

# 基于稀疏表示的 THz 信号分类方法

李 静<sup>1</sup> 谢维信<sup>1,2</sup> 裴继红<sup>2</sup>

(1. 西安电子科技大学工程学院, 陕西 西安 710071; 2. 深圳大学智能信息研究所, 广东 深圳 518060)

**摘 要:** 基于几何代数的太赫兹时域光谱(Terahertz time domain spectroscopy, THz-TDS)信号分析的研究表明: THz 信号可表示为几何代数结构下的实矢量, 同类 THz 信号矢量位于与其物质相对应的二维特征子空间中, 并且在研究同厚度目标样品或非涅耳损耗可忽略的情况下, 同类 THz 信号矢量间具有线性相关的特性。THz 信号矢量可以用已知同类的信号矢量的线性组合来表示。在此基础上, 本文从信号的稀疏表示方法出发, 采用已知 THz 信号矢量构建冗余字典, 对 THz 信号矢量进行基于冗余字典的稀疏表示, 并将 THz 信号分类问题描述为线性方程组的求稀疏解问题。本文提出了基于稀疏表示的 THz 信号分类方法。该方法首先基于已知类别的 THz 信号矢量构建冗余字典, 然后对待分类信号运用最小化  $l_1$  方法来求解线性方程组中的系数, 最后根据稀疏系数来确定该信号矢量的类别标识。本文详细讨论了该方法中的每个步骤: 基于 THz 信号矢量的几何分布和代数结构特性, 提出了冗余字典的优化构建方法; 通过引入噪声因素, 对信号分类问题的线性方程组模型进行了修正; 在确定类别标识的问题上分别提出了以最大系数和最小残差作为分类依据的方法。实验验证了本文方法的可行性和有效性。

**关键词:** 太赫兹时域光谱信号; 稀疏表示; 压缩感知; 几何代数

**中图分类号:** TP391 **文献标识码:** A **文章编号:** 1003-0530(2012)05-0645-08

## THz-TDS Signal Classification via Sparse Representation

LI Jing<sup>1</sup> XIE Wei-xin<sup>1,2</sup> PEI Ji-hong<sup>2</sup>

(1. School of Electronic Engineering, Xidian University, Xi'an, Shanxi 710071, China;

2. Intelligent Information Institute, Shenzhen University, Shenzhen, Guangdong 518060, China)

**Abstract:** It is demonstrated by the THz-TDS (terahertz time domain spectroscopy) signal analysis based on geometric algebra that: THz signals can be represented as real vectors under the framework of the geometric algebra, vectors corresponding to the same substance belong to the intrinsic 2-dimensional feature subspace of that substance, and especially when the samples studied are of the same thickness value or when the Fresnel loss can be neglected, vectors of the same substance are linearly dependent to each other. Vectors of THz signals can be represented as a linear combination of the signal vectors correspondingly from the same class. Based on that, from the view of the signal sparse representation, THz signal vectors can be represented in an overcomplete dictionary whose base elements are the “known” signal vectors themselves. Such that, the signal classification can be modeled as the problem of finding the optimal sparse solution to linear equations. A signal classification method via the sparse representation is presented. In the method, an overcomplete dictionary is constructed using the “known” THz signal vectors. And for one test signal, the optimal sparse coefficients of the linear equations are obtained efficiently via the  $l_1$ -minimization. Finally, the class of the test sample is determined based on the coefficients. Each step of the method is discussed in detail: an optimal construction method of the overcomplete dictionary is developed based on the geometrical distribution and the algebraic structure properties of signal vectors, the model of the signal classification is modified to account for possibly noise, and the test sample is classified using the criterion of either the maximized coefficient or the minimized residual. Feasibility and effectiveness of the method is confirmed by experiments presented.

**Key words:** THz-TDS; sparse representation; compressive sensing; geometric algebra

## 1 引言

太赫兹(Terahertz,简称 THz)波通常是指频率在 0.1-10THz (1THz =  $10^{12}$ Hz)范围内的电磁波,其波段位于微波和红外线之间,属于远红外线和亚毫米波范畴<sup>[1]</sup>。太赫兹时域光谱(THz time domain spectroscopy, THz-TDS)技术直接测量的是 THz 脉冲的电场随时间演化的电场强度。对 THz 脉冲的时域波形进行傅里叶变换,可以得到电场的振幅和相位信息,从而可以获得测量物质的吸收率、折射率以及复折射率等物理参数。这些参数提供了辨识样品的物理化学成分及结构等的重要信息。

