

无机盐混合溶液的太赫兹光谱定量分析

何明霞^{1,2,3}, 孙珑玲^{1,2,3}, 陈 达³, 黄志轩³, 刘立媛^{2,4}, 赵晋武^{1,2,3}, 张洪桢^{1,2,3}

1. 天津大学精密测试技术及仪器国家重点实验室, 天津 300072
2. 天津大学太赫兹研究中心, 天津 300072
3. 天津大学精密仪器与光电子工程学院, 天津 300072
4. 天津大学光电信息技术教育部重点实验室, 天津 300072

摘要 太赫兹生物医学是当前光谱研究领域的前沿热点,其主要难点在于如何在有效避免水分干扰的同时,实现复杂生物体系组分的精准分析。太赫兹光谱产生于分子振动的信息,其吸收谱较弱,吸收峰严重重叠,且多组分复杂样品的太赫兹光谱往往不是各组分光谱的简单叠加,难以用传统的峰高、峰面积标定技术进行定量计算。但采用多元校正技术可以方便地实现太赫兹光谱的定量分析,使太赫兹光谱成为一种快速、简便且适用范围广泛的分析技术。以KCl和NaCl的无机盐混合体系为典型研究体系,两种组分的浓度范围均为 $0.1\sim 2\text{ mol}\cdot\text{L}^{-1}$,浓度间隔为 $0.1\text{ mol}\cdot\text{L}^{-1}$ 。获取20组浓度配比不同的混合溶液的吸收系数和折射率,巧妙利用水溶液体系中无机金属离子的水合氢键作用,由此采集无机盐溶液体系的太赫兹时域光谱,提取各组分的特征信息,建立多尺度数据驱动的定量分析模型,有望实现水溶液中无机金属离子的定量分析。针对太赫兹光谱数据规模大、基质干扰强及数据关联复杂等特点,构建复杂二维小波变换、多变量筛选、贝叶斯数据挖掘、深度学习和数据关联性分析技术为一体的算法数据库,由此构建基于多尺度数据驱动的太赫兹光谱解析方法。论文依据正交实验的原则,构建具备良好数据结构特征的混合溶液数据集,引导后续的光谱解析方法准确提取无机金属离子水合氢键信息。在此基础上,发展自适应算法,寻找光谱数据变量与浓度间的关系,并采用变量筛选技术,从原始光谱数据中提取无机盐水合氢键的特征信息,最终构建浓度与特征信息之间的数据驱动模型。计算结果表明,KCl和NaCl组分的预测误差分别为8.0%和9.1%,能有效满足大部分应用的检测精度要求。多尺度数据驱动模型方法充分利用太赫兹光谱信号的时域和频域多尺度特性,实现数据预处理与多元校正的一体化运算以避免重要信息丢失,具备高度自适应特征。因此,基于数据驱动建模的太赫兹光谱分析新方法为太赫兹生物医学研究提供了新思路。

关键词 混合溶液; 太赫兹时域光谱; 数据驱动建模; 定量分析

中图分类号: O433.4 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2019)12-3731-06

引言

近年来,太赫兹生物医学研究成为光谱研究领域的前沿热点。太赫兹(THz)波是指频率 $0.1\sim 10\text{ THz}$,相应波长为 $30\ \mu\text{m}\sim 3\text{ mm}$ 的电磁波,在电磁波谱中位于微波和红外辐射之间^[1]。由于生物体对太赫兹波具有独特的响应,生物分子的振动及转动能级多处于太赫兹波段,因此有望对生物分子的组成进行分析鉴定。然而由于生物体中含有大量的水分,极易导致太赫兹光谱受到严重的干扰。如何在水干扰的情况下,准确提取复杂生物体系中的组分信息,成为太赫兹研究

领域的一个热点、难点。

目前基于太赫兹光谱技术的混合物定量分析多集中于固体物质,鲜有针对混合水溶液的定量分析。水溶液体系与人们的生活息息相关,尤其在化工、农业、医学等领域。在多数应用环境中,溶液中包含的溶质往往不只有一种。为了严格控制生产的质量,混合溶液中各组分的含量检测已成为一项不可缺少的环节,也为太赫兹生物医学应用提供了良好的理论基础支撑。

