基于差分-主成分分析-支持向量机的有机化合物太赫兹吸收光谱识别方法

刘俊秀[1]，杜彬[1]，邓玉强[2]，张建文[3]，祝海江[1]*

[1]北京化工大学信息科学与技术学院，北京 100029；

[2]中国计量科学研究院光学所，北京 100029；

[3]北京化工大学化学工程学院，北京 100029

摘要 针对有机化合物的太赫兹时域光谱数据，提出了一种基于差分-主成分分析(PCA)-支持向量机(SVM)的有机化合物识别方法。基于物质样本的太赫兹时域信号计算得到太赫兹吸收光谱，对 0.2～2.5 THz 频率区间内的数据进行特征提取。在特征提取中，提出了基于差分数据的样本容量扩充方法，并结合 PCA 进行了特征的提取。利用 SVM 建立了提取的特征与物质类别对应关系的数学模型，并根据建立的模型对未知样本进行了识别研究。利用所提方法对 15 种有机化合物的太赫兹光谱数据进行了识别，正确识别率为 93.33%。将所提方法与线性判别分析法及吸收峰频率-幅值法进行了对比，结果表明基于差分-PCA-SVM 的有机化合物识别方法的正确识别率最高。

关键词 太赫兹技术；光谱学；差分数据；主成分分析；支持向量机

中图分类号 O436 **文献标识码** A **doi**：10.3788/CJL201946.0614039

Terahertz-Spectral Identification of Organic Compounds Based on Differential PCA-SVM Method

Liu Junxiu[1], Du Bin[1], Deng Yuqiang[2], Zhang Jianwen[3], Zhu Haijiang[1]*

[1]*College of Information Science & Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*;

[2]*Optics Division, National Institute of Metrology, Beijing* 100029, *China*;

[3]*College of Chemical Engineering, Beijing University of Chemical Technology, Beijing* 100029, *China*

Abstract This paper proposes a method for identifying organic compounds by applying a differential principal-component-analysis (PCA)-support-vector-machine (SVM) to the terahertz time-domain spectral data. First, the terahertz absorption spectrum is calculated according to the terahertz time-domain signal of the material sample; then, the features of the data in the frequency range of 0.2-2.5 THz are extracted. During the feature extraction, an expansion-of-sample-size method based on differential data is proposed and combined with the PCA method to achieve the feature extraction. Finally, the SVM is used to establish a mathematical model for the corresponding relationship between the extracted features and the material category, and the unknown samples are identified according to this model. The terahertz-spectral data of 15 organic compounds are identified using the proposed method, and the correct recognition rate is 93.33%. The experimental results show that the correct recognition rate of organic compounds by the proposed method is the highest when compared with those by the linear-discriminant analysis method and the absorption peak frequency-amplitude method.

Key words terahertz technology; spectroscopy; differential data; principal component analysis; support vector machine

OCIS codes 040.2235; 070.4560; 160.4890

收稿日期：2019-01-21；**修回日期**：2019-02-18；**录用日期**：2019-04-08

基金项目：国家自然科学基金(61672084,11834777)

* **E-mail**：zhuhj@mail.buct.edu.cn

1 引　言

太赫兹波是一种频率在 0.1～10 THz 之间的电磁波，位于远红外光和微波之间[1]。相比于目前相对成熟的红外波段微波和 X 射线，太赫兹波有其独特的优点：1）太赫兹波有很强的穿透性，能够穿透纸板、塑料和布料等常见的包装材料，这为对包装后物品进行检测识别提供了可能；2）太赫兹波的能量很低，在穿透物质的过程中几乎不产生电离辐射[2]，因此对于安检等接触人体的情况，太赫兹波在安全性能上更胜一筹；3）许多化学物质在太赫兹波段有其独特的信息，通过分析太赫兹光谱可以获取物质对光波的吸收特点，从而实现对物质更好的识别与区分。

