# 基于扩散映射的太赫兹光谱识别

倪家鹏1,沈 韬1,2\*,朱 艳2,李灵杰1,毛存礼1,余正涛1

昆明理工大学信息工程与自动化学院,云南昆明 650504
昆明理工大学材料科学与工程学院,云南昆明 650093

**摘 要**特征提取对于太赫兹光谱识别来说至关重要。传统方法是通过人工选取太赫兹光谱中差异性较大的吸收峰作为特征进行光谱识别,但当部分物质在太赫兹波段没有明显波峰、波谷等光谱图形特征时,这种方式便不再适用。为此,研究人员利用统计学习与机器学习方法对高维太赫兹光谱数据进行降维和特征提取。由于物质的太赫兹光谱数据各维度呈现非线性,尤其是当不同物质的太赫兹光谱曲线整体非常相似时,线性处理方法易产生较大误差。针对这一问题,提出了一种基于扩散映射(DM)的太赫兹光谱识别方法。扩散映射能在保持数据内在几何结构的同时对其进行非线性降维,提取的流形特征区分度较高,对数据还有聚类效果。首先用 S-G 滤波器对 Alloxazine 等 10 种物质的太赫兹光谱样本进行滤波,并用三次样条插值法对截取相同频段后的光谱样本进行统一分辨率处理;然后利用 DM 将高维太赫兹光谱数据映射到低维特征空间并提取太赫兹光谱的流形特征;最后用多分类支持向量机(M-SVM)对十种物质的太赫兹透射光谱进行分类。实验结果表明,相比于主成分分析(PCA)和等距映射(ISOMAP),使用 DM 提取的太赫兹光谱流形特征具有更高的区分度,而且 DM 可以直接得到太赫兹光谱数据本征维数的估计值,这为相似太赫兹光谱的快速精准识别提供了一条新的途径。

关键词 太赫兹光谱; 流形学习; 谱方法; 扩散映射; 非线性降维 中图分类号: O433.5 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2017)08-2360-05

引 言

有机化合物及生物大分子在太赫兹波段具有指纹特性, 而且太赫兹波的光子能量很低,用于物质检测时不会发生有 害的光致电离<sup>[1]</sup>。随着太赫兹辐射光源及探测器等硬件方面 的迅速发展,太赫兹在物质识别和无损检测等领域的应用也 逐渐增多。国内外许多实验表明,很多化学或生物物质在太 赫兹波段存在明显的吸收峰<sup>[2-3]</sup>,基于此特性,研究人员可 快速有效地对物质进行识别。如朱思原等<sup>[4]</sup>、李利龙等<sup>[5]</sup>利 用太赫兹光谱中差异很大的吸收峰分别对四种青霉素类抗生 素、七种植物油和两种调和油成功进行识别。然而还有部分 物质在太赫兹波段没有明显波峰、波谷等光谱图形特征,此 类物质便很难运用传统的人工选择的方式来确定合适的特 征。加之物质的太赫兹光谱数据维度很高,若不对光谱数据 进行降维和特征提取,太赫兹光谱的识别将会异常困难且要 耗费大量时间。近年来,统计学习与机器学习方法被用来分 析和处理太赫兹光谱数据。Zhang 等<sup>[6]</sup>使用主成分分析法 (principal components analysis, PCA)对八种转基因棉花种 子的太赫兹光谱进行主成分分解,提取贡献率较大的前三个 主成分作为太赫兹光谱特征用于区分上述物质; Zhan 等<sup>[7]</sup> 使用聚类分析(cluster analysis, CA)结合主成分分析法对七 种来自不同国家和油田的原油的太赫兹光谱数据进行计算, 用层级树状图和第一主成分得分柱状图将这七种原油的异同 展现出来并进行识别。主成分分析法通过对原始数据进行线 性变换,选择方差贡献率较大的主成分实现维数约简,但太 赫兹光谱数据各维度呈现非线性,这种线性处理方法易产生 较大误差。更为关键的是,以上对太赫兹光谱的处理方法都 是针对光谱曲线整体差异较大的情况,当不同物质的谱线十 分相似时便不能做到准确鉴别。