太赫兹光谱技术在混合物定量检测中,主要应用于固体混合物组分定量分析。王迎新等^[2]基于太赫兹吸收光谱拟合的定量检测方法,针对无水茶碱和乳糖一水合物组成的固体

收稿日期: 2018-10-24, 修订日期: 2019-02-06

基金项目: 国家自然科学基金项目(61675151)资助

作者简介: 何明霞, 1965年生, 天津大学精密仪器与光电子工程学院教授 e-mail: hhmmxx@tju.edu.cn

混合物,采用经典最小二乘回归建立校正模型,对其组分浓度进行解析;陈涛等应用太赫兹时域光谱技术结合化学计量学方法对多组分药物混合物中药物活性成分和药用辅料的含量进行了定量分析研究。水在太赫兹频段具有独特的分子键振动模式(氢键拉伸和弯曲),因此其对太赫兹波具有强烈的吸收特性^[3],导致水溶液太赫兹定量分析方法的研究尚处于起步阶段。Ueno 等^[4]通过太赫兹时域光谱对氨基酸混合样品进行定量分析,计算的浓度误差低于 20%。大量的研究表明,水溶性金属离子通过影响水合氢键的作用,间接影响水分子的振动和转动光谱^[5],进而为太赫兹时域光谱定量分析提供了良好的基础。

基于太赫兹时域光谱技术,利用适合于混合溶液物系的多尺度数据驱动定量算法,对不同浓度配比的 NaCl 和 KCl 混合溶液进行物系组分的定量分析。多尺度数据驱动定量算法由多尺度建模方法^[6]发展而来,该算法通过分析水溶液的太赫兹时域光谱特征及内在规律,准确解析溶液中 Na⁺ 和 K⁺ 等离子浓度,有望推动太赫兹时域光谱技术在混合溶液定量分析中的应用,并为后续的生物医学研究提供有力支撑。

1 实验部分

1.1 装置

实验使用天津大学太赫兹研究中心实验室自行搭建的 8F 太赫兹时域光谱(Terahertz time-domain spectroscopy, THz-TDS)系统。该系统原理图如图 1 所示。

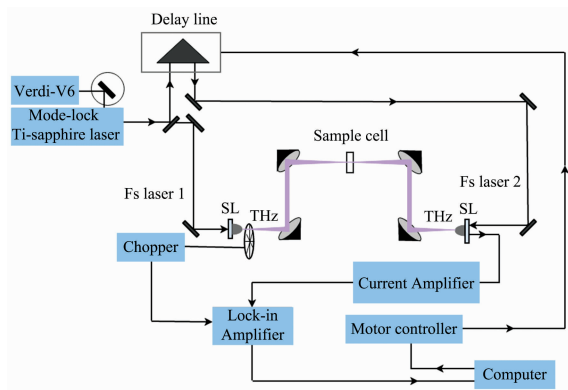


图 1 太赫兹时域光谱系统原理图

Fig. 1 Terahertz time domain spectroscopy system

钛宝石激光器 Mantis 产生重复频率为 100 MHz、中心波长为 800 nm、脉宽约为 20 fs 的超短脉冲序列^[7],经过分束器分为泵浦光及探测光两束光。由于来自同一束光波,故探测光和泵浦光具有很好的相干性。泵浦光光路中,斩波片的调制频率为 370 Hz,经过斩波片对光波的调制和衰减片对光波的衰减,泵浦光以 10 mW 左右的功率入射至发射天线上,产生太赫兹波。在太赫兹波对被测样品进行光谱信息探测的环节,光波进入由四个抛物面镜组成的 8F 系统,被测样品位于其中两个焦距相同的抛物面镜的共焦点位置处,透过待测样品的 THz 波携带样品信息后被太赫兹接收天线吸

收,转化成与太赫兹辐射场强大小成正比的电信号。

由于瞬态的电流信号强度较弱,故需要对其进行信号的滤波和放大,降低系统采集数据过程中的干扰和噪声信号,提高太赫兹时域光谱系统的信噪比,以获得有用的样品信息。在 THz-TDS 系统中,首先利用电流放大器对电流信号进行初步放大,而后利用锁相放大器实现对电流信号的滤波和再次放大,获取噪声干扰较少的有用信号。最终由计算机接收数据并进行数据的初步处理分析,以时域信号的形式进行储存,得到太赫兹时域光谱曲线。

实验所用的液体样品池为 Hellma Analytics 公司的可拆卸比色皿 106-QS,其内腔厚度为 0.1 mm,内宽为 9 mm,容量为 26 μ L,外形尺寸为 45 mm \times 12.5 mm \times 2.6 mm。实验中,太赫兹时域光谱系统的参考信号带宽约为 2.0 THz,环境温度保持在 20 $^{\circ}$ C 左右,相对湿度保持在 3% 以下。