太赫兹时域光谱可用于物质的定性识别，其传统方法是提取物质在太赫兹吸收光谱图中的吸收峰，但也存在部分物质在吸收光谱图中没有明显的吸收峰的情况。对于在太赫兹波段具有较明显吸收峰的物质，Chen 等[3]从物质的分子结构特性出发，选择一种手性物质——酒石酸（$C_4H_6O_6$），对其两种不同的异构体进行区分识别。应用密度泛函理论，对不同异构体的 $C_4H_6O_6$ 单分子进行理论模拟计算，对比了计算得到的光谱与实验得到的光谱的吸收峰频率位置，并结合不同分子振动模式进行分析，实现了对 $C_4H_6O_6$ 的两种不同异构体的区分。Trofimov 等[4]提出了采用整体相关标准方法判断某一频率的吸收峰是否合理，进而确定物质的种类。而对于部分没有明显吸收峰的物质，利用其吸收峰频率对物质种类进行判断存在困难和误差。Ni 等[5]采用一种流形学习方法，即扩散映射方法，将高维的太赫兹吸收光谱数据映射到低维空间，通过在低维空间内提取数据特征，并建立多分类支持向量机（SVM）对 10 种物质（5 种吸收峰较为明显的物质和 5 种吸收峰较不明显且整体谱线比较近似的物质）进行识别，准确率可达 96%。Liu 等[6]也采用了一种流形学习——等距映射（ISOMAP）算法，结合 SVM 提出了 ISOMAP-SVM 算法，并测量三种不同厂家生产的感冒药的吸收光谱，先用 ISOMAP 算法进行吸收光谱数据的降维，再用 SVM 进行建模。每种感冒药的样本数较多，而较多的样本数会影响 SVM 建模的速度，因此提出了邻接图法对 SVM 中的支持向量进行预选，该方法可缩短建模训练时间，且几乎不影响分类效果。

在物质光谱的识别方法方面，Xie 等[7]将实际测量的物质吸收光谱与标准数据库中的特征吸收光谱进行匹配，通过比较峰值面积偏差并设定阈值，判断被测的样品中是否含有标准库内的危险物质，可实现对爆炸品的识别。Wang 等[8]对应用于颜料识别的几种吸收光谱匹配识别方法进行了对比，对编辑距离（ED）算法的判定条件进行改进，提出了自适应阈值编辑距离的光谱匹配方法，减少了识别判断过程中算法过于敏感的问题，改善了匹配效果。Zhang 等[9]针对含有转基因成分和非转基因成分的大豆油，使用主成分分析（PCA）法在物质的吸收光谱中提取 8 个主成分，将其作为 SVM 的输入并建立模型，采用交叉验证的方式，实现对几种转基因和非转基因大豆油的识别。Pohl 等[10]针对吸收光谱进行多元数据分析处理，采用 PCA 法对 12 种物质，包括纯物质与混合物，进行特征提取，并利用偏最小二乘回归实现组分的定量分析，实现了物质的分类。

针对有机化合物的识别，本文提出了基于差分-PCA-SVM 的识别方法。在特征提取方面，传统的 PCA 方法需要对一种样品的多组数据同时进行特征提取，而本文针对小样本数据，包括待识别的物质只有一组测量数据的情况，提出了差分-PCA-SVM 的方法。对需要进行特征提取的数据先进行差分处理以扩充样本容量，结合 PCA 进行特征提取，并采用 SVM 建立分类模型。在识别的过程中，输入为从样品的太赫兹吸收光谱图中提取的特征，输出为物质的类别。首先采用 SVM 对物质特征建立分类模型，实现物质特征与类别的对应关系。针对未知样本的检测，先提取其太赫兹吸收光谱图的特征，用径向基神经网络判断物质是否属于需要识别物质中的一种，如果是，则进一步使用模型判断物质类别。

2 方法提出

基于差分-PCA-SVM 的识别方法的流程主要包括数据预处理、特征提取、数据建模和物质识别 4 部分，其整个有机化合物识别流程如图 1 所示。

在数据的获取中，使用由中国计量科学院自主研发的透射型太赫兹光谱仪测量得到太赫兹时域光谱数据。这种仪器可激发宽频太赫兹波，频率范围为 0.1～3.0 THz，光斑直径为 1 mm。由太赫兹光谱仪激发的太赫兹脉冲透射过样品，得到带有样品信息的太赫兹信号，经过探测器检测得到随时间变化的电场强度脉冲信号，即太赫兹时域光谱数据。在数据预处理部分，分别对样品信号和背景信号进行傅里叶变换，得到两组太赫兹频域数据。根据该样品与参考信号的频域数据，并利用光学参数的计算公式，得到太

赫兹信号中体现物质特征的吸收系数和折射率等光学参数。在特征提取部分，一条吸收曲线包含多个数据点，不利于后续的建模，因此，在此部分主要对吸收系数曲线进行特征提取。这里先用差分数据法扩充样本容量再结合 PCA 方法进行特征的提取。在数据建模和物质识别部分，选择一部分已有的物质数据作为已知样本，采用 SVM 法建立数据特征与物质类别之间的分类模型，并训练径向基神经网络，输出值表明已知样本均在数据库中。选择另一部分已有的物质数据作为未知样本，在提取数据特征后，先用径向基神经网络判断其是否为已知数据库中的一种：如果是，则用建立好的 SVM 模型输出物质的类别；如果不是，则不能给出其具体类别。