目前对太赫兹时域光谱的表示方法、信号光谱的分类和辨识等研究已经有了一些相关的研究工作。一般是借用标量数据来表示 THz 信号,例如:时域信号最大峰值、信号光谱的波峰或波谷的频率值等。还有的研究应用小波分析方法对 THz 信号进行噪声分析,以降低噪声对分类方法的影响<sup>[2]</sup>。然而这些标量数据处理方法均有其局限性。在实际的应用中,利用波形特征的分析方法对信噪比要求很高,且不适于处理波形较相似的信号;而在小波分析中,最优母函数的选取问题尚未得到很好的解决<sup>[3]</sup>。主分量分析是多元数据分析中的一种有效手段,也被应用于 THz-TDS 信号分类问题的研究中,例如,应用 KL 变换提取主成分作为特征,基于最大欧氏距离进行分类处理<sup>[4]</sup>。还有的研究则是基于图像匹配的方法,将被测物体的时频谱图与数据库中的已知物质的时频谱图(指纹谱)进行匹配,依据匹配的相似度来识别被测物体。Fukunaga 等人提出空间模式分量分析方法<sup>[5]</sup>,以吸收光谱作为指纹谱来辨识物质。该方法要求高稳定的 THz 系统建立不同物质的指纹谱数据库。另外有些研究采用模式分类方法,应用马氏距离分类法和支持向量机的分类方法。这类监督学习方法,需要大量的学习样本进行分类器学习,而且模型的训练会影响分类效果。近年来,THz 信号的几何代数分析研究<sup>[6]</sup>则将 THz 信号描述为几何代数结构下的实矢量形式,通过矢量分解来实现 THz 信号分类。对高维 THz 信号矢量进行基于一维特征子空间(单矢量)或二维特征子空间(双矢量)的射影分解或共形分解方法,将高维信号矢量映射为特征空间中的低维矢量,从而在低维空间中实现信号的辨识和分类等应用。

传统的信号表示如离散余弦变换和小波变换等,

它们是基于“正交基”的展开,属于非冗余的正交变换。近年来研究人员在改变传统信号表示方面取得了很大的进展。超完备信号稀疏表示方法始于 1993 年, Mallat 等人首次提出了应用超完备冗余字典对信号进行稀疏分解的思想,并引入了匹配追踪(marching pursuit, MP)算法<sup>[7]</sup>。1998 年 Donoho 等人提出了基追踪(basis pursuit, BP)算法<sup>[8]</sup>,稍后在 2001 年又给出了基于 BP 算法的稀疏表示具有惟一解的边界条件,提出了字典的互不相干性概念。2006 年, Candés 和 Donoho 等人从信号分解和逼近理论基础正式提出了压缩感知的概念<sup>[9,10]</sup>。压缩感知的先验条件是信号必须在某种变换下可以稀疏表示。它要求信号经投影变换后,绝大部分变换系数的绝对值很小,所得到的变换向量是稀疏或近似稀疏的。信号稀疏表示的变换基可以根据信号本身的特点选取,常用的有离散余弦变换基、快速傅立叶变换基、离散小波变换基、Curvelets 基、Gabor 基以及冗余字典等<sup>[11]</sup>。压缩感知理论中的核心问题是信号重构算法,即由  $m$  次测量向量重构长度为  $n$  ( $m \ll n$ ) 的稀疏信号的过程。研究人员提出了一系列求得次最优解的算法,主要包括最小  $l_1$  范数法、匹配追踪系列算法、迭代阈值法等<sup>[10]</sup>。压缩感知理论带来了信号采样理论的变革,已被应用于压缩成像、图像识别、模式分类等领域。并且,在 THz 研究中,已经有基于压缩感知的 THz 单像素成像系统和基于压缩感知的 THz 图像重构等研究成果。

THz 信号几何代数分析的研究表明:在几何代数结构下,THz 信号可以描述为实矢量形式,而同类 THz 信号矢量位于与其物质相对应的二维特征子空间中,并且在所研究的目标物质是相同厚度的或非涅耳损耗可以忽略的情况下,同类 THz 信号矢量间具有线性相关特性。本文从信号稀疏表示方法出发,基于已知 THz 信号矢量构建冗余字典,并将 THz 信号矢量表示为多矢量线性组合的形式,通过构建线性方程组模型来研究 THz 信号的分类问题。在此基础上,本文提出基于稀疏表示的 THz 信号分类方法。该方法首先基于已知类别的 THz 信号矢量构建冗余字典,将待分类信号矢量表示为冗余字典中原子的线性组合,然后运用压缩感知理论中  $l_1$  最小化问题的求解方法来得到线性方程组的系数,最后根据稀疏系数来确定该信号矢量的类别标识。

本文第二节介绍了 THz-TDS 几何代数分析模型;第三节研究了 THz 信号的稀疏表示方法和信号分类模

型,提出了基于稀疏表示的信号分类方法,并详细讨论了该方法各步骤;第四节给出THz信号分类实验,验证了本文方法的有效性。

## 2 THz信号矢量

当利用脉冲THz技术测量物质的光谱时,首先获得通过自由空间(空气)的THz脉冲得时域波形,称为参考波形。然后,测量透射经过样品之后的时域波形,称为信号波形。分别对参考波形和信号波形进行傅里叶变换,得到参考光谱 $\hat{E}_{\text{ref}}(\omega_i)$ 和信号光谱 $\hat{E}_{\text{sam}}(\omega_i)$ (角频率 $\omega_i, i=1, 2, \dots, m$ )。在THz-TDS透射系统中,由信号与参考光谱相比较可得到物质的频域复传递函数 $\hat{h}_i = \hat{E}_{\text{sam}}(\omega_i) / \hat{E}_{\text{ref}}(\omega_i)$ 。