1.2 样品制备

在化学和生物体系中,无机金属阳离子 Na⁺, K⁺, Ca²⁺ 和 Mg²⁺ 对分子识别、蛋白质和核酸的结构与功能起着十分重要的作用。对于摩尔浓度相同的各溶质,阳离子化合价越高,则溶液中相应的阴离子浓度越高,而对于无机盐混合溶液,阴离子浓度的大小将直接影响各组分在水中的溶解程度。实验选取阳离子化合价较低的 NaCl 和 KCl 为混合溶液的两种组分,以获得更大浓度范围的混合溶液。

实验选取分析纯纯度大于 99.5% 的氯化钠固体粉末和氯化钾固体粉末,按照正交实验的设计原则,利用精度为 0.01 mg 的分析天平和规格为 10 mL 的容量瓶配制相应浓度的溶液。经过试验,确定两种溶质混合时在水中的最大溶解浓度均为 2 mol \cdot L⁻¹ 左右。同时,由于水对太赫兹的吸收较强,若浓度间隔过小,则易导致两种溶质的光谱信息被水覆盖,因此实验中两组分的浓度间隔取 0.1 mol \cdot L⁻¹。

首先,分别配制浓度为 4 mol \cdot L⁻¹ 的 NaCl 溶液和 KCl 溶液;其次,按照相应比例混合两种溶液,并借助超纯水,依据正交实验的原则,配制 20 种不同浓度配比的混合溶液。最后,使得组分 NaCl 和 KCl 的浓度范围均为 0.1~2 mol \cdot L⁻¹,浓度间隔均为 0.1 mol \cdot L⁻¹。

1.3 方法

使用透射式 8F 太赫兹时域光谱系统测量 NaCl 和 KCl 混合溶液样品的太赫兹时域光谱。利用移液枪将被测溶液移至比色皿中,每个待测样品取 3 个样,每个样进行 1 次参考信号的测量和 3 次样品信号的测量,以排除取样误差和系统误差。故每个浓度配比的样品均有 9 组数据,取其平均值应用于后续算法分析。

2 结果与讨论

2.1 混合物系光谱实验数据采集

采集浓度为 2 mol \cdot L⁻¹ 的 NaCl 溶液、浓度为 2 mol \cdot L⁻¹ 的 KCl 溶液及超纯水的标准吸收谱图和折射率谱图,如图 2 所示。

通过对比图中三种不同体系的谱图曲线可以发现,对于吸收系数,相同浓度的 KCl 溶液和 NaCl 溶液有明显差异,

且明显大于水的吸收系数；对于折射率，三种体系存在差异，且在 1.3 THz 左右 KCl 溶液和 NaCl 溶液出现显著差异。这为进一步分析提供了理论依据。

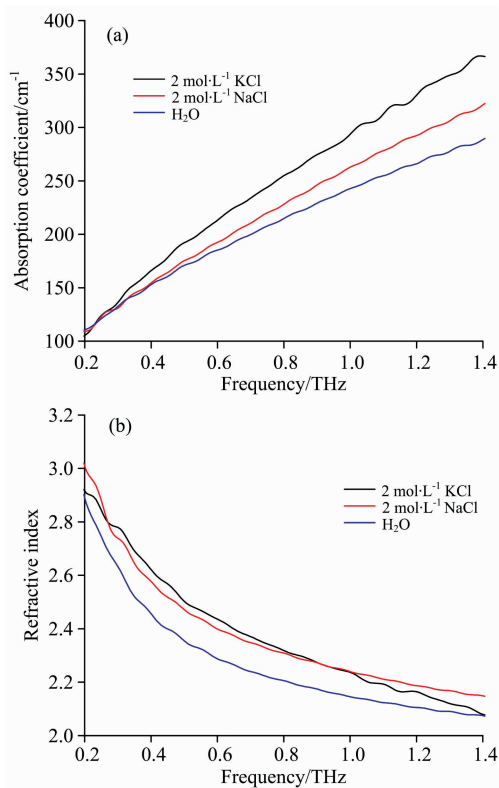


图 2 KCl 溶液、NaCl 溶液、超纯水的标准谱图

(a): $2 \text{ mol} \cdot \text{L}^{-1}$ KCl, $2 \text{ mol} \cdot \text{L}^{-1}$ NaCl 和 H_2O 的吸收系数;