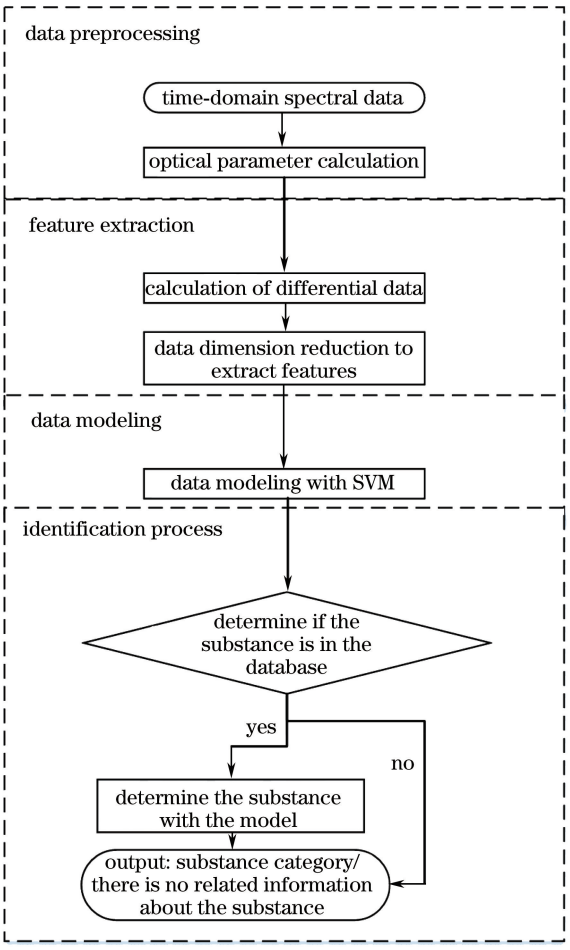

图 1　有机化合物识别流程

Fig. 1　Flow chart of organic compound identification

2.1　数据预处理

本小节主要描述根据太赫兹时域光谱数据计算得到表征样品特性的光学参数，并结合三次样条曲线插值拟合平滑曲线，使数据点分辨率达到统一。所用的数据为由中国计量科学院自主研发的透射型太赫兹光谱仪测量得到的随时间变化的电场强度脉冲信号。空气中存在的水蒸气对太赫兹辐射有一定的吸收作用，这会对实验结果产生较大影响，因此该测量过程需要在氮气环境中进行。

在处理过程中，需要两组太赫兹信号：一组为在不放置样品的条件下，只在背景环境下测量并记录得到的信号，记为背景信号；另一组为在放置样品的条件下，使太赫兹波透射样品并记录得到的信号，记为样品信号。对每一个样品的背景信号和样品信号的两组时域信号进行傅里叶变换，得到背景信号和样品信号的频域谱，分别记为 $E_{\mathrm{b}}(\omega)$和$E_{\mathrm{s}}(\omega)$，其中 ω 为频率。根据这两个频域信号，计算得到复透射函数 $H(\omega)$为

$$H(\omega)=\frac{E_{\mathrm{s}}(\omega)}{E_{\mathrm{b}}(\omega)}。\tag{1}$$

采用幅值和相位角来表示$H(\omega)$：

$$H(\omega)=P(\omega)\exp[-\mathrm{j}\varphi(\omega)],\tag{2}$$

式中：$\varphi(\omega)$表示相位角；$P(\omega)$表示幅值。材料的光学参数通常用复折射率 $\tilde{n}(\omega)$来描述，即

$$\tilde{n}(\omega)=n(\omega)-\mathrm{j}\kappa(\omega),\tag{3}$$

式中：实数部分 $n(\omega)$表示折射率；虚数部分 $\kappa(\omega)$表示消光系数。利用吸收系数 $\alpha(\omega)$可以计算样品对太赫兹波的吸收程度，即

$$\alpha(\omega)=\frac{2\omega\kappa(\omega)}{c},\tag{4}$$

式中：c 为真空中的光速。

忽略太赫兹波在样品内多次反射的情况，厚度为 d 的样本的折射率 $n(\omega)$和吸收系数分别满足

$$n(\omega)=\frac{\phi(\omega)c}{\omega d}+1,\tag{5}$$

$$\alpha(\omega)=\frac{2\omega\kappa(\omega)}{c}=\frac{2}{d}\ln\left\{\frac{4n(\omega)}{P(\omega)[n(\omega)+1]^2}\right\}。\tag{6}$$

