz 频率范围内折射率、吸收系数、介电常数实部和虚部随频率变化的对比图，从图中可以看出海水 A 样品的十次测量结果略有差异，其光学参数随频率变化趋势大致相同。图 3 为不同水样的折射率、吸收系数、介电常数实部和虚部对比图，从图 3 可以看出，纯净水与其他四种水吸收系数、介电常数都有较明显不同，其他四种水样的折射率、吸收系数和介电常数差异较小，仅靠光学参数谱线无法区分不同的水样。

### 2.2 主成分分析

主成分分析(principal component analysis，PCA)是一种数学统计方法[9]。由于变量之间具有一定的相关性，因此变量之间可能存在一些重叠信息[10]。PCA 用于将一组可能相关的变量转换成一组线性不相关的变量，这组线性不相关的变量称为主成分(Principal components，PCs)。PCs 是原始变量的线性组合，其个数小于原始数据的个数。为了减少光谱的数据冗余，提高模型效率，对样品折射率、吸收系数、介电常数实部和虚部在 0.2～1.0THz 波段的原始数据进行主成分分析，降维后前 3 项主成分的累计方差贡献率分别为 98.992%，99.722%，99.242%和 99.762%，可以近似解释所有原始数据。图 4 和图 5 分别是基于不同光学参数的 PCA 二维和三维得分图，从图中可以看出，吸收系数、介电常数实部和虚部的二维和三维 PCA 得分图无法区分自来水、河水和海水，而折射率的二维和三维 PCA 得分图可以明显的区分不同水样，并且基于折射率的三维 PCA 得分图聚类效

果最好，可以通过聚类情况区分不同的水样。由于 PCA 结果取决于原始数据，上述聚类结果说明了实验样本中纯净水、自来水、河水和海水这四种水样的折射率光谱特性具有一定的差异，而海水 A 和海水 B 的折射率特性相近；自来水、河水和海水的吸收系数、介电常数实部和虚部的光谱特性相近，这一结果与光谱测量结果相符。