(b): $2 \text{ mol} \cdot \text{L}^{-1}$ KCl, $2 \text{ mol} \cdot \text{L}^{-1}$ NaCl 和 H_2O 的折射率

Fig. 2 Terahertz spectroscopy results of each component

(a): Absorption coefficient of $2 \text{ mol} \cdot \text{L}^{-1}$ KCl, $2 \text{ mol} \cdot \text{L}^{-1}$ NaCl and water; (b): Refractive index of $2 \text{ mol} \cdot \text{L}^{-1}$ KCl, $2 \text{ mol} \cdot \text{L}^{-1}$ NaCl and water

选取频谱信号效果较好的 0.2~1.4 THz 频段的数据，以样品的吸收系数和折射率为研究对象，如图 3 所示。可知各浓度配比的混合溶液的吸收系数随频率的增加而增大，且各吸收系数曲线的曲率不同；各浓度配比的混合溶液的折射率随频率的增加而减小。

2.2 基于特征提取的数据驱动校正模型

2.2.1 模型建立

对于获取的太赫兹光谱数据，选取其中的一部分进行样品光谱的建模，另一部分进行模型的验证。对于建模所用的样品光谱数据，首先，利用背景噪声扣除、基线漂移校正、多尺度校正等算法进行光谱的预处理；其次，利用主成分分析空间偏除法、SIMCA 模型预测值法等判别异常光谱；最后，设置参数，利用 SNV 和 MSC 等背景和散射校正方法、多尺度数据驱动算法等进行多尺度建模。对于模型验证所用的样品光谱数据，将其带入前述多尺度建模算法中，利用多元统计过程控制 (MSPC) 技术、统计量判断标准法等动态评价与自动更新算法，进行样品光谱的验证。由此，对比预测

集与实际值的误差，获得该多尺度数据驱动定量分析模型的误差。

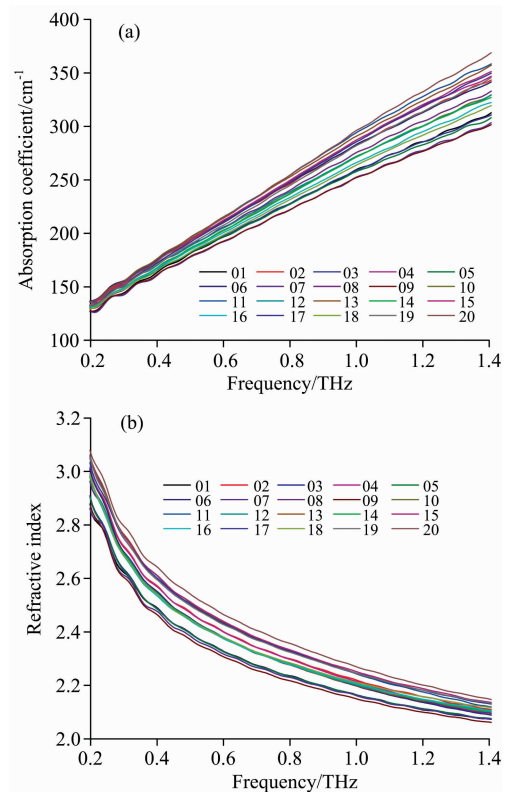


图 3 混合溶液太赫兹光谱实验数据结果

(a): NaCl 与 KCl 混合溶液吸收系数;

(b): NaCl 和 KCl 混合溶液折射率

Fig. 3 Terahertz spectroscopy results of 20 sets of the mixed solution

(a): Absorption coefficient of the mixed solution;

(b): Refractive index of the mixed solution

本文将 20 组样品采用 KS 分组算法，选择其中 12 组样品作为校正集进行数据驱动模型的建立，剩余的 8 组光谱验证集用于验证定量模型的预测结果。

2.2.2 多元定量校正算法分析过程

(1) 混合溶液总浓度定量校正模型建立及分析

Na^+ 和 K^+ 对水中氢键的作用，会造成盐溶液和纯水在吸收系数和折射率上的差异。从总体上看，三种体系的折射率都随频率的升高而下降，但随频率下降的速率不一致。因此，可以将折射率作为数据对象，建立 NaCl 和 KCl 混合溶液总浓度的定量校正模型，该模型校正集的拟合效果和验证集的预测效果如图 4 和表 1 所示。