扩散映射(diffusion maps, DM)<sup>[8]</sup>是近些年学者们提出 的一种较新颖的流形学习谱方法,该方法能在保持数据内在 几何结构的同时对其进行非线性降维,最终提取到的流形特 征具有较高的区分度,而且对数据还有聚类效果,所以比较

收稿日期: 2016-07-15,修订日期: 2016-11-08

基金项目:国家自然科学基金项目(61671225,61302042)资助

作者简介: 倪家鹏, 1991年生, 昆明理工大学信息工程与自动化学院硕士研究生 e-mail: nijiapeng618@126.com

<sup>\*</sup> 通讯联系人 e-mail: shentao@kmust.edu.cn

适用于太赫兹光谱的分类与识别。本文针对部分化合物透射 光谱无明显图形特征、谱线整体相似而难以识别的问题,提 出一种基于扩散映射的太赫兹光谱识别方法。首先利用扩散 映射将太赫兹光谱数据映射到低维特征空间,然后在特征空 间中提取太赫兹光谱的流形特征,最后采用多分类支持向量 机(multi-class support vector machine, M-SVM)对太赫兹光 谱进行分类与识别。

## 1 基于扩散映射的太赫兹光谱识别方法

## 1.1 理论基础

### 1.1.1 流形学习谱方法

流形学习(manifold learning)<sup>[9]</sup>是基于拓扑学与微分几 何的非线性高维数据处理方法,旨在发掘高维数据中可能存 在的嵌套性低维光滑流形。在流形学习中, 谱方法<sup>[10]</sup>有着广 泛的应用。从几何角度分析, 谱方法的思想是在去除冗余成 分的同时尽可能保留数据在高维空间中的局部几何信息。近 些年,在机器学习、数据挖掘和计算机视觉等领域,基于谱 思想的流形学习方法[11-12]具有很高的关注度。通常将流形学 习谱方法分为线性和非线性两大类别。线性谱方法,如 PCA、多维尺度分析(multidimensional scaling, MDS)等传统 的线性投影技术;非线性谱方法,如等距映射(isometric mapping, ISOMAP)、局部线性嵌入(locally linear embedding, LLE)、局部切空间排列算法(locality tangent space alignment, LTSA)等基于谱图的谱方法。这些方法都是用关 联矩阵的最大(或最小)特征谱所对应的特征向量的数目作为 高维数据本征维数的估计值[13]。本征维数反映了表征高维 数据所需独立坐标或自由参数的最小数目,估计的准确性将 决定提取的低维特征对于保留高维数据有用信息的效果。 1.1.2 扩散映射

扩散映射是 2006 年由 Coifman 等提出的一种流形学习 谱方法,在由数据点构建的图上定义一个 Markov 随机游走, 在一定时间步长的随机游走之后得到任意两个数据点之间接 近度的距离函数,并定义为扩散距离。扩散映射的主要思想 就是在尽可能保持扩散距离的前提下获取数据的低维流形结 构。

对给定的 D 维数据集  $X = \{x_1, x_2, \dots, x_N\}, x_i \in R^D$  (*i* =1, 2, …, N), 从中提取 d 维(0 < d < D)的流形特征 Y。首 先根据数据点构建一个与之对应的带有权值的图 G, 使用高 斯核函数计算图中节点之间的权值,并由所有权值组成一个 权值矩阵 W, 其元素为

$$\boldsymbol{W}_{ij} = \exp\left(-\frac{\parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2}{2\sigma^2}\right) \tag{1}$$

其中, σ为高斯核宽度参数。

然后对矩阵 W 进行归一化处理,使 W 的每行元素之和 单位化为 1,从而得到归一化后的权值矩阵 P<sup>(1)</sup>,其元素为

$$\mathbf{P}_{ij}^{(1)} = \frac{\mathbf{W}_{ij}}{\sum_{i} \mathbf{W}_{ik}} \tag{2}$$

**P**<sup>(1)</sup>是数据点之间的一步转移概率矩阵,其中每一行表

示对应数据点随机游走到其他数据点的概率。而对于 t 步转移,即经过  $t 步 随机 游走之后的转移概率矩阵为 <math>P^{(i)} = (P^{(1)})^{i}$ ,此时  $x_{i} = x_{j}$ 之间的扩散距离定义为

$$D^{(t)}(x_i, x_j) = \sqrt{\sum_k \frac{(\mathbf{P}_{ik}^{(t)} - \mathbf{P}_{jk}^{(t)})^2}{\varphi(x_k)}}$$
(3)