2.2　特征提取

样本的吸收系数与频率有关，一条吸收系数谱线包含多个数据点。在这一部分，主要对吸收系数进行特征提取，以减少数据维数。在此利用 PCA 法结合差分数据实现对数据的特征提取。

利用 2.1 节描述的方法可以得到表示样品特征的吸收系数谱线，即吸收系数随频率变化的曲线，吸收系数数据为 $1\times n$ 维矩阵。不同的频率点对应不同的吸收系数。在整个频率区间内，某些频率点的吸收系数与频率的关系曲线表现出波峰，即样品物质在该频率点下有特征吸收峰。针对单一谱线不利于进行特征提取的问题，采用差分数据结合 PCA 法

的方式进行特征提取，即先用差分数据对单一样品的一维吸收系数谱线进行数据容量扩充，再用 PCA 法进行特征提取。

对原始的吸收系数数据进行一阶差分处理，每一个频率点的一阶差分数据计算方法为

$$\Delta y_1 = y(x+1) - y(x), \tag{7}$$

式中：x 表示所研究频率点对应的频率；$y(x)$表示该频率点下的吸收系数；$x+1$ 表示下一个频率点处对应的频率；Δy_1 表示一阶差分数据。对一阶差分数据再次进行差分处理，得到的二阶差分数据 Δy_2 为

$$\Delta y_2 = \Delta y_1(x+1) - \Delta y_1(x) = y(x+2) - 2y(x+1) + y(x), \tag{8}$$

并依此得到三阶差分数据 Δy_3 为

$$\Delta y_3 = \Delta y_2(x+1) - \Delta y_2(x) = y(x+3) - 3y(x+2) + 3y(x+1) - y(x)。 \tag{9}$$

原吸收系数数据与三组差分数据组成 $4\times(n-3)$维矩阵 $\boldsymbol{X}$，三阶差分处理实现了对样本容量的扩充，再采用 PCA 法对这组数据进行特征提取。

PCA 法特征提取是通过将多个变量转化为少数几个变量来实现的。每个主成分都是原始变量的线性组合，并且足够表示原始变量的大部分信息。将上一步骤得到的数据矩阵 $\boldsymbol{X}$ 的每列视为一个变量，取其平均数得到均值矩阵 $\boldsymbol{X}_{\text{mean}}$。使用原始数据减去平均数，再对数据进行标准化处理后得到 $\tilde{\boldsymbol{X}}$：

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{X}_{\text{mean}}。 \tag{10}$$

计算协方差矩阵，并求协方差矩阵的 n 个特征值 λ_i 及其对应的特征向量 $\boldsymbol{p}_i$，i 为序号。按照特征值大小降序排列得到

$$\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}, \tag{11}$$

$$\boldsymbol{p} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n\}。 \tag{12}$$

按照特征值大小占特征值总和的百分比确定其累积贡献率，当成分数对应的特征值贡献率达到 90%时，即

$$\lambda_1 + \lambda_2 + \cdots + \lambda_m > 0.9, \tag{13}$$

式中：m 表示主成分的个数($m<n$)。对应的特征矩阵 $\boldsymbol{P}=[\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m]$，特征提取的结果为 $\tilde{\boldsymbol{X}}\boldsymbol{P}$。

2.3　数据建模

在数据建模这一部分主要采用 SVM 的方式建立分类模型，即建立提取的特征与对应物质种类的对应关系。通过映射关系将每种物质提取出的特征向量映射到特征空间，以更好地区分物质。特征向量映射与特征空间之间的映射关系为

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b, \tag{14}$$

式中：$\boldsymbol{x}$ 表示吸收系数光谱数据中提取出来的特征向量；$f(\boldsymbol{x})$表示对应 $\boldsymbol{x}$ 进行映射后的新空间；$\boldsymbol{w}$ 和 b 分别表示映射到的新空间超平面的法向量及对应的截距，即 $\boldsymbol{w}$ 控制分割超平面的方向，b 控制分割超平面的位置。

针对多分类建模问题，设计一对一的分类器。利用该分类器实现映射后，两类样本之间的间距达到最大，即 L 达到最小，L 的表达式为

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2。 \tag{15}$$

为保证能将全部的数据点正确分类在分割面的两侧，L 需满足

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n}\beta_i[y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1], \tag{16}$$