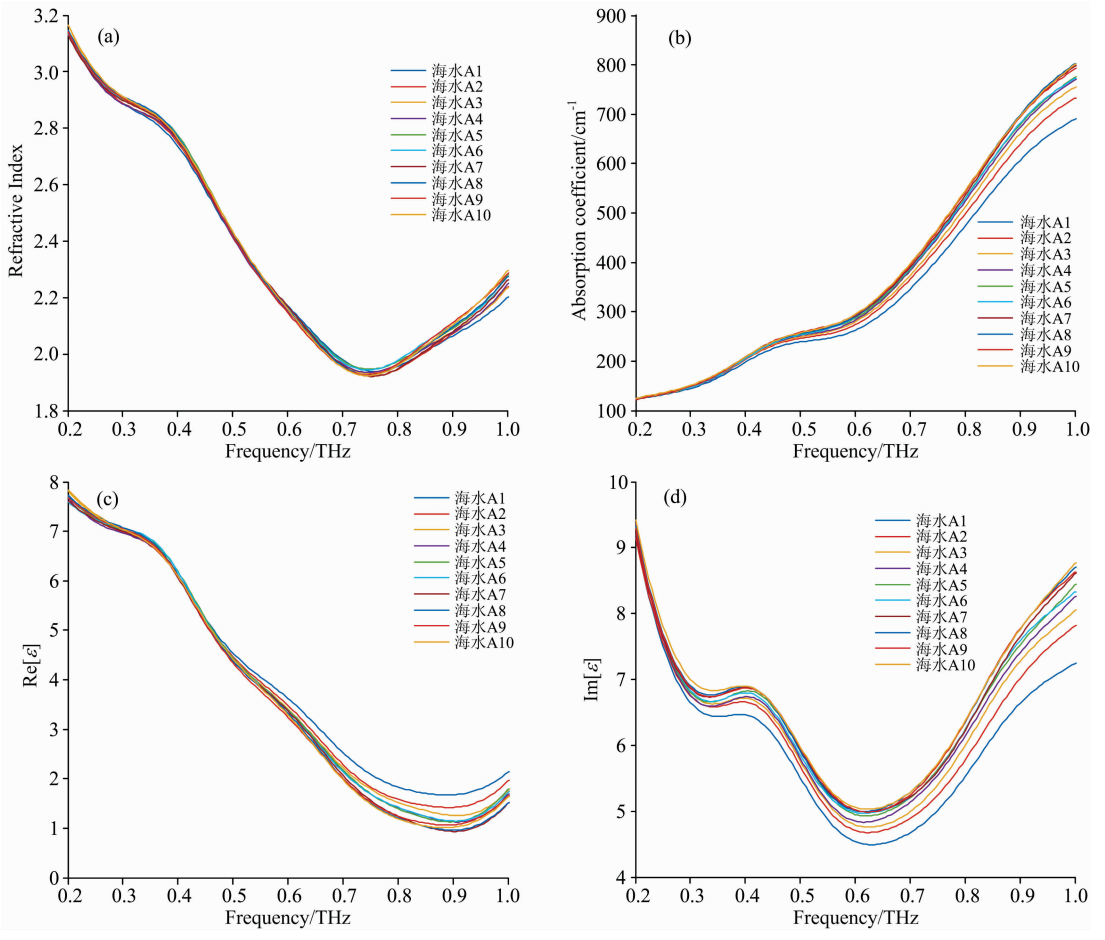


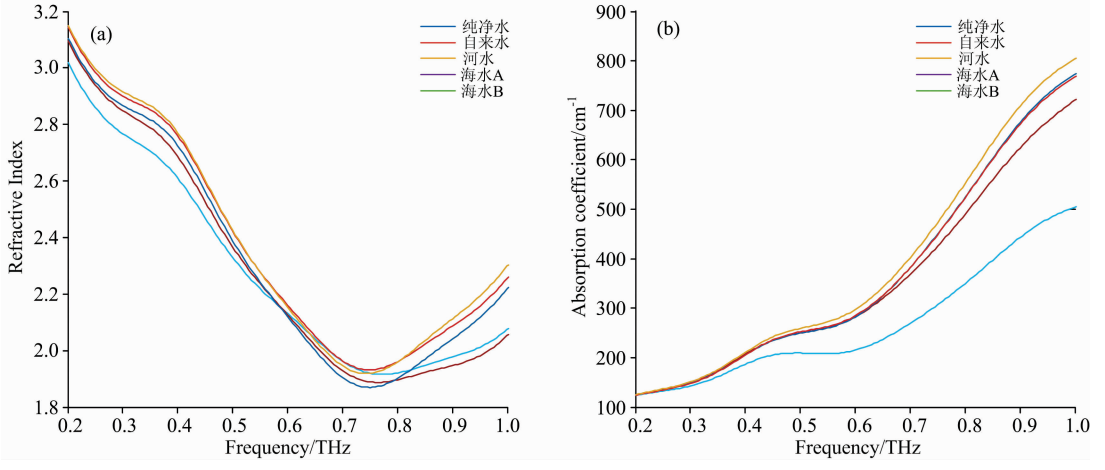

**图 2　海水 A 样品在 0.2～1.0 THz 范围内的光学参数**

(a)：折射率；(b)：吸收系数；(c)：介电常数实部；(d)：介电常数虚部

**Fig. 2　Comparison of optical parameters of sea water A sample in the range of 0.2～1.0 THz**

(a)：Refractive index；(b)：Absorption coefficient；

(c)：Real part of dielectric constant；(d)：Imaginary part of dielectric constant



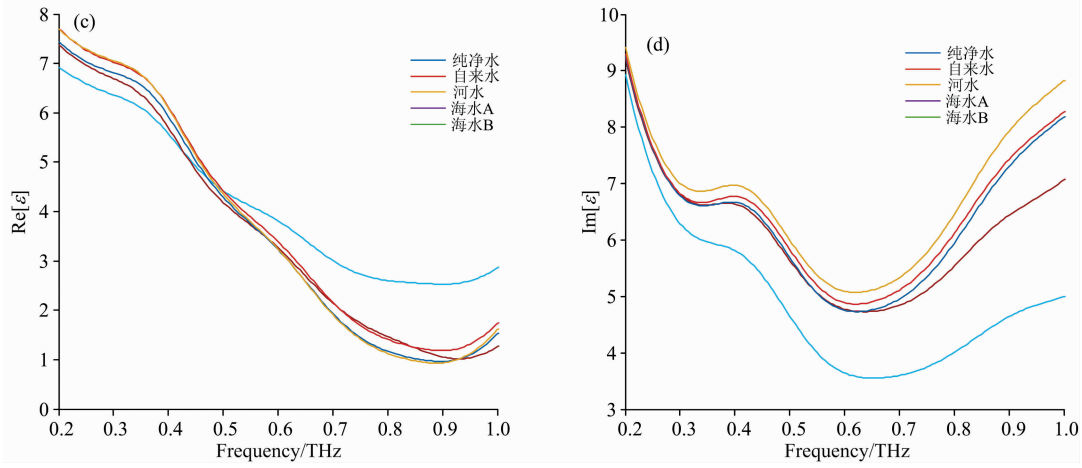



**图 3 纯净水、自来水、河水、海水 A 和海水 B 五种样品在 0.2～1.0 THz 范围内的光学参数对比**

(a)：折射率；(b)：吸收系数；(c)：介电常数实部；(d)：介电常数虚部

**Fig. 3 Optical parameter comparison of five samples in the range of 0.2～1.0 THz for purified water, tap water, river water, seawater