文章编号：0258-7025(2009)Supplement 1-0283-03

应用支持向量机对抗生素药物太赫兹吸收谱的识别

赵树森 梁美彦 沈京玲
（首都师范大学物理系，北京 100048）

摘要 采用太赫兹时域光谱(THz-TDS)技术对 2 种常见消炎用粉针类药品和 4 种胶囊、片剂、冲剂类药品进行实验研究并得到它们在 0.2～2 THz 频率范围的特征吸收光谱，对这些光谱进行了分析比较，用支持向量机(SVM)对太赫兹吸收光谱进行了识别分类。首先，用归一化预处理后的 6 种药品的太赫兹吸收光谱训练 libsvm 模型；然后，选用与训练光谱不同时间测得的 6 种药品太赫兹吸收光谱作为检测光谱，经过归一化预处理后分别输入到训练好的 libsvm 模型中进行识别。研究表明，各种药物由于其化学成分的不同，显示出不同的太赫兹吸收光谱。用支持向量机可以实现对不同种类药品的识别。这些结果为太赫兹光谱技术用于药品的检测和识别提供了另一种有效的方法。

关键词 光谱学；药品识别；太赫兹吸收；光谱；支持向量机
中图分类号 O433.4 **文献标识码** A **doi**：10.3788/CJL200936s1.0283

Identification of Terahertz Absorption Spectra of Medicines Using Support Vector Machines

Zhao Shusen Liang Meiyan, Shen Jingling
(*Department of Physics, Capital Normal University, Beijing* 100048, *China*)

Abstract On the base of absorption spectra obtained in the range from 0.2 to 2 THz of six different medicines using terahertz time-domain spectroscopy (THz-TDS) technique, the THz absorption spectra of medicines were identified successfully by support vector machines (SVM). Firstly, absorption spectra of the six different medicines, which were pretreated by normalized unit, were used to train libsvm program. Secondly, absorption spectra of the medicines which were measured on different date, pretreated by normalized unit too, were identified by the libsvm and the identification rate of 100% was achieved. The results indicated that it is feasible to apply SVM on identification of medicines and providing an effective method in the secure inspection and identification for medicines.
Key words spectroscopy; medicines identification; terahertz absorption spectra; support vector machines

1 引 言

青霉素类和头孢菌素类抗生素是最具有代表的两类 β-内酰胺类抗生素，这两类抗生素具有广谱抗菌、灭菌能力强等特性，是临床常用的药品。但由于抗生素种类繁多，结构复杂，目前市场上的药品流通中来源广、渠道多，为了确保临床用药安全有效，抗生素的鉴别和质量保证是有效用药的关键。

青霉素和头孢菌素已报道的测定方法有分光光度法，HPLC 法，红外光谱鉴别法等[1~4]。太赫兹光谱以其具有的指纹特性，使光谱用于药物鉴别成为可能[5]。

支持向量机(SVM) 是 Vapnik 等[6]提出的一类新型机器学习方法。由于其出色的学习性能，该技术已成为机器学习界的研究热点，并在很多领域都得到了成功的应用。在红外光谱波段 SVM 应用广泛[7,8]，在太赫兹波段 Xiaoxia Yin 等[9]用 SVM 识别核酸的太赫兹吸收光谱，而国内尚没有报道 SVM 在太赫兹波段的类似应用。本文用支持向量机以 6 种抗生素的 THz 相对吸收光谱(以下称吸收光谱)作为训练数据，对 6 种抗生素不同时间的吸收光谱数据进行识别，识别率达到 100%，得到了预期的效果。

2 样品的 THz 吸收光谱

2.1 实验装置

基于 THz 时域光谱技术的透射式光谱实验装

基金项目：北京市教育委员会科技发展计划重点项目(KM200910028005)资助课题。
作者简介：赵树森(1983－)，男，硕士研究生，主要从事太赫兹光谱检测方面的研究。E-mail：winter-sen@163.com
导师简介：沈京玲(1957－)，女，教授，主要从事太赫兹光谱、非线性光学及混沌等方面的研究。
E-mail：sjl-phy@mail.cnu.edu.cn（通信联系人）