式中：β_i 表示拉格朗日乘数，$\beta_i \geqslant 0$；$\boldsymbol{x}_i$ 表示要分类的数据点；y_i 表示根据映射函数得到的值。引入核函数以解决高维计算困难的问题，主要对两个向量进行内积的计算。在此选择径向基函数作为核函数。核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$可表示为

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma\,|\boldsymbol{x}_i - \boldsymbol{x}_j|^2), \tag{17}$$

式中：γ 为径向基函数的宽度参数；$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别表示序号为 i 和 j 对应的向量。通过核函数映射后，(16)式可写为

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n}\beta_i\{y_i[\boldsymbol{w}^{\mathrm{T}}K(\boldsymbol{x}_i) + b] - 1\}。 \tag{18}$$

使(18)式中 L 最小的 $\boldsymbol{w}$ 和 b 即为最佳模型的映射函数参数。

2.4　识别方法

本小节主要介绍有机化合物的识别方法。假设测试样本包括库内数据和库外数据，先用径向基神经网络对测试样本进行数据库内外的判别，通过径向基神经网络的输出判断该物质是否为已知物质的一种。输出值为 1，表示未知物质在已知的数据库中；输出值为-1，则表示未知物质不在已知的数据库里。

对于识别结果为数据库内物质的样本情况，根据建立的 SVM 模型给出该样本的具体类别；对于识别结果为数据库外物质的样本，由于没有相关物质的数据信息，暂时不能给出具体的类别。

3　实验过程

3.1　数据获取

数据获取部分首先是进行样品的准备。选取

20 种物质，分别对其进行研磨处理。具体步骤如下：用酒精将研磨棒和研磨钵擦拭干净，称取 400 mg的样品放入研磨钵内，将其研磨成粉末状，再称取 800 mg 的聚乙烯粉末放入研磨好的样品中，搅拌使其充分混合。对混合后的样品进行压片处理，固定成型为厚度在 1.00～1.60 mm 之间、直径为15 mm左右的圆形薄片。

在太赫兹时域光谱数据的测量过程中，首先检查设备的光路系统是否完好，并调试仪器。然后不断通入氮气，以排除空气中水蒸气对太赫兹波的吸收。实验测量中，测试时间为 40 ps，宽频太赫兹波段为 0～3 THz。接着测量两组不放置样品时的太赫兹信号，作为背景环境下的参考信号。最后依次将每一个样品固定在支架上，使太赫兹波能够垂直透射其样本中心，分别对各个样品进行测量，记录其太赫兹时域光谱数据。

3.2　数据预处理

在数据预处理这一部分，主要对 20 种有机酸和危险品等化学物质的时域光谱进行傅里叶变换处理，包括熵炸药（TATP）、黑火药（BP）、硝酸脲（UN）、硝酸铵（AN）、草酸（脱水）、草酸（二水）、硬脂酸、硬脂酸钠、柠檬酸、苯甲酸、柠檬酸钠、尿嘧啶和山梨酸等。根据频域谱和 2.1 节中的(1)～(6)式计算光学参数，得到其各自的吸收光谱图，并通过小波变换及三次样条曲线进行平滑处理。

首先，基于样品和背景的时域光谱并通过傅里叶变换计算得到其对应的频域谱。不同的有机化合物的时域光谱及其对应的频域谱不同，图 2 描述了其中一组苯甲酸及背景信号的时域谱和频域谱，图中实线代表背景信号，虚线代表样品信号。由于样品对太赫兹波的吸收作用，太赫兹波在透射样品后出现一定的时间延迟，电场强度也有一定的下降。根据频域谱计算不同物质的吸收光谱，并进行对比分析，实验中选择 0.2～2.5 THz 频率区间内的吸收光谱数据作为特征区间。小波变换部分的主要目的是滤除曲线上的噪声，选择 sym6 小波作为小波基函数，采用固定阈值的方式去除高频干扰。采用三次样条插值拟合的方法平滑数据点，使频率分辨率统一为 0.01 THz。

图 3 是 AN、BP、TATP 和 UN 这 4 种危险化合物在 0.2～2.5 THz 频率范围内的吸收光谱图，可以看出不同种类物质的吸收系数谱图也不相同，这为物质的识别提供了参考。其中AN和UN有比较明显的吸收峰，而 BP 和 TATP 的吸收谱没有比较明显的吸收峰，但在数值上有一定的差别。