A and seawater B**

(a)：Refractive index；(b)：Absorption coefficient；(c)：Real part of dielectric constant；(d)：Imaginary part of dielectric constant

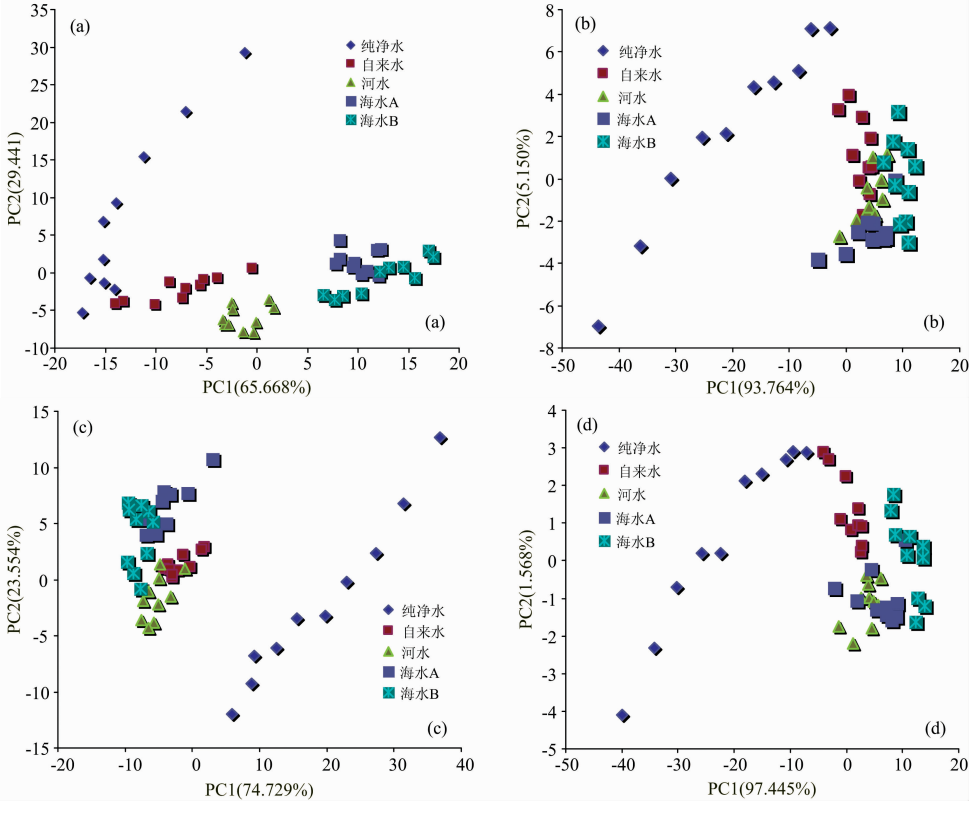



**图 4 纯净水、自来水、河水、海水 A 和海水 B 在 0.2～1.0 THz 范围内的光学参数在第一、二主成分上的得分**

(a)：折射率得分；(b)：吸收系数得分；(c)：介电常数实部得分；(d)：介电常数虚部得分

**Fig. 4 The scores of the optical parameters of purified water, tap water, river water, sea water A and sea water B in the range of 0.2～1.0 THz on the first and second principal components**

(a)：Score of refractive index；(b)：Score of absorption coefficient；

(c)：Score of dielectric constant real part；(d)：Score of dielectric constant imaginary part