令空间 $\mathcal{V}_{2m}$ 为 $2m$ -维实矢量空间,它的正交基为 $\mathbf{e}_i, i=1, 2, \dots, 2m$ ,其对应的几何代数为 $\mathcal{G}_{2m}^{m,m,0}$ ,该频率相关的复传递函数可以表示为空间 $\mathcal{V}_{2m}$ 中的实矢量: $\mathbf{h} = \sum_{i=1}^m \hat{h}_i \mathbf{e}_i + \sum_{i=1}^m \ln |\hat{h}_i| \mathbf{e}_{m+i}$ 。

基于THz-TDS信号的物理机理和几何代数理论,本文作者们研究分析了THz-TDS信号矢量的几何分布和代数结构特性<sup>[6]</sup>。该研究表明:

(1) 在所研究的目标物质是相同厚度的情况下,或者是菲涅耳损耗可以忽略的情况下,由相同复折射率的目标物质所获得的THz-TDS信号所确定的信号矢量间具有线性相关的特性。在这类情况下,THz-TDS信号矢量具有射影等价特性。

(2) 每种物质的复折射率唯一确定空间 $\mathcal{V}_n$ 中的一个“二级面片”,该面片定义了空间 $\mathcal{V}_n$ 中一个二维子空间 $\mathcal{V}_2$ ,THz-TDS信号矢量位于与样品物质相对应的二级面片(2-blade)上。其中,物质所对应的面片可以采用通过同物质但不同厚度的两个THz-TDS信号矢量来计算得到。

令函数 $\text{id}(\mathbf{h})$ 为信号矢量 $\mathbf{h}$ 的类别标识,参数 $\lambda_i$ 为实数,则以上研究结果可简述为:

(1) 在样品均是相同厚度的或非涅耳损耗可忽略的情况下,同类THz信号矢量间具有线性相关的特性,即 $\text{id}(\mathbf{h}_1) = \text{id}(\mathbf{h}_2) \Leftrightarrow \mathbf{h}_1 = \lambda_1 \mathbf{h}_2$ ;

(2) 同类THz信号矢量位于唯一的二维子空间,该子空间可以由同质但不同厚度的两个THz-TDS信号矢量来描述,即

$$\text{id}(\mathbf{h}_1) = \text{id}(\mathbf{h}_2) = \text{id}(\mathbf{h}) \Leftrightarrow \mathbf{h} = \lambda_1 \mathbf{h}_1 + \lambda_2 \mathbf{h}_2,$$

其中 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 为同类不同厚度的信号矢量。

因此,THz信号矢量可以用有限个同类信号矢量的线性组合表示。

一般的,高维的THz信号矢量可以表示为单个或多个信号矢量的线性组合形式,即

$$\mathbf{h} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \dots + \lambda_n \mathbf{v}_n.$$

该线性表达式仅在对应 $\text{id}(\mathbf{h}) = \text{id}(\mathbf{v}_i)$ 项的参数 $\lambda_i$ 为非零系数,它是THz信号矢量的一种稀疏表示方法。从信号稀疏表示的角度来看,以已知THz信号矢量来构建冗余字典,上式则是THz信号矢量基于冗余字典的一种线性稀疏表示方法。

## 3 基于稀疏表示的分类方法

THz信号分类中的一个基本问题是基于已知物质的 $n$ 条THz信号来识别新信号所源自的物质成分。假设THz信号矢量空间为 $\mathcal{V}_m$ ,每条THz信号矢量有 $m$ 个采样点<sup>1</sup>,它可表示为列向量形式 $\mathbf{v} = [v_1, v_2, \dots, v_m] \in \mathbb{R}^m$ ;在已知类别的信号中,有 $n_i$ 条THz信号是源自第 $i$ 类物质样品的,则每条信号表示为列向量 $\mathbf{v}_i = [v_{i,j}]$ , $j=1, 2, \dots, m$ ,而这 $n_i$ 条信号可以表示为矩阵形式 $\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \dots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ 。基于 $k$ 类已知物质的 $n$ 条THz信号,冗余字典可被定义为矩阵 $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k] = [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \dots, \mathbf{v}_{k,n_k}] \in \mathbb{R}^{m \times n}$ 。

### 3.1 THz信号的稀疏表示

给定源自第 $i$ 类物质的新信号 $\mathbf{y}$ ( $\text{id}(\mathbf{y}) = i$ ,并且 $\mathbf{y} \in \mathbb{R}^m$ ),及冗余字典中该物质所对应的原子 $\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \dots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ ,则新信号样本 $\mathbf{y}$ 可近似表示为这些原子们的线性展开形式:

$$\mathbf{y} = \alpha_{i,1} \mathbf{v}_{i,1} + \alpha_{i,2} \mathbf{v}_{i,2} + \dots + \alpha_{i,n_i} \mathbf{v}_{i,n_i},$$

其中, $\alpha_{i,j} \in \mathbb{R}^{n_i}, j=1, 2, \dots, n_i$ 。

由于新信号所源自的物质类别标识 $i$ 是未知的。因此,待分类信号 $\mathbf{y}$ 可以表示为冗余字典中所有原子的线性展开形式:

$$\mathbf{y} = \mathbf{A} \mathbf{x}_0 \in \mathbb{R}^m \quad (1)$$

其中, $\mathbf{x}_0 = [0, \dots, 0, \alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,n_i}, 0, \dots, 0]^T \in \mathbb{R}^n$ 。