置图如图 1 所示。产生 THz 波的抽运光源是重复频率为 82 MHz 的锁模钛宝石激光器，产生飞秒激光的中心波长为 810 nm，脉宽为 100 fs，平均功率为 980 mW。实验中，锁相放大器积分时间是 100 ms，信号时域峰值处的信噪比可达到 600。实验时样品附近的空气湿度低于 4%，温度为 23 ℃。

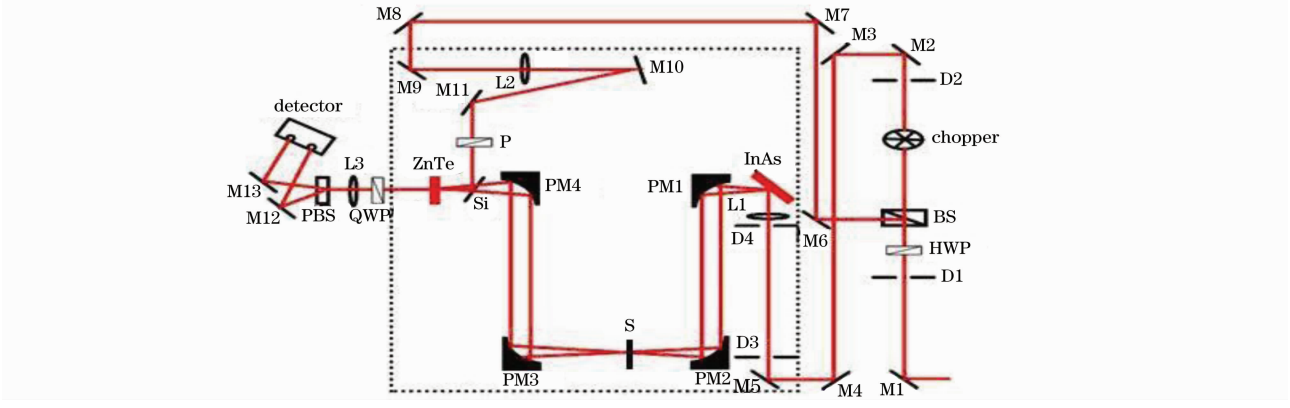

图 1 太赫兹时域光谱实验装置图

Fig. 1 Schematic setup of the THz-TDS

2.2 样品制备

将各种样品取一定质量，利用压片机以 5 T 左右的压力压制成片状样品。在不同的时间对样品进行测量。

2.3 样品吸收光谱

在光谱识别过程中，只需输入吸收光谱的归一化数据，因此从实验数据中提取相对吸收系数即可。如果不考虑边界处的能量损失，样品相对吸收系数可以用公式 $\alpha(\omega)=\ln[A_r(\omega)/A_s(\omega)]$ 求得，其中 $A_r(\omega)$ 和 $A_s(\omega)$ 分别为 THz 参考信号和样品信号时域谱傅里叶变换以后的振幅。注射用青霉素钠(1)、青霉素 V 钾片(2)、阿莫西林胶囊(3)、头孢氨苄颗粒(4)、头孢羟氨苄片(5)、注射用头孢噻肟钠(6) 6 种抗生素的吸收光谱如图 2 所示。