其中,  $\varphi(x_i) = \frac{m_i}{\sum_i m_j}$ ,  $m_i = \sum_j p_{ij}$ 。随着 t 的增加, 数据集

的局部几何信息被整合在一起。

从式(3)可以看出,图中节点越密集,数据点之间随机 游走的步伐就越多,扩散距离就越小。扩散距离考虑了两点 之间所有路径的作用,相比测地距离具有更强的鲁棒性。

最后,以保持扩散距离为前提,提取数据的低维流形 Y。 根据 Markov 随机游走的谱图理论,可由(4)式的 d 个主特征 向量组成 Y。

$$\boldsymbol{P}^{(t)}\boldsymbol{Y} = \boldsymbol{\lambda}\boldsymbol{Y} \tag{4}$$

值得注意的是,数据点构建的是全局连接图,谱分解得 到的最大特征值(即 $\lambda_1$ =1)是平凡的,因此要舍弃其对应的 特征向量 $\nu_1$ 。那么,Y可由其余d个主特征向量构成

$$\mathbf{Y} = (\lambda_2 \, \nu_2 \, , \, \lambda_3 \, \nu_2 \, , \, \cdots \, , \, \lambda_{d+1} \, \nu_{d+1})^{\mathrm{T}} \tag{5}$$

#### 1.2 模型的构建

1.2.1 流形特征提取

流形给出了一种表现同类事物共性的新形式<sup>[14]</sup>,即可 由其样本数据所位于的流形结构来体现。流形学习谱方法可 以在低维特征空间中体现流形上不同区域的差异,提取的流 形特征也有利于进一步的聚类或分类任务。传统的线性谱方 法,如 PCA 在处理非线性的太赫兹光谱数据时,不能有效逼 近光谱数据点所位于的流形。因此,本文利用扩散映射这种 非线性流形学习谱方法提取太赫兹光谱内在流形特征,考虑 到实际计算的复杂度和准确性,一般采用一步转移概率矩阵 (即 *t*=1)。具体步骤如下:

Step 1 对给定的太赫兹光谱数据进行预处理,构建高 维观测样本集  $X = \{x_1, x_2, \dots, x_N\}, x_i \in R^D$  (*i*=1, 2, …, *N*);

Step 2 使用扩散映射方法得到样本集的 Markov 转移 概率矩阵  $P^{(1)}$ 并对其进行谱分解。将特征值从大到小排列:  $\lambda_1 = 1 > \lambda_2 \ge \cdots \ge \lambda_N \ge 0$ ,舍弃平凡特征值  $\lambda_1 = 1$  及其对应的 常量特征向量  $\nu_1$ ;

Step 3 根据以下判定规则得到高维太赫兹光谱数据本 征维数的估计值,进而在该估计值所对应维数的特征空间中 提取太赫兹光谱的流形特征。

(1)若  $P^{(1)}$ 的前 d+1(d+1 < N)个特征值之间数值相差 不大且都接近于 1,同时  $\lambda_{d+1} \ge \lambda_{d+2}$ ,则本征维数的估计值为 d;

(2)若  $P^{(1)}$ 的特征值不符合(1)的条件,则计算相邻特征 值的差值,若某个特征值  $\lambda_{d+2}$ 突然减小,即  $\lambda_{d+1} - \lambda_{d+2} \gg \lambda_k$  $-\lambda_{k+1}(d+1 < k \leq N)$ ,且  $\lambda_{d+2}$ 的值也很小,同样可得到本征 维数的估计值为 d。

## 1.2.2 分类

支持向量机(SVM)是一种二分类的分类器,它的机理是 在两类样本之间构建一个最优分类超平面,即在保证分类精 度的条件下使该超平面两侧与最近样本点的距离最大化。针 对多分类问题,采用一对一(one-versus-one)策略,在任意两 种待分类样本之间均设计一个二分类器,并最终构造一个多 分类支持向量机(M-SVM)。分类时,根据每一个 SVM 的判