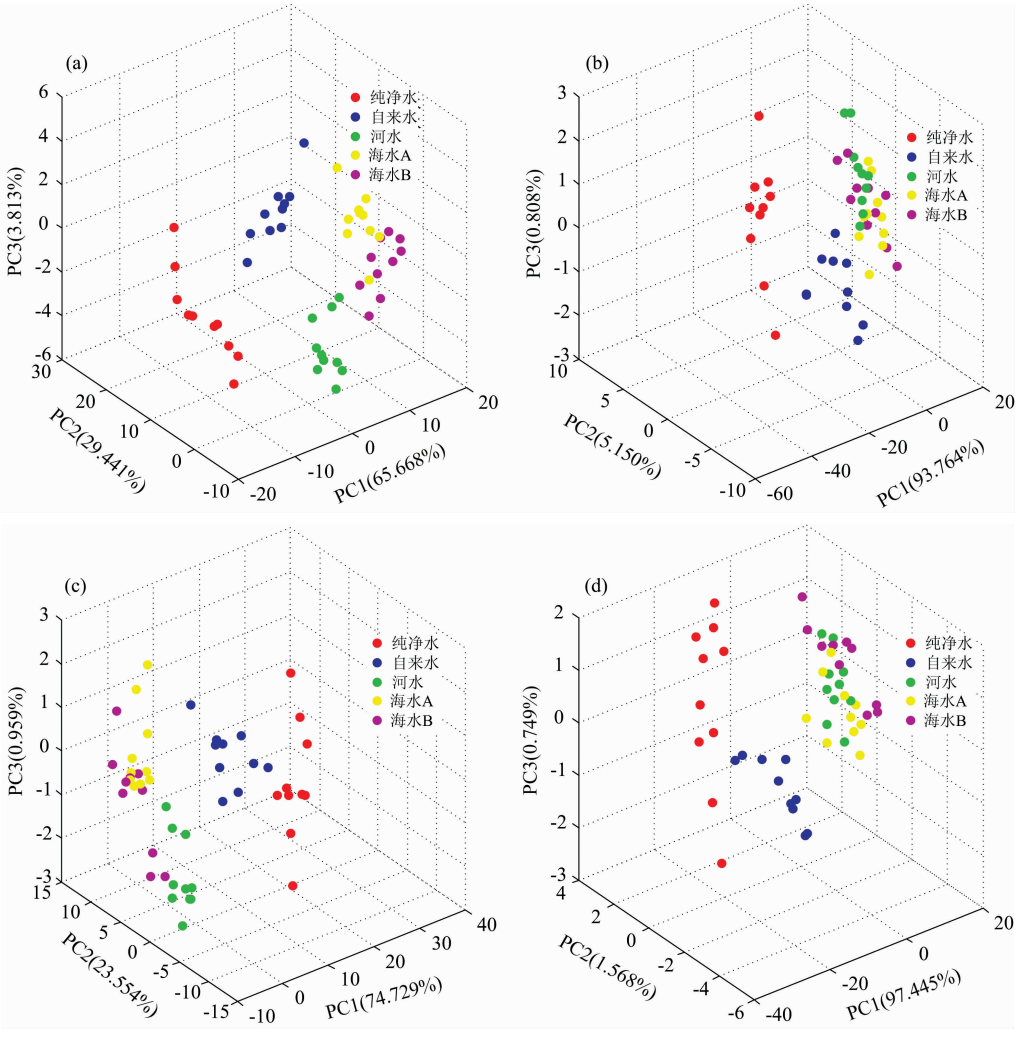



**图 5　纯净水、自来水、河水、海水 A 和海水 B 在 0.2～1.0 THz 范围内的光学参数在前三个主成分上的得分**

(a)：折射率得分；(b)：吸收系数得分；(c)：介电常数实部得分；(d)：介电常数虚部得分

**Fig. 5　The scores of the optical parameters of purified water, tap water, river water, sea water A and sea water B in the range of 0.2～1.0 THz on the first three principal components**

(a): Score of refractive index; (b): Score of absorption coefficient;
(c): Score of dielectric constant real part; (d): Score of dielectric constant imaginary part