这样,THz信号分类问题就转化为线性等式求解问题。在高维信号空间,该等式的求解问题为最近邻问题(nearest neighbor, NN)或是最近子空间问题(nearest subspace, NS)。如果线性方程组未知数的数目小

<sup>1</sup>在本文工作中,仅使用复光谱的幅度信息,即 $v_j = \ln |\hat{h}_j|, j=1, 2, \dots, m$ 。

于等于方程的数目,即  $n \leq m$ ,则这个方程组仅有唯一解。而如果线性方程组欠定,即未知数的数目超过方程的数目,一般而言它有无数个解。但是如果这个欠定方程组只有唯一的一个稀疏解,那么可以利用压缩感知理论和方法来寻找这个解。

在 THz 信号分类应用中,相对于 THz 信号的有效频率采样点的数目  $m$ ,已知类别数目  $k$ (如 RIKEN 数据库)或冗余字典原子的数目  $n$ (如下文实验)都会出现大于  $m$  的情况,即此时等式(1)是欠定线性方程组。考虑到,在等式(1)的系数向量  $\mathbf{x}_0$  中,与待分类信号  $\mathbf{y}$  异类的原子们所对应的系数为零,因此系数向量  $\mathbf{x}_0$  是稀疏的。等式(1)的求解可以采用基于  $l_1$  最小的稀疏求解方法。

如果线性等式的解向量  $\mathbf{x}_0$  足够稀疏,则该等式的求解可转化为  $l_1$  最小化问题<sup>[12]</sup>:

$$(l_1): \hat{\mathbf{x}}_1 = \arg \min (\|\mathbf{x}\|_1), \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x} \quad (2)$$

这是基追踪(Basis Pursuit, BP)问题。

### 3.2 THz 信号分类方法

基于 THz 信号的稀疏表示方法,本文提出的 THz 信号分类方法(the sparse representation classification, SRC)框架如下:

- ① 输入:待分类信号向量  $\mathbf{y}$ ,已知信号向量的冗余字典:  $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k] = [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \dots, \mathbf{v}_{k,n_k}] \in \mathbb{R}^{m \times n}$ ,并对  $\mathbf{A}$  进行标准化处理(单位化矩阵中的列向量);
- ② 求  $l_1$  最小化问题的解  $\hat{\mathbf{x}}_1 = [\alpha_{i,j}]$ ;
- ③ 输出:类别标识  $\text{id}(\mathbf{y})$ 。

下面,本文对 SRC 方法中各步骤展开讨论。

首先是冗余字典中原子的选取方法。原子可以简单地在已知类别的信号向量中随机选取。然而,基于 THz 信号几何代数分析的结果可以得到冗余字典原子的优化选取方法。由于相同复折射率的目标物质所获得的 THz-TDS 信号所确定的信号向量间具有线性相关的特性,所以当对应某类物质仅选择一条信号向量时,应当选取该类中与其他向量间线性相关度最大的信号向量作为原子,即该向量与同类信号向量间内积之和最大。类似的,由于源自同质的 THz 信号向量均位于与样品物质相对应的二级面片(2-blade)上(即二维特征子空间中),而相对于线性相关度高的向量们而言,相关度低的向量能更好地描述二维子空间。因此当某类物质选择两条以上向量时,应当至少选取两条相关度最低的同类信号向量作为原子,其余原子则随机选取。基于以上讨论,本文提出冗余字典原子的优化选

取方法如下:给定第  $i$  类物质的已知信号向量集  $[\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \dots]$ ,则冗余字典中对应第  $i$  类物质的原子们为

$$\mathbf{A}_i = [\mathbf{v}_{i,j_1}, \mathbf{v}_{i,j_2}, \dots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$$

其中,如果  $n_i = 1, j_1 = \arg \max_j |\mathbf{v}_{i,j} \cdot \mathbf{v}_{i,l}| / (|\mathbf{v}_{i,j}| |\mathbf{v}_{i,l}|)$ ;

如果  $n_i \geq 2$ ,

$$(j_1, j_2) = \arg \min_{(j_1, j_2)} |\mathbf{v}_{i,j_1} \wedge \mathbf{v}_{i,j_2}| / (|\mathbf{v}_{i,j_1}| |\mathbf{v}_{i,j_2}|),$$

并且  $j_l = \text{randi}(n_i), l = 3, 4, \dots, n_i$ ,函数  $\text{randi}(z)$  为不大于  $z$  的随机整数。

当 THz 系统信噪比大,THz 信号稳定的情况下,THz 信号的稀疏表示为等式(1),其  $l_1$  最小化的求解方法为等式(2)BP 方法。然而,由于各次获取 THz 信号时,采样的时间间隔不同或时间窗不同,以及实验环境存在的空气湿度温度等因素的影响,需要考虑噪声因素。因此,等式(1)可修正为