定结果对输入样本进行投票,得票最多的类别就是输入样本 所属的类别。于是,用提取的流形特征结合多分类支持向量 机构建分类模型,最终实现太赫兹光谱的识别。

采用扩散映射提取太赫兹光谱的流形特征与利用 M-SVM对太赫兹光谱进行识别的流程如图1所示。

## 2 实验与结果

#### 2.1 数据

由于不同太赫兹光谱系统设定的带宽和采样频率等参数 并不一致,以及受到实验条件的影响,最终获得的太赫兹光 谱会含有噪声且光谱分辨率不尽相同。针对这些问题,首先 利用 S-G 滤波器对太赫兹光谱进行滤波,然后截取相同频段 的光谱,利用三次样条插值法得到统一分辨率的光谱数据。

实验以 Alloxazine, Anthraquinone, Carbazole, Cuprous oxide, Inositol, Maltotriose, Maltotetraose, Maltopentaose, Malthexaose 和 Maltoheptaose的太赫兹透射光谱为例,对这 十种物质各自测量 100 次得到的太赫兹光谱样本进行滤波, 截取 0.9~6 THz 频段,经过三次样条插值处理后得到统一 分辨率的光谱数据集,每条光谱曲线的数据点均为 6 349 个。 从每种物质滤波后的 100 个光谱样本中随机抽取一条光谱曲 线如图 2 所示。





## 2.2 设置

根据光谱曲线的整体差异,将这十种物质的太赫兹透射 光谱分为 dataset-1 和 dataset-2。dataset-1 中五种物质的光谱 曲线差异很大,谱线均有明显的波峰(谷); dataset-2 中五种 物质的光谱曲线相似度很高,谱线均无明显图形特征。然后 由 dataset-1 与 dataset-2 组成 dataset-3,即 dataset-3 是这十 种物质的光谱集合。

经过迭代实验分析, DM 中高斯核宽度参数 σ 设定为

0.7。对样本集 dataset-1, dataset-2 和 dataset-3 的转移概率 矩阵进行谱分解,分别将 3 组特征值中最大的前 15 个特征 值从大到小排序为:

 $\lambda_{dataset-1} = \begin{bmatrix} 1.0, 0.999 \ 9, 0.999 \ 8, 0.999 \ 6, 0.999 \ 3, \\ 0.360 \ 9, 0.360 \ 8, 0.360 \ 3, 0.360 \ 2, 0.360 \ 1, 0.130 \ 6, \\ 0.130 \ 5, 0.130 \ 4, 0.130 \ 3, 0.130 \ 2, \cdots \end{bmatrix};$ 

$$\begin{split} \lambda_{\text{dataset-2}} = [ \ 1. \ 0, \ 0. \ 999 \ 9, \ 0. \ 998 \ 1, \ 0. \ 969 \ 1, \ 0. \ 907 \ 1, \\ 0. \ 257 \ 6, \ 0. \ 186 \ 7, \ 0. \ 186 \ 5, \ 0. \ 179 \ 3, \ 0. \ 134 \ 7, \ 0. \ 079 \ 1, \end{split}$$

0.056 9, 0.050 3, 0.050 2, 0.039 3, ...];

 $\lambda_{dataset-3} = \begin{bmatrix} 1. \ 0, \ 0. \ 999 \ 9, \ 0. \ 999 \ 7, \ 0. \ 999 \ 6, \ 0. \ 999 \ 4, \\ 0. \ 999 \ 1, \ 0. \ 998 \ 5, \ 0. \ 998 \ 1, \ 0. \ 969 \ 1, \ 0. \ 907 \ 1, \ 0. \ 361 \ 0, \\ 0. \ 360 \ 8, \ 0. \ 360 \ 3, \ 0. \ 360 \ 2, \ 0. \ 360 \ 1, \ \cdots \ ]_{\circ}$ 

将平凡的最大特征值1舍弃之后可以很明显看出,  $\lambda_{dataset-1}$ 和 $\lambda_{dataset-2}$ 的第2到第5个特征值相差不大且都接近于 1,同时 $\lambda_5 - \lambda_6 \gg \lambda_6 - \lambda_7$ ;  $\lambda_{dataset-3}$ 的第2到第10个特征值相差 不大且都接近于1,并且 $\lambda_{10} - \lambda_{11} \gg \lambda_{11} - \lambda_{12}$ 。根据1.2.1小节 所述映射理论,dataset-1,dataset-2和 dataset-3中的太赫兹 光谱数据本征维数的估计值分别为4,4和9,这样便可在相 应维数的低维空间中提取太赫兹光谱的流形特征。