### 2.3　支持向量机分析

支持向量机(support vector machines，SVM)是一种机器学习算法，它在解决小样本、非线性和高维模式识别问题时具有独特的优势[11]。由于折射率的三维主成分得分图分类效果最好，因此选取样品折射率的前三个主成分作为输入数据输入到 SVM 中建立分类模型。在 SVM 中，数据集分为两类，一类是训练集，一类是测试集。实验共测得 5 种水样，每种水样各测 10 组，共 50 组数据。每种水样中随机抽取 6 组数据(共 30 组)作为训练集，剩余的 20 组数据作为测试集。

为了提高分类模型的预测精度，需要使用优化算法来优化惩罚参数 $c$ 和径向基函数核参数 $g$[12]。引入遗传算法(genetic algorithm，GA)、网格搜索(grid search，GridSearch)和粒子群(particle swarm optimization，PSO)三种优化算法来搜索参数 $c$ 和 $g$ 的最佳组合[13]，从而选出准确率最高的优化算法建立 PCA-SVM 分类模型。表 1 为 PCA-SVM 结合三种优化方法的结果对比，其中 GA、GridSearch 和 PSO 三种优化算法的训练集准确率都达到了 100%，测试集准确率分别为 99.5%，99.0%和 100%。图 6 为 PSO 优化算

**表 1　PCA-SVM 结合遗传、网格搜索和粒子群三种优化方法对比**

**Table 1　Comparison of three optimization methods of PCA-SVM combined with GA, Gridsearch and PSO**

| 优化方法 | 最优参数 $c$ | 最优参数 $g$ | 种群数量 | 迭代次数 | 训练集准确率/% | 测试集准确率/% |
|---|---|---|---|---|---|---|
| 遗传算法 | 1.675 4 | 5.966 5 | 20 | 200 | 100 | 99.5 |
| 网格搜索法 | 1.414 2 | 2 | 20 | 200 | 100 | 99.0 |
| 粒子群算法 | 3.154 9 | 12.589 0 | 20 | 200 | 100 | 100 |

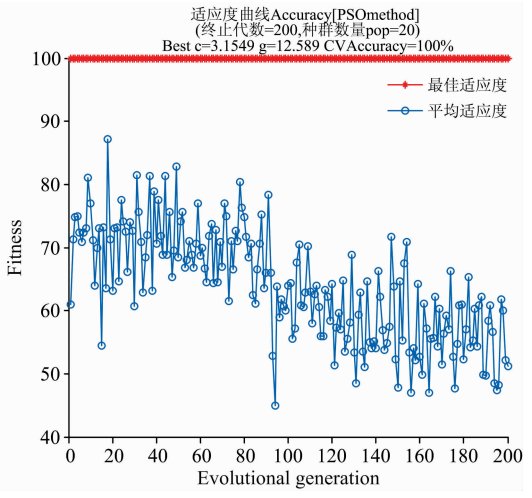



**图 6 PSO 优化算法的适应度曲线**
**(最优参数 $c=3.154\ 9$, $g=12.589$)**
**Fig. 6 Fitness curve of PSO**
**(optimal parameter $c=3.154\ 9$, $g=12.589$)**

法的适应度曲线，可以看出当惩罚参数 $c$ 为 3.154 9，核函数参数 $g$ 为 12.589 时，训练集和测试集的准确率均达到 100%。结果表明，PSO 构建的 PCA-SVM 分类模型效果最优，可以对不同水样品进行很好的分类鉴别。

## 3 结 论

利用 THz-ATR 技术，测得纯净水、自来水、河水、海水 A 和海水 B 五种水样品在 0.2～1.0 THz 频段下折射率、吸收系数和介电常数。采用 PCA 对折射率原始数据进行降维和特征提取，将提取后的前三个主成分输入到 SVM 中建立分类模型。引入 GA、GridSearch 和 PSO 算法对 SVM 参数进行优化。三种算法的优化识别率分别为 99.5%，99.0%和 100%。结果表明，利用 PSO 优化算法基于折射率构建的 PCA-SVM 分类模型识别不同的水样准确率达到了 100%。因此，利用 THz-ATR 技术结合 PCA-SVM 分类模型有望用于水资源质量的快速检测。