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{e} \quad (3)$$

其中,噪声向量  $\mathbf{e} \in \mathbb{R}^m$ ,且噪声向量为能量有限的,即  $\|\mathbf{e}\|_2 < \varepsilon$ 。则该等式的解可转化为基追踪去噪(Basis Pursuit denoise, BPd)问题求解<sup>[12]</sup>:

$$(l_{1s}): \hat{\mathbf{x}}_1 = \arg \min (\|\mathbf{x}\|_1), \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon \quad (4)$$

在本文实验中,具体地采用 SPGL1 算法<sup>[12]</sup>实现 BP 和 BPd 方法。

最后是类别标识的确定方法。SRC 算法是基于  $l_1$  最小化问题的解来确定类别标识的。因此,可以将最大系数或最小残差作为分类依据,即

$$\text{id}(\mathbf{y}) = \arg \max_i |\alpha_{i,j}| \text{ 或 } \text{id}(\mathbf{y}) = \arg \min_i \|y - \hat{y}_i\|_2,$$

其中  $\hat{y}_i = \sum_{j=1}^{n_i} \alpha_{i,j} \mathbf{v}_{i,j}, i = 1, 2, \dots, k$ 。

在下文实验中简称:SRCmdBP 方法和 SRCmdBPd 方法为基于随机冗余字典,分别采取 BP 和 BPd 求解  $l_1$  最小化问题的 SRC 方法;而 SRCoptBP 方法和 SRCoptBPd 方法为基于优化冗余字典,分别采取 BP 和 BPd 求解  $l_1$  最小化问题的 SRC 方法。

## 4 实验结果

本文实验采用 Ekspla 公司 THz-TDS 透射系统,该系统的频域信噪比<sup>[13]</sup>为 85.4。实验中 BPd 算法中参数  $\varepsilon$  均取值 0.05。

### 4.1 THz 信号分类实验

实验采用六类不同物质的样品,共采集了 67 组 THz 信号,获得 67 个信号向量。其中,每个信号向量的有效频率域选取 0.2 ~ 1THz,频率采样点数目  $m$  为

28. 分别以  $n_i=1, 2, \dots, 5$  多矢量 ( $i=1, 2, \dots, k$ ) 构建冗余字典, 本文 SRC 算法的分类成功率见表 1。

表 1 THz 信号的分类成功率(%)

Tab.1 Success rates of THz signal classification (%)

以最大系数作为分类依据					
$n_i$	1	2	3	4	5
SRCOptBP	96.72	100	97.96	100	97.30
SRCOptBPd	96.72	100	97.96	100	100
SRCmdBP	94.79	92.33	94.86	95.49	97.03
SRCmdBPd	93.87	95.02	96.16	97.30	97.51
以最小残差作为分类依据					
$n_i$	1	2	3	4	5
SRCOptBP	96.72	100	100	100	100
SRCOptBPd	96.72	100	100	100	100
SRCmdBP	95.31	95.85	97.02	97.21	97.51
SRCmdBPd	94.36	97.20	96.78	98.00	97.84

表 1 中各 SRC 方法的成功率均大于 92%, 这表明本文 SRC 算法是可行的。并且, 采用本文优化冗余字典的 SRCOptBP 方法和 SRCOptBPd 方法, 它们的成功率最小值为 96.72%, 且均高于对应的随机冗余字典方法。这表明本文提出的冗余字典原子的优化选取方法是有效的。相对于以最大系数作为分类依据而言, 以最小残差作为分类依据的本文方法的成功率均出现了提高。

在以上实验中  $m=28$ , 仅当  $n_i=5, n=k \times n_i=30$  时, 信号分类方程组为欠定的 ( $m < n$ ), 它的解为优化稀疏解。而  $n_i$  为其他取值时, 信号分类方程组为过定的, 仅有唯一解。考虑到在实际应用中, 特别是 THz 图像信号的分类处理中, 为减少计算量往往仅选取少量的频率采样值进行处理。因此, 我们对信号进行等间隔的下采样, 有效频率域仍为 0.2 ~ 1THz。图 1 给出了不同频率采样点数目下 SRCOptBP 方法和 SRCOptBPd 方法(以最小残差作为分类依据)的分类成功率。

由图 1 可见, 当信号的频率点数目  $m \geq 10$  并且冗余字典原子数目(基的数目)  $n \geq 12$  ( $n_i \geq 2$ ) 时, 本文方法的成功率在 90% 以上。当  $m=5, 6, 7$  时, 本文方法的成功率不稳定。这是由于随着所使用的频率采样点数目减少, 用于分类处理的信号中所包含的信息也随之减少, 仅依据几个非特定的采样点数据是无法对 THz 信号进行分类的。而当  $n=6$  ( $n_i=1$ ) 时, 仅在  $m=28$  和