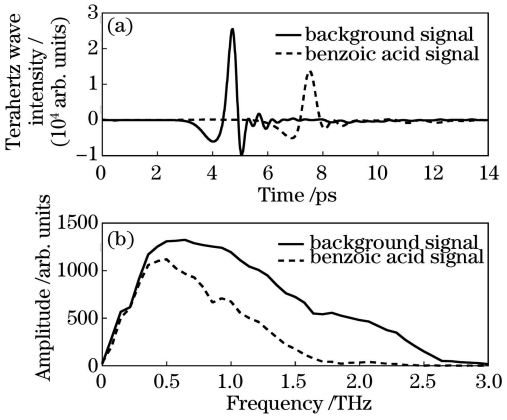

图 2　苯甲酸和背景信号的时域谱和频域谱。(a)时域谱；(b)频域谱

Fig. 2　Time-domain and frequency-domain spectra of benzoic acid and background signal. (a) Time-domain spectra; (b) frequency-domain spectra

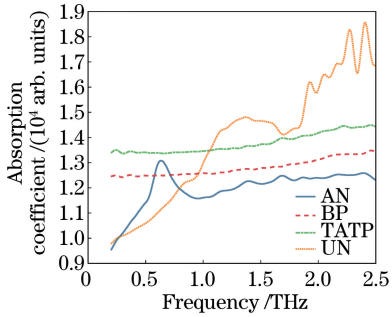

图 3　4 种危险物质的吸收系数图

Fig. 3　Absorption coefficient diagrams of four dangerous substances

3.3　特征提取

特征提取主要是对上一步骤得到的吸收系数谱进行特征提取。基于差分-PCA 方法进行降维的实验过程如下：首先对吸收系数谱数据进行样本容量的扩充，计算每个样本的吸收系数谱线（数据为 1×231 维），根据原始吸收系数谱数据计算一阶差分数据（1×230 维），依次计算其二阶差分数据和三阶差分数据，并将其与原始吸收系数构成 4×228 维矩阵。图 4 展示了苯甲酸的一阶差分数据光谱图。图 5 展示了 4 种危险品物质的一阶差分数据的对比图。为了更好地区分曲线，图 5 采用了双坐标的形式，横坐标为频率，纵坐标为一阶差分数据，其中左纵坐标对应 AN 和 UN 两种物质，右纵坐标对应 BP 和 TATP 两种物质。

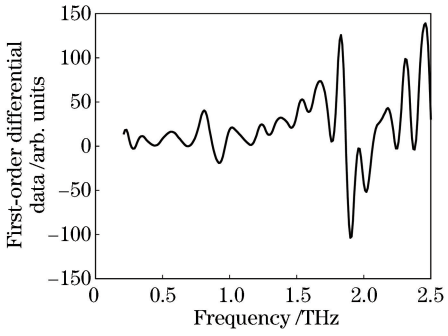

图 4　苯甲酸的一阶差分数据图

Fig. 4　First-order differential data of benzoic acid

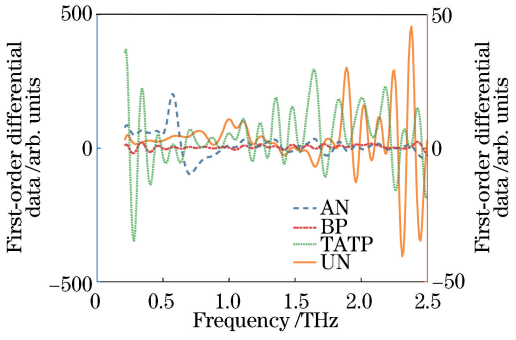

图 5　4 种危险物质的一阶差分数据图

Fig. 5　First-order differential data of four dangerous substances

表 1　基于差分-PCA 降维的特征提取结果

Table 1　Feature extraction results based on differential PCA

Substance	One-dimensional	Two-dimensional	Three-dimensional	Four-dimensional
TATP	−13.60	4.52	4.54	4.54
Black powder	−14.50	4.83	4.83	4.83
UN	−15.59	5.19	5.20	5.20
AN	−15.94	5.29	5.33	5.33
Dehydrated oxalic acid	−13.09	4.34	4.38	4.37
Dihydrated oxalic acid	−13.82	4.59	4.62	4.62
Stearic acid	−10.77	3.58	3.60	3.60
Sodium Stearate	−11.47	3.81	3.83	3.83
Citric acid monohydrate	−13.28	4.4	4.44	4.44
Citric acid dehydration	−15.59	5.17	5.21	5.21
Dihydrated sodium citrate	−10.94	3.64	3.65	3.65
Dehydrated sodium citrate	−14.81	4.93	4.94	4.94
Sodium citrate	−10.26	3.40	3.43	3.43
Uracil	−11.68	3.88	3.91	3.90
Sorbic acid	−14.38	4.76	4.81	4.81