## References

[1] Cheng Yayun, Wang Yingxin, Niu Yingying, et al. Optics Express, 2020, 28(5): 6350.
[2] HU Jun, LIU Yan-de, SUN Xu-dong, et al(胡 军, 刘燕德, 孙旭东, 等). Laser & Optoelectronics Progress(激光与光电子学进展), 2020, 57(7): 073002.
[3] HAN Xue, SU Bo, ZHANG Cun-lin(韩 雪, 苏 波, 张存林). Journal of Terahertz Science and Electronic Information Technology(太赫兹科学与电子信息学报), 2015, 13(4): 536.
[4] Hirori H, Yamashita K, Nagai M, et al. Japanese Journal of Applied Physics. Part 2, Letters & Express Letters, 2004, 43(10A): L1287.
[5] Nagai M, Yada H, Arikawa T, et al. International Journal of Infrared and Millimeter Waves, 2006, 27(4): 505.
[6] Newnham D A, Taday P F. Applied Spectroscopy, 2008, 62(4): 394.
[7] Shiraga K, Ogawa Y, Suzuki T, et al. Applied Physics Letters, 2013, 102(5): 053702.
[8] NIE Mei-tong, XU De-gang, WANG Yu-ye, et al(聂美彤, 徐德刚, 王与烨, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(7): 2016.
[9] Lian Feiyu, Xu Degang, Fu Maixia, et al. IEEE Transactions on Terahertz Science and Technology, 2017, 7(4): 378.
[10] Zhang Huo, Li Zi, Chen Tao, et al. Optik-International Journal for Light and Electron Optics, 2017, 138: 95.
[11] Guo Jin, Deng Hu, Liu Quancheng, et al. Journal of Spectroscopy, 2020, 2020: 8811467.
[12] Liang Jie, Guo Qijia, Chang Tianying, et al. Optik, 2018, 174: 7.
[13] ZHANG Wen-tao, LI Yue-wen, ZHAN Ping-ping, et al(张文涛, 李跃文, 占平平, 等). Infrared and Laser Engineering(红外与激光工程), 2017, 46(11): 1125004.

# Water Quality Analysis Based on Terahertz Attenuated Total Reflection Technology

CAO Qiu-hong, LIN Hong-mei, ZHOU Wei, LI Zhao-xin, ZHANG Tong-jun, HUANG Hai-qing, LI Xue-min, LI De-hua*
Qingdao Key Laboratory of Terahertz Technology, College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China

**Abstract** With the growth of population and the rapid development of society, the problem of water shortage and water pollution have become increasingly serious. As an important aspect of water pollution assessment, water quality classification has a more prominent significance and role. Based on terahertz attenuated total reflection (THz-ATR) spectrum and pattern recognition technology, a water quality analysis model is proposed in this paper. The terahertz time-domain spectroscopy system and the attenuated total reflection module were used to measure the terahertz attenuated total reflection spectra of five water samples, including pure water, tap water, river water, seawater A and seawater B. The refractive index, absorption coefficient, real and imaginary parts of the dielectric constant of five water samples in the frequency range of 0.2～1.0 THz were obtained using the optical parameter extraction model. Principal Component Analysis (PCA) was applied to conduct refractive index reduction and feature extraction, and two-dimensional score charts of the first and second principal components and three-dimensional score charts of the first three principal components of the samples were made respectively. It can be seen that the principal component score chart based on the index of refraction can clearly distinguish different water samples. In order to further classify different water samples accurately, the data after dimension reduction is input into a support vector machine to construct a water sample classification model. Three-fifths of each water sample is randomly selected as the training set, and the remaining data is used as the test set. At the same time, three optimization algorithms, grid search (GridSearch), genetic algorithm (GA) and particle swarm algorithm (PSO) are introduced to optimize the parameters of the support vector machine. The results show that the optimal parameters $c$ and $g$ of the support vector machine based on the grid search algorithm are 1.414 2 and 2.0, respectively, with an accuracy of 99.0%; the optimal parameters $c$ and $g$ of the support vector machine based on the genetic algorithm are 1.675 4 and 5.966 5, respectively, which are accurate The rate is 99.5%; the optimal parameters $c$ and $g$ of the support vector machine based on particle swarm optimization are 3.154 9 and 12.589 respectively, and the accuracy rate is 100%. It can be seen that the optimal parameters obtained by different optimization algorithms are different, and all the SVM classification models constructed can achieve correct classification, and the classification accuracy is above 99.0%. The research results show that the PCA-SVM classification model based on the refractive index constructed by the particle swarm optimization algorithm has the best effect and can accurately identify different water samples, laying a foundation for water quality classification.

**Keywords** Terahertz; Attenuated total reflection; Principal component analysis; Support vector machine

(Received Dec. 29, 2020; accepted Apr. 6, 2021)

* Corresponding author