$m=10$  处, 本文方法的成功率在 90% 以上。这表明在信号频率点数目少的情况下需要采用多个同类矢量来构建冗余字典, 而不宜仅使用单矢量。

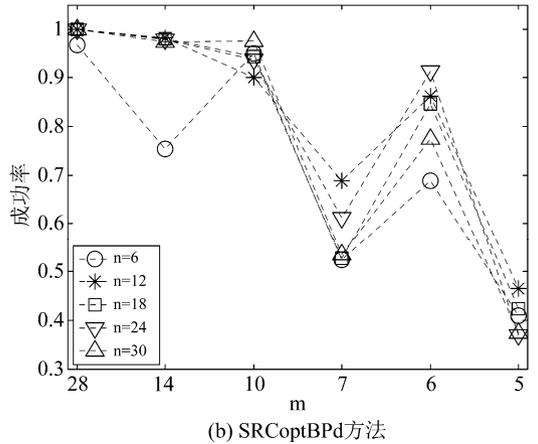
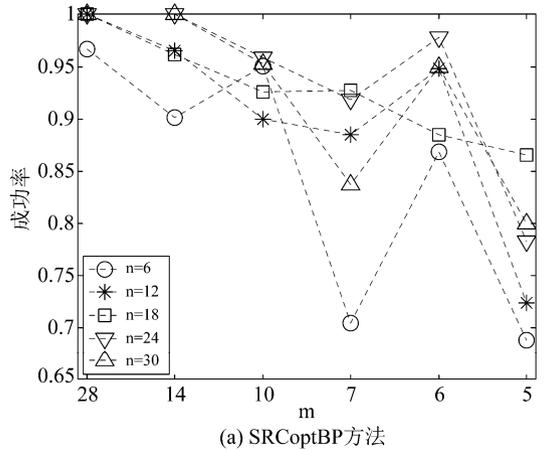


图 1 不同频率采样点数目(m)的信号分类成功率

Fig.1 Success rates of the signal classification with different number (m) in the frequency sampling

### 4.2 不同分辨率的 THz 信号分类实验

实验使用源自 23 类不同物质的 240 组 THz-TDS 信号, 其中各组信号之间存在时域采样间隔不统一和时间窗不统一的问题。通过在频域内对各信号矢量进行统一频率间隔的线性插值, 从而获得具有相同频率分辨率的 THz 信号矢量集  $S_0$ 。其中, 每个信号矢量统一选取的频率域为 0.2 ~ 1THz, 频率采样点数目  $m$  为 28。在信号集  $S_0$  中, 由于各信号的原频率分辨率不同, 通过人工频域插值而引入了误差。这导致同类信号间差异较实验 4.1 中的明显。实验分别以  $n_i=1, 2, \dots, 5$  的多矢量 ( $i=1, 2, \dots, k$ ) 构建冗余字典, 采用本文 SRC 算法的分类成功率见表 2。

在本实验中, 仅当  $n_i=1$  时,  $m > n$ ; 其余  $n_i$  取值时,

信号分类方程组的解为优化稀疏解。由实验结果表2可见,类似于上节实验,采用本文优化冗余字典的SRCoptBP方法和SRCoptBPd方法,它们的成功率均高于对应的随机冗余字典方法的。这同样验证了本文冗余字典的优化组建方法是有效的。并且,在 $n_i \geq 2$ 时,优化冗余字典的分类成功率均高于98%,而随机冗余字典的分类成功率均高于93%。它们均明显大于单矢量冗余字典( $n_i = 1$ )的成功率。这表明当信号间存在的误差较大时,单矢量不足以表示某类物质的信号矢量,此时应该采用多矢量来构建冗余字典。

表2 THz信号集 $S_0$ 的分类成功率(%)

Tab.2 Success rates(%) of THz signal classification for the signal set  $S_0$

以最大系数作为分类依据					
$n_i$	1	2	3	4	5
SRCoptBP	89.86	99.48	99.43	100	100
SRCoptBPd	88.48	99.48	99.43	100	99.26
SRCmdBP	85.37	93.87	97.73	97.26	98.46
SRCmdBPd	83.29	94.66	96.01	98.06	98.60
以最小残差作为分类依据					
$n_i$	1	2	3	4	5
SRCoptBP	89.86	98.97	100	100	100
SRCoptBPd	90.32	100	98.28	99.35	100
SRCmdBP	84.70	93.76	97.18	97.16	98.20
SRCmdBPd	82.19	93.61	95.17	97.06	97.87

然而,相对于以最大系数作为分类依据而言,以最小残差作为分类依据的分类方法的成功率并没有出现上文实验中明显地提高,它们的成功率仅相当于以最大系数为依据的分类方法结果。这是由于当同类信号间差异明显的情况下,同类信号矢量间的欧氏距离并不能很好地度量矢量间的相似性<sup>[14]</sup>。而残差的定义是2阶范数,在冗余字典的 $n_i = 1$ 的情况下,它为待分类信号矢量和某类样本信号矢量之间的欧氏距离;而在 $n_i \geq 2$ 时,它为待分类信号矢量和某类样本信号矢量的线性组合之间的欧氏距离。因此,在同类信号之间差异较为明显时,以最小残差为分类依据的分类方法的成功率相对于最大系数分类方法没有明显的优势。