作为对比实验,分别采用流形学习的线性谱方法 PCA 以及非线性谱方法 ISOMAP 对太赫兹光谱数据进行降维处 理并提取流形特征。PCA 主成分贡献率设置为 0.96; ISO-MAP 所要估计的本征维数与 DM 的结果保持一致。

从每种物质的 100 个光谱样本中随机抽取 30 个样本作 为测试集,剩下的作为训练集。M-SVM 选用径向基核函数, 经过十折交叉验证,确定惩罚系数 c=1 024,核函数参数 g =1 024。实验时,分别将训练集的原始数据、PCA 主成分、 ISOMAP 以及 DM 提取的流形特征输入 M-SVM,建立分类 模型,然后用测试集对分类模型进行检测,重复检测 20 次, 分类准确率由分类正确样本数与总样本数的比值得出。

不同谱方法提取的流形特征的分类准确率如表1所示。

表 1 不同谱方法提取的流形特征的分类准确率 Table 1 Classification accuracy of manifold features extracted by different spectral methods

	准确率/%		
	dataset-1	dataset-2	dataset-3
原始数据	86.83	88.17	85.33
PCA	92.17	90	88.83
ISOMAP	95.5	94	92.5
DM	100	98.5	96

由表1可知,相比于原始数据,使用 PCA, ISOMAP 和 DM 提取的光谱流形特征在 M-SVM 分类器中的识别率较 好,M-SVM 分类器在 dataset-1 上的识别率要高于 dataset-2。由于 dataset-2 中五种物质的太赫兹光谱曲线整体十分相 似,导致 M-SVM 分类器更容易产生误判。对比实验结果表 明,DM 提取的光谱流形特征区分度更高,其中,PCA 在处 理非线性的太赫兹光谱数据时,会使光谱样本之间产生较多 混叠,尤其是当谱线整体差异很小时,这种情况更为严重, 这也就导致得到的主成分的分类效果很不理想。

实验时发现,在不同本征维数估计值所对应维数的特征 空间中,ISOMAP提取的流形特征的识别率有所差异。图 3

### References

为 ISOMAP 和 DM 分别将 3 个太赫兹光谱数据集映射到低 维空间时,不同本征维数估计值对 M-SVM 分类器识别率的 影响。



Fig. 3 Classification accuracy of manifold features extracted from different dimensions of space

数据的本征维数是对数据进行建模所需自由变量的最少数目。过高估计将会引入许多伪成分,过低估计则会丢失大量有价值的信息。本征维数估计准确与否将直接影响在低维空间中提取的特征的效果。从图 3 中可以看出, ISOMAP 需要通过迭代的方式调整本征维数的估计值以使得分类结果最优,这不利于太赫兹光谱的快速精准识别,而 DM 则可以直接得到 3 个太赫兹光谱数据集中光谱数据本征维数的估计值分别为 4,4 和 9,均低于 ISOMAP 的最优估计值 11,10 和 12。而且 DM 在相应低维空间提取的流形特征的识别效果也更好,这是因为 DM 中扩散距离考虑了两个数据点之间所有路径的贡献,相比 ISOMAP 中的测地距离对噪声干扰的鲁棒性更强。

## 3 结 论

针对部分化合物太赫兹透射光谱无明显图形特征、谱线 整体相似而难以识别的问题,本文提出一种基于扩散映射的 太赫兹光谱识别方法。利用扩散映射这种流形学习谱方法直 接对太赫兹光谱数据集进行学习,提取太赫兹光谱的流形特 征,使用多分类支持向量机对太赫兹光谱进行识别。相比流 形学习线性谱方法,扩散映射能够有效发掘高维非线性太赫 兹光谱数据内嵌的低维流形结构,并能直接得到其本征维数 的估计值。该方法提取的流形特征区分度较高,对于谱线整 体差异较大或是非常相似的太赫兹光谱均有很好的识别效 果。所以,基于扩散映射的太赫兹光谱识别方法在未来光谱 数据识别领域具有很好的应用前景。

[2] MENG Zeng-rui, ZHANG Wei-bin, DU Yu(孟增睿,张伟斌,杜 宇). Acta Physica Sinica(物理学报), 2015, 64(7): 214.