对扩充后的数据采用 PCA 法进行提取特征。通过计算发现：当主成分为 1 时，累积贡献率可以达到 99%。因此，实验中选择用 PCA 法将 4×228 维数据矩阵降维到 4×1 维矩阵。表 1 显示了基于差分-PCA 降维的特征提取结果。第 1 列为有机化合物的名称，第 2～5 列为提取出的代表该物质的 4 维数据。

3.4　实验结果及分析

3.4.1　模型测试实验结果

本节基于 3.3 节中提取出的物质特征和对应的物质类别，并利用 SVM 法进行建模。在此选择了包括 4 种危险品在内的 15 种物质的 30 个样本，所选择的物质主要包括危险品以及一些常用的食品添加剂，也包括了三组不同化学状态下的物质，如水柠檬酸与脱水柠檬酸。在模型训练中，选择 15 个样本作为模型训练集，另外 15 个样本作为测试集。选用径向基函数作为模型核函数。

在建模过程中，将实际物质的名称转换为数学表达式，即用数字 1，2，…，15 代表 15 种物质。输入为 3.3 节中提取的特征数据，输出为目标值，如 1 或 2 等。SVM 建模的过程即为建立物质特征与物质类别的对应输出关系。

在建模训练中，对每两种物质建立一对一的分类模型，则对于 15 种物质，建立的分类模型总数为 105。SVM 训练的参数列在表 2 中。在测试过程中，选用吸收峰处的频率和幅值(F&A)作为特征，以及选择线性判别分析(LDA)降维方法得到的特征作为 SVM 的输入，并对建模后的效果进行了对比。对测试集的 15 种样本的模型识别效果进行验证，其识别结果如表 3 所示。表中第 1 列为测试样本个数，第 2 列为特征提取方法，第 3 列为识别出的样本个数。在这三种方法中，对于利用吸收峰的频率-幅值进行特征提取建模的方法，由于部分物质没有明显的吸收峰，在判断峰值点时易造成较大误差，

从同一种物质的不同样本提取出的特征存在较大差别，这不利于后续的建模与识别。对比识别率相对较高的差分-PCA 和 LDA 法可知，使用差分-PCA 法时的降维识别效果最好。

表 2　SVM 训练的参数

Table 2　Parameters obtained from training in SVM

Method	Differential PCA	LDA	F&A
Penalty factor	8	32	16
Parameter γ	4.88×10^{-4}	16	16

表 3　验证测试集的 15 种样本的模型识别效果的识别结果

Table 3　Model identification results of 15 kinds of samples in test dataset

Number of test samples	Feature extraction method	Number of identified samples	Accuracy rate /%
15	Differential PCA	15	100
15	LDA	13	86.67
15	F &A	9	60

3.4.2　物质识别结果及分析

在第 3.4.1 节中建立了 15 种物质的分类模型，并将用于建模的物质数据信息作为已知物质数据库内的信息，通过验证其测试样本，识别出属于数据库内的物质。本节主要对 3.4.1 节中验证测试集的 15 种样本的模型识别效果的识别结果进行分析。对于已知物质数据库内的物质，可以根据模型给出物质的类别，而对数据库外的物质，由于没有相关物质的信息，无法给出其类别。因此，需要先对数据库外的物质进行判别排除。

首先建立径向基神经网络对数据库外的物质进行判别。对于一个需识别其物质类别的未知样本，需先判断它是否为已知数据库内的物质，如果是则用已建立模型判断其具体类别，如果不是则无法给出其具体的类别。

根据 3.4.1 节用于建模的 15 种物质数据对径向基函数神经网络进行训练，输入为采用差分 GPCA 法提取的特征。通过迭代修改权值、隐藏层数据中心点以及宽度参数，使数据库内物质输出为 1，训练输出值为 1 表示物质在数据库中。

在识别方法的测试实验过程中，选择了 15 种与已知数据库内物质种类相同的样本以及 5 种数据库内没有的物质作为样本测试集，包括衣康酸、季戊四醇、柠檬酸三钠、硬脂酸和十八醇，对其识别效果进行验证。径向基神经网络识别效果如表 4 所示。由于库外柠檬酸三钠与库内黑火药的吸收系数曲线整体较为近似，且径向基神经网络的训练测试存在一定误差，测试实验中存在错误识别的情况。在整体效果上，所提方法能够先排除数据库外的物质，对于识别为库内的物质，根据模型测试过程的结果，可以实现对物质种类的正确识别。