类似地,我们还进行了多次仿真噪声实验。采用 $S_0$ 集中的信号矢量构建冗余字典,而待分类信号集则为该信号矢量集中的矢量们分别加上高斯噪声后得到的仿

真矢量集 $S_1$ 。在仿真实验中,高斯噪声均为0均值,标准差为原信号幅度均值的 $\lambda_j$ 倍,即 $\sigma_j = \lambda_j \text{mean}(\mathbf{v}_i)$ ,且 $\lambda_j \in [0.001, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04]$ ,对应信噪比值分别为61.4, 47.5, 41.4, 38.0, 35.4, 33.5, 31.9, 30.6, 29.9。采用优化冗余字典的本文方法对仿真矢量集 $S_1$ 进行分类,在不同噪声情况下分类成功率的平均值见表3。SRCoptBP方法和SRCoptBPd方法(以最大系数作为分类依据)在不同噪声情况下的分类结果见图2。

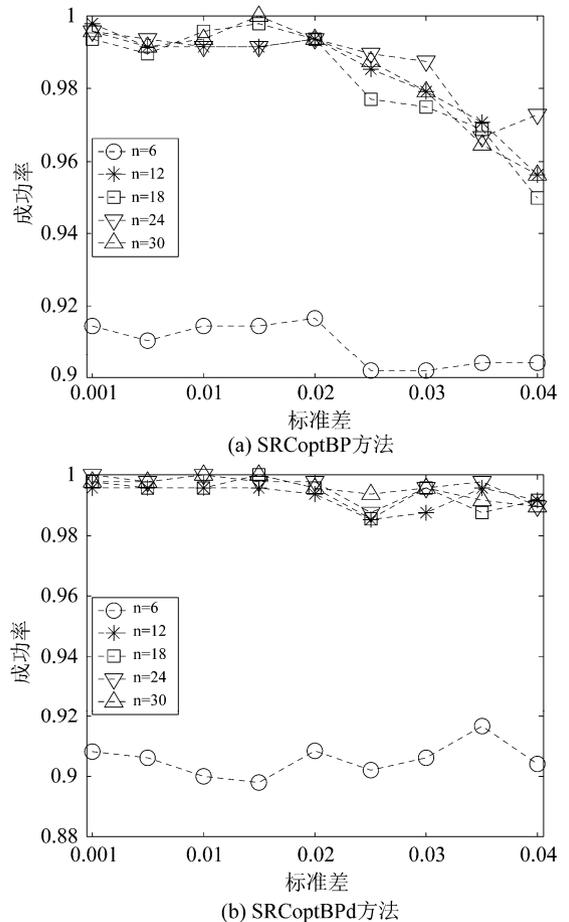


图2 仿真信号集 $S_1$ 的分类成功率

Fig.2 Averages of the classification success rates for the simulated signal set  $S_1$

在表3中,以最大系数或最小残差作为分类依据,它们的分类成功率平均值非常地接近。这同样表明:以最小残差作为分类依据,是基于欧氏距离准则;在同类信号之间差异明显时,分类方法不会明显改善。由表3数据和图2可见,在不同的 $n$ 取值中,当 $n = 6$ ( $n_i = 1$ )时,即以单矢量构建冗余字典,分类成功率在各噪

声环境下是最低的,而当  $n$  取其他值时, SRCoptBP 方法和 SRCoptBPd 方法的成功率最小值分别为 95% 和 98.5%。这是由于在待分类信号中引入了噪声,噪声对单一矢量的影响较大。因此,在噪声环境下不宜采用单矢量构建冗余字典。同时,由图 2 可见,以最大系数作为分类依据, SRCoptBPd 方法比 SRCoptBP 方法更为稳定。这是由于 SRCoptBPd 方法是基于噪声修正模型(3)的,考虑了噪声因素。所以,在噪声环境中 SRCoptBPd 方法相对稳定。

表 3 仿真信号集  $S_1$  分类成功率的平均值(%)

Tab. 3 Averages of the classification success rates(%) for the simulated signal set  $S_1$

以最大系数作为分类依据					
$n_i$	1	2	3	4	5
SRCoptBP	90.93	98.43	98.24	98.70	98.47
SRCoptBPd	90.56	99.30	99.40	99.61	99.58
以最小残差作为分类依据					
$n_i$	1	2	3	4	5
SRCoptBP	91.09	98.40	98.59	98.98	98.75
SRCoptBPd	90.95	99.17	99.00	98.87	98.96

## 5 结论

通过 THz 脉冲,人们可获得被测量物质的吸收率、折射率以及复折射率等物理参数等,从而辨识样品的物理化学成分及结构等。基于 THz 信号的物质辨识和分类是 THz 重要应用之一。

基于 THz 信号几何代数分析模型,本文从 THz 信号矢量的线性表示方法出发,将高维的 THz 信号矢量表示为多个信号矢量的线性组合形式,从而对 THz 信号矢量进行了稀疏表示。在此基础上,本文将 THz 信号分类问题描述为线性方程组的求稀疏解问题,提出了基于稀疏表示的 THz 信号分类方法。该方法首先采用已知类别的 THz 信号矢量作为原子来构建冗余字典,将待分类信号矢量表示为原子的线性组合,然后运用压缩感知理论中  $l_1$  最小化问题的求解方法来得到线性方程组的系数,最后根据稀疏系数来确定该信号矢量的类别标识。在本文方法中,基于 THz 信号矢量的几何分布和代数结构特性,提出了冗余字典原子的优化选取方法;通过引入噪声因素,对信号分类问题的线性方