<sup>[1]</sup> WANG Guo, WANG Wei-ning(王 果,王卫宁). Acta Physico-Chimica Sinica(物理化学学报), 2012, 28(7): 1579.

- [3] Qin J, Xie L, Ying Y. Analytical Chemistry, 2014, 86(23): 11750.
- [4] ZHU Si-yuan, ZHANG Man, SHEN Jing-ling(朱思原,张 曼,沈京玲). Infrared and Laser Engineering(红外与激光工程), 2013, 42 (3): 626.
- [5] LI Li-long, XIANG Yang, WU Lei(李利龙,向 洋,吴 磊). High Power Laser and Particle Beams(强激光与粒子束), 2013, 25(6): 1566.
- [6] Zhang W, Nie J, Tu S. Optical and Quantum Electronics, 2015, 47: 3533.
- [7] Zhan H, Wu S, Bao R. Fuel, 2015, 143: 189.
- [8] Coifman R R, Lafon S. Applied and Computational Harmonic Analysis, 2006, 21(1): 5.
- [9] Luca Rossi, Andrea Torsello, Edwin R Hancock. Pattern Recognition, 2015, 48(11): 3357.
- [10] HUANG Yun-juan, LI Fan-zhang(黄运娟,李凡长). Journal of Software(软件学报), 2013, 24(11): 2656.
- [11] Izquierdo-Verdiguier E, Jenssen R, Gomez-Chova L. Neurocomputing, 2015, 149(5): 1299.
- [12] Wang L, Wang K, Li R. Iet Computer Vision, 2015, 9(5): 655.
- [13] Horvath D, Ulicny J, Brutovsky B. Connection Science, 2016, (1): 1.
- [14] Lisboa P J G, Martin-Guerrero J D, Vellido A. Expert Systems with Applications, 2015, 42(22): 8982.

## **Terahertz Spectroscopic Identification with Diffusion Maps**

NI Jia-peng<sup>1</sup>, SHEN Tao<sup>1, 2</sup>\*, ZHU Yan<sup>2</sup>, LI Ling-jie<sup>1</sup>, MAO Cun-li<sup>1</sup>, YU Zheng-tao<sup>1</sup>

- Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China
- 2. Faculty of Materials Science and Engineering, Kunming University of Science and Technology, Kunming 650093, China

Abstract Feature extraction is the key issue for the identification of terahertz spectroscopy. For the traditional method, it is identified by different absorption peaks as the features that extracted through manual method. However, for many materials, there are no apparent spectral graphics features in the terahertz band, such as peaks, valleys and etc. To this end, the researchers reduce the dimension from the high-dimensional terahertz spectroscopy data and extract the features through statistics learning and machine learning methods. Linear method is easy to cause greater error due to the nonlinear nature of terahertz spectroscopy data, especially when different materials of spectrum curves are very similar. To address this issue, a novel terahertz spectroscopy identification approach with Diffusion Maps (DM) was studied in this paper. Diffusion Maps can realize nonlinear dimensionality reduction while maintaining the internal geometry of the data. In addition, the manifold features extracted by the method have good discrimination and clustering performance. Firstly, S-G filter and cubic spline interpolation were used to smooth and uniform the resolution of terahertz transmission spectra of ten kinds of substances in the same frequency band. Secondly, high-dimensional data of terahertz spectra is mapped to the low-dimensional feature space by using DM so that we can extract the manifold features of terahertz spectroscopy. Finally, a Multi-class Support Vector Machine (M-SVM) classifier is applied to classify these terahertz spectra. Experimental results show that, compared with Principal Component Analysis (PCA) and Isometric Mapping (ISOMAP), manifold features of terahertz spectroscopy extracted by DM have higher degree of differentiation. Besides, DM can get the estimation of intrinsic dimension of terahertz spectra directly. So this proposed method provides a novel approach to identify similar terahertz spectrum quickly and accurately.

Keywords THz spectroscopy; Manifold learning; Spectral method; Diffusion Maps; Nonlinear dimensionality reduction

(Received Jul. 15, 2016; accepted Nov. 8, 2016)

\* Corresponding author