表 4　径向基神经网络识别效果

Table 4　Recognition performance of radial basis neural network

Sample test set	Number of test samples	Number of identified samples	Accuracy rate /%
Substance in database	15	14	93.33
Substance out of database	5	4	80.00

4　结　论

针对单一物质的吸收系数谱线不利于进行特征提取的问题，提出了一种基于差分-PCA-SVM 的有机化合物识别方法。该方法先利用差分数据的方式扩充单一物质的谱线数据的样本容量，并结合 PCA 法进行特征提取，然后利用 SVM 建立数学模型，最后实现了 15 种有机化合物的正确识别。实验中对比了所提出的方法、线性判别分析法和特征吸收峰幅值频率-幅值法的识别效果。结果表明：线性判别分析法提取特征需要对数据库内几种物质同时进行特征提取，在更新数据库内物质时需要重新计算代表物质的特征；采用吸收峰的频率-幅值法时，由于在提取峰值点的位置时容易出现误差，建模识别的效果不理想；所提出的基于差分-PCA-SVM 的有机化合物识别方法的正确识别率最高。

参　考　文　献

[1] Cheng D Z. Controllability of switched bilinear systems [J]. IEEE Transactions on Automatic Control, 2005, 50(4): 511-515.

[2] Rudd J V, Johnson J L, Mittleman D M. Quadrupole radiation from terahertz dipole antennas [J]. Optics Letters, 2000, 25(20): 1556-1558.

[3] Chen T, Zhang C J, Xu C P. Study onterahertz time-domain spectroscopy of tartaric acid isomers [J]. Laser & Optoelectronics Progress, 2017, 54(8): 081202.
陈涛, 张超杰, 许川佩. 酒石酸异构体的太赫兹时域光谱研究[J]. 激光与光电子学进展, 2017, 54(8): 081202.

[4] Trofimov V, Varentsova S. An effective method for substance detection using the broad spectrum THz

signal: a "terahertz nose"[J]. Sensors, 2015, 15(6): 12103-12132.

[5] Ni J P, Shen T, Zhu Y, *et al*. Terahertz spectroscopic identification with diffusion maps[J]. Spectroscopy and Spectral Analysis, 2017, 37(8): 2360-2364.
倪家鹏, 沈韬, 朱艳, 等. 基于扩散映射的太赫兹光谱识别[J]. 光谱学与光谱分析, 2017, 37(8): 2360-2364.

[6] Liu K, Li B, Zeng X X, *et al*. Method of terahertz time-domain spectroscopy classification based on mani-fold learning and support vector machine[J]. Computer Engineering and Applications, 2015, 51(24): 141-144, 175.
刘坤, 李飚, 曾祥鑫, 等. 基于流形学习和支持向量机的太赫兹谱分类[J]. 计算机工程与应用, 2015, 51(24): 141-144, 175.

[7] Xie Q, Yang H R, Li H G, *et al*. Explosive identification based on terahertz time-domain spectral system[J]. Optics and Precision Engineering, 2016, 24(10): 2392-2399.
解琪, 杨鸿儒, 李宏光, 等. 基于太赫兹时域光谱系统的爆炸物识别[J]. 光学 精密工程, 2016, 24(10): 2392-2399.

[8] Wang K, Wang H Q, Yin Y, *et al*. Pigment spectral matching recognition method based on adaptive edit distance[J]. Laser & Optoelectronics Progress, 2018, 55(11): 113004.
王可, 王慧琴, 殷颖, 等. 基于自适应编辑距离的颜料光谱匹配识别方法[J]. 激光与光电子学进展, 2018, 55(11): 113004.

[9] Zhang W T, Li Y W, Zhan P P, *et al*. Recognition of transgenic soybean oil based on terahertz time-domain spectroscopy and PCA-SVM[J]. Infrared and Laser Engineering, 2017, 46(11): 1125004.
张文涛, 李跃文, 占平平, 等. 基于太赫兹时域光谱技术与 PCA-SVM 的转基因大豆油鉴别研究[J]. 红外与激光工程, 2017, 46(11): 1125004.

[10] Pohl A, Deßmann N, Dutzi K, *et al*. Identification of unknown substances by terahertz spectroscopy and multivariate data analysis[J]. Journal of Infrared, Millimeter, and Terahertz Waves, 2016, 37(2): 175-188.