程组模型进行了修正;在确定类别标识的问题上分别提出了以最大系数和最小残差作为分类依据的方法。

实验结果表明:本文提出的优化冗余字典组建方法的分类成功率高于随机冗余字典方法;在信噪比值高时,分类方法选择最小残差作为分类依据比以最大系数作为分类依据更有效;在信噪比值低时,应对每类物质至少采用两个已知的信号矢量来组建冗余字典(即对应每类应有至少两个原子),不宜采用单个矢量。实验结果验证了本文方法在等间隔采样的 THz 信号分类问题中的可行性和有效性。在压缩域进行 THz 信号分类是本文的后续研究的方向之一。

## 参考文献

- [1] 谢维信,裴继红,THz 信号处理与分析的研究现状和发展展望[J]. 电子学报,2007,35(10):1973-1979.  
W. X. Xie, J. H. Pei, Review of Terahertz signal processing and analysis [J]. Chinese Journal of Electronics, ACTA Electronica sinica, 2007, 35(10), 1973-1978. (in Chinese)
- [2] Roberto K. H. Galvão, Sillas Hadjiloucas, John W. Bowen, Clarimar J. Coelho, Optimal discrimination and classification of THz spectra in the wavelet domain[J]. OPTICS EXPRESS, OSA, 2003, 11(12):1462-1472.
- [3] Pei J. H. Ye P. L. and Xie W. X., Optimal wavelet analysis for THz-TDS pulse signals [C] // Photonics and Optoelectronics Meetings. Wuhan, China:POEM, 2008.
- [4] Sillas Hadjiloucas, Roberto K. H. Galvão, John W. Bowen, Analysis of spectroscopic measurements of leaf water content at terahertz frequencies using linear transforms [J]. Journal of the Optical Society of America A, 2002, 19(12):2495-2509.
- [5] Fukunaga Kaori, Ogawa Yuichi, Hayashi Shinichiro, and Hosako Iwao, Application of terahertz spectroscopy for character recognition in a medieval manuscript [J]. IEICE Electronics Express, 2008, 5(7):223-228.
- [6] Xie W. X. Li J. Pei J. H. THz-TDS Signal Analysis and Substance Identification via the Conformal Split [J]. SCIENCE CHINA, Information Sciences, 2012, 55(1):49-63.
- [7] Mallat S. G. and Zhang Z. Matching pursuits with time-frequency dictionaries [J]. IEEE Transactions on Signal Processing. 1993, 41:3397-3415.

- [8] Chen S B, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit [J]. *SIAM Journal on Scientific Computing*, 1998, 20(1):33-61.
- [9] Donoho D. L. Compressed sensing [J]. *IEEE Transactions on Information Theory*. 2006, 52(4): 1289-1306.
- [10] Candés E. J., Romberg J. and Tao T., Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information [J]. *IEEE Transactions on Information Theory*, 2006, 52(2):489-509.
- [11] Rauhut H, Schnass K V P. Compressed sensing and redundant dictionaries [J]. *IEEE Transactions on Information Theory*, 2008, 54(5):2210-2219.
- [12] E. van den Berg, Friedlander M. P., Probing the Pareto frontier for basis pursuit solutions [J]. *SIAM J. on Scientific Computing*, 31(2):890-912.
- [13] Naftaly M., Dudley R., Methodologies for determining the dynamic ranges and signal-to-noise ratios of terahertz time-domain spectrometers [J]. *Optics Letters*, 2009, 34(8):1213-1215.
- [14] 李静, 谢维信, 裴继红, THz-TDS 信号的相似性度量及其应用 [J]. *信号处理*, 2011, 27(10):1457-1463.  
Li J., Xie W. X., Pei J. H., Similarity metrics and their applications of THz-TDS signals [J]. *Signal Processing*, 2011, 27(10):1457-1463. (in Chinese)

### 作者简介



李静(1976-),女,生于湖南长沙,西安电子科技大学电子工程学院在读博士研究生,1998年毕业于中南大学信息工程学院获学士学位,2006年毕业于深圳大学信息工程学院并获硕士学位。主要研究方向为:信号分析和处理、图像分析与处理等。E-mail:lijingsz@szu.edu.cn



谢维信(1941-),男,生于广东花都,教授,博士生导师,1965年毕业于西安军事电讯工程学院,1981-1983年和1989-1990年分别在美国宾夕法尼亚大学访问学者和访问教授,现为深圳大学学术委员会主任,中国电子学会信号处理分会主任委员。研究方向为:信号处理、智能信息处理、雷达目标识别、太赫兹时域光谱(THz-TDS)信号和图像分析等。E-mail:wxxie@szu.edu.cn



裴继红(1966-),男,生于甘肃省武威市,1989年毕业于北京航空航天大学获学士学位,分别于1994年和1998年在西安电子科技大学毕业并获硕士学位和博士学位,现为深圳大学教授。主要研究方向为:智能信息处理、模式识别、视频图像分析、太赫兹时域光谱(THz-TDS)信号和图像分析。E-mail:jhpei@szu.edu.cn