其中，蓝色的线为标准线，表示理论预测浓度与实际实验浓度一致；红色的点为样本的预测效果点，其越接近蓝线表示预测效果越好。

其中，RMSEP 为均方根误差，其表征的是误差的绝对量，其值越小表示拟合或预测效果越好； R^2 为决定系数，其值越接近 1，表示预测结果越好。

由图 4 和表 1 可知，样本预测效果点均在标准线附近，

校正集和验证集的平均相对误差较小,分别为 5.1% 和 2.9%,故多尺度数据驱动定量分析模型拥有良好的预测效果,具有很高的可靠性。

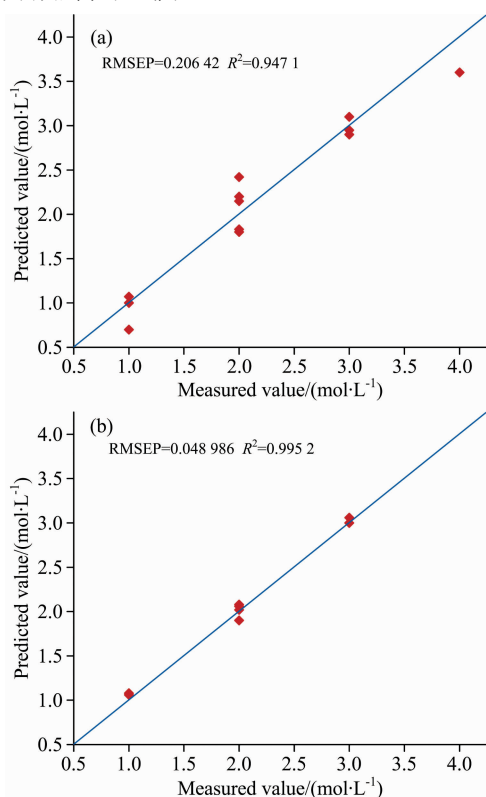


图 4 KCl 和 NaCl 混合溶液总浓度模型分析结果

(a): 总浓度校正集拟合结果; (b): 总浓度验证集预测结果

Fig. 4 Total concentration model analysis results of KCl and NaCl mixed solution

(a): Fitting result of calibration set;
(b): Fitting result of validation set

表 1 混合溶液总浓度模型预测效果表

Table 1 Prediction results of total concentration model

	RMSEP	R^2	平均相对误差/%
校正集	0.206	0.947	5.1
验证集	0.049	0.995	2.9

(2) 混合溶液各组分浓度定量校正模型建立及分析

吸收谱含有的 KCl 和 NaCl 的独立信息较多,因此,将吸收系数作为数据对象,建立混合溶液各组分浓度的定量校正模型。对吸收系数与浓度的相关性进行分析,发现吸收谱中含有 KCl 的特征信息,而 NaCl 的特征信息难以通过相关性分析直接获得,需要引入标准的水太赫兹谱作为参照提供吸收谱中包含的 NaCl 和 KCl 的波动信息。

为了进一步提取混合溶液太赫兹时域光谱中的 NaCl 和 KCl 浓度信息,论文采用数据驱动建模方法,准确提取各混合溶液实验样品与超纯水的吸收特征信息,并进行归一化,进而得到标准校正后的特征光谱信息。相关结果如图 5 所示。

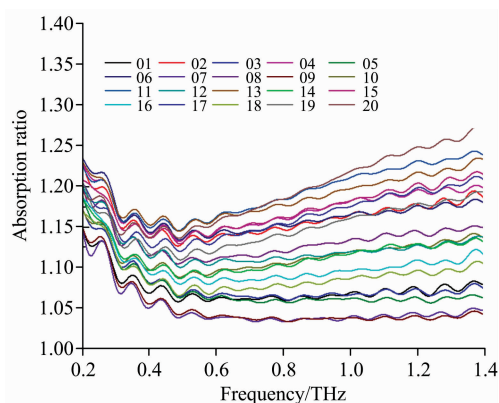
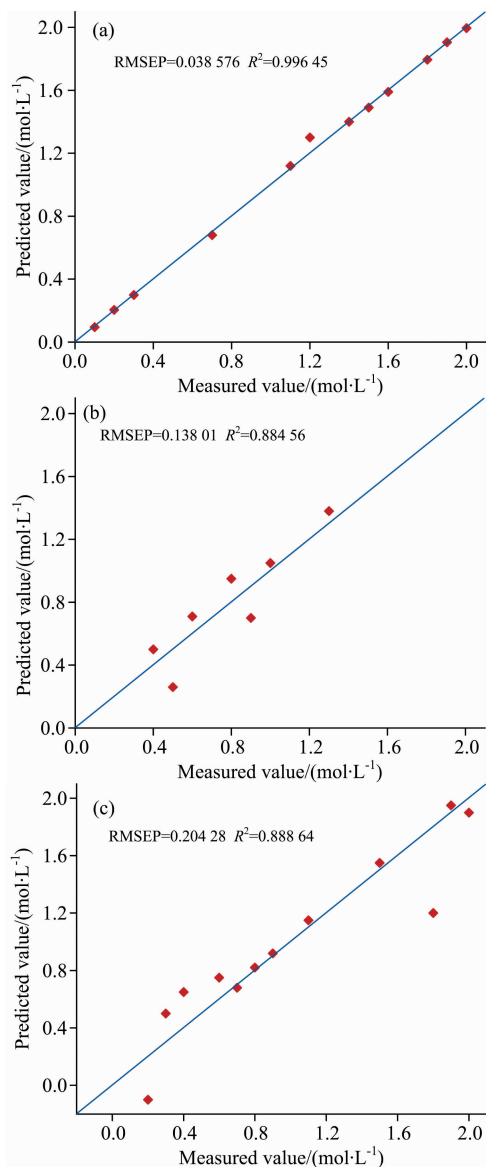


图 5 校正后的特征光谱信息

Fig. 5 Characteristic spectra after calibration

采用多尺度数据建模技术时,通过对比原始光谱和提取的光谱特征,可以发现 KCl 和 NaCl 组分相关的定量信息,其定量模型的预测模型效果如图 6 和表 2 所示。



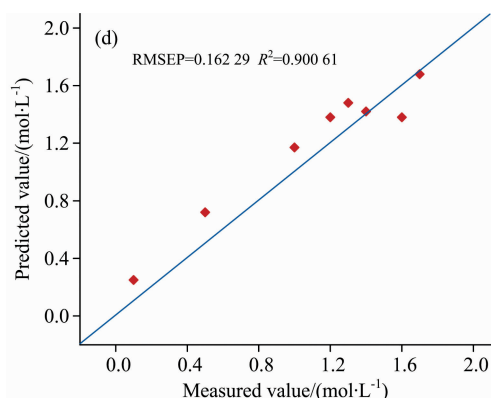


图 6 KCl 和 NaCl 混合溶液各组分浓度模型分析结果

(a): KCl 浓度校正集拟合结果; (b): KCl 浓度验证集预测结果;
(c): NaCl 浓度校正集拟合结果; (d): NaCl 浓度验证集预测结果

Fig. 6 Component concentration model analysis results of the mixed solution

(a): The fitting result of KCl's concentration of calibration samples;
(b): The predicted result of KCl's concentration of validation samples;
(c): The fitting result of NaCl's concentration of calibration samples;
(d): The predicted result of NaCl's concentration of validation samples

由图 6 和表 2 可知, 对于 KCl 成分, 样本预测效果点均在标准线附近, 校正集和验证集的平均相对误差较小, 分别为 2.5% 和 8.0%, 表明 K^+ 的水合作用较强, 诱导了较强的太赫兹时域光谱变化量; 对于 NaCl 成分, 样本预测效果点大致在标准线附近, 存在少数几个点偏离标准线较远, 校正集和验证集的平均相对误差分别为 11.2% 和 9.1%, 说明

Na^+ 的水合作用较弱, 导致太赫兹时域光谱变化量较小。此外, 由于太赫兹本身对水的吸收较强, 使得混合溶液体系的特征信号相对比较微弱, 故这两种组分在高浓度时的预测精度优于低浓度的预测精度。

表 2 KCl 和 NaCl 混合溶液各组分浓度预测效果表

Table 2 Prediction effect of the component concentration model

		RMSEP	R^2	平均相对误差/%
KCl	校正集	0.039	0.996	2.5
	验证集	0.138	0.885	8.0
NaCl	校正集	0.204	0.889	11.2
	验证集	0.162	0.901	9.1

3 结 论

以 KCl 和 NaCl 混合溶液为典型体系, 验证太赫兹时域光谱技术在溶液定量分析中的应用潜力, 为太赫兹生物医学研究提供了良好的理论支撑。计算结果表明, 太赫兹时域光谱技术可以精准捕获不同金属离子的水合氢键作用, 进而为无机盐体系的水溶液定量分析提供了良好的基础。然而, 由于太赫兹时域光谱对水的吸收较强, 高浓度无机盐的分析精度优于低浓度的无机盐组分。计算结果表明, 多尺度数据驱动模型在混合无机盐溶液的光谱信号分析中得到了良好的光谱解析和信息提取效果, 并具备优异的自适应性, 有望促进太赫兹时域光谱检测技术在生物医学领域的推广和应用。

References

- [1] HE Ming-xia, GUO Shuai(何明霞, 郭 帅). Journal of Electronic Measurement and Instrument(电子测量与仪器学报), 2012, 26(8): 663.
- [2] WANG Ying-xin, KANG Ke-jun, CHEN Zhi-qiang, et al(王迎新, 康克军, 陈志强, 等). Journal of Tsinghua University · Science and Technology(清华大学学报 · 自然科学版), 2009, 49(2): 161.
- [3] JIANG Qiang, WANG Yue, WEN Zhe, et al(蒋 强, 王 玥, 文 哲, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(4): 1049.
- [4] Ueno Y, Rungsawang R, Tomita I, et al. Analytical Chemistry, 2006, 78(15): 5424.
- [5] Stefan Funkner, Gudrun Niehues, Diedrich A, et al. Journal of the American Chemical Society, 2012, 134(2): 1030.
- [6] CHEN Da, LU Fan, LI Qi-feng(陈 达, 卢 帆, 李奇峰). Nanotechnology and Precision Engineering(纳米技术与精密工程), 2017, (2): 121.
- [7] LAI Hui-bin, HE Ming-xia, TIAN Tian, et al(赖慧彬, 何明霞, 田 甜, 等). Acta Optica Sinica(光学学报), 2018, 38(6): 0630001.

Quantitative Analysis of Mixed Inorganic Salt Solution Based on Terahertz Spectroscopy

HE Ming-xia^{1,2,3}, SUN Long-ling^{1,2,3}, CHEN Da³, HUANG Zhi-xuan³, LIU Li-yuan^{2,4}, ZHAO Jin-wu^{1,2,3}, ZHANG Hong-zhen^{1,2,3}

1. State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China

2. Center for Terahertz Waves, Tianjin University, Tianjin 300072, China

3. School of Precision Instrument and Optoelectronics Engineering, Tianjin University, Tianjin 300072, China

4. Key Laboratory of Opto-Electronics Information and Technical Science, Ministry of Education, Tianjin University, Tianjin 300072, China

Abstract Terahertz biomedicine is getting more and more concentration, especially in the field of spectroscopic research. Its main difficulty lies in how to achieve accurate component analysis of complex biological system as well as effectively avoid water interference. Terahertz spectrum contains the information of molecular vibration. However, its absorption spectrum is weak and overlaps seriously. Therefore, it is difficult to use traditional calibration techniques for quantitative calculation, such as the peak height and peak area. Adopting multivariate correction method makes terahertz spectrum a fast, simple and widely applicable way to carry out quantitative analysis. In this paper, the mixed aqueous salt solution of KCl and NaCl is taken as a typical system to be studied. The concentration of each component varies from 0.1 to 2 mol · L⁻¹, with an interval of 0.1 mol · L⁻¹. Due to the hydrated hydrogen bond of inorganic metal ions, terahertz time-domain spectrum data can be collected to analyze each component quantitatively. Based on the orthogonal experiment principle, we constructed the data set with outstanding structure characteristics to accurately extract the hydrogen bond information by spectral analysis. Here, the adaptive algorithm is developed to find the relationship between the spectral data and the concentration, and the variable screening technology is adopted to extract the characteristic information of different components from the original spectral data. Finally, we build the data-driven model between the concentration and the characteristic information. The calculation results show that the prediction errors of KCl and NaCl components are 8.0% and 9.1% respectively, which can effectively meet the requirements of detection accuracy for most applications. Therefore, the new method of data-driven modeling terahertz spectrum analysis can provide a new way for terahertz biomedical research.

Keywords Mixed solution; Terahertz time-domain spectrum; Data-driven modeling; Quantitative analysis

(Received Oct. 24, 2018; accepted Feb. 6, 2019)