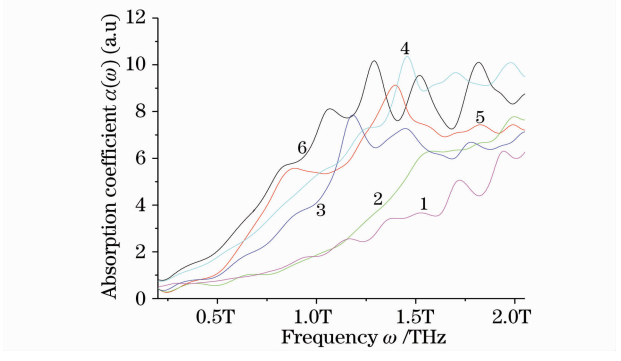

图 2 6 种样品的吸收光谱

Fig. 2 Absorption spectra of six samples

3 支持向量机

3.1 支持向量机基本原理

支持向量机[10,11]是以结构化风险最小化(SRM)代替常用的经验风险最小化(ERM)作为优化准则，其基本思想就是首先通过用内积函数定义的非线性变换将输入空间变换到一个高维空间，在这个空间中求(广义)最优分类面。如果采用不同的内积核函数将导致不同的支持向量机算法，目前采用较多的 3 类核函数包括多项式内积函数、径向基函数和 S 型内积函数。在形式上 SVM 的分类函数类似于一个神经网络，输出是中间节点的线性组合，每个中间节点对应一个支持向量，。如图 3 所示。

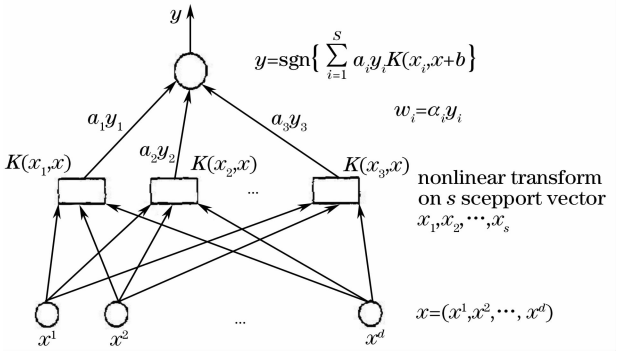

图 3 支持向量机算法

Fig. 3 Algorithm of SVM

3.2 程序及其实现

在一般的监督学习方法中，包括两个数据集，一个用于构造分类器，称为训练样本集；另一个用于检验分类器的性能，称为测试样本集。利用编制好的 libsvm 对吸收光谱进行分类。首先自己编程对实验数据进行预处理，先用得到的实验数据计算出抗生素的吸收光谱，然后对吸收光谱进行归一化。由于各种样品的吸收光谱的有效频宽不一致，因此取几种样品有效频宽重合的频率段，即 0.2～2 THz

(共计 61 个数据点)。然后将这些吸收光谱按照 libsvm 的格式输入训练样本集数据,即输入吸收光谱数据样本和对应数据编号。下一步是用网格搜索法确定模型的参数,这种搜索法就是先定义参数取值范围和步长,然后将参数进行排列组合,分别对模型进行训练,得到识别训练数据结果最好的模型参数。之后用这些参数训练并建立模型,最后用训练好的模型对测试样本集进行识别。

其中,libsvm 的核函数一共有 4 种,线性核函数、多项式核函数、径向基核函数和 S 型内积函数。由于对于一般数据,不同的核函数可以得到性能相近的结果,故在实验中取的是默认值径向基核函数。

4　实验结果及分析

用注射用青霉素钠(1)、青霉素 V 钾片(2)、阿莫西林胶囊(3)、头孢氨苄颗粒(4)、头孢羟氨苄片(5)、注射用头孢噻肟钠(6)这 6 种抗生素的吸收光谱作为训练样本(样品后面的数字为训练标签),并且再用其他时间测量的 6 种抗生素吸收光谱作为识别样本。

原始输出数据见表 1,表格中第一列为待识别数据期望输出的结果,第二列为实际输出结果。由识别结果表 2 可以看出,用 6 种抗生素作为训练集建立的模型去识别不同时间抗生素数据,是完全可以识别出来的。

表 1　输出数据

Table 1　Output data

Expected output	Actual output	Expected output	Actual output	Expected output	Actual output
3	3	1	1	2	2
3	3	1	1	2	2
4	4	6	6	5	5
4	4	6	6	5	5

表 2　分类结果

Table 2　Classification result

Sample	Classification result
Amoxicillin	Amoxicillin
Cefalexin	Cefalexin
Carbenicillin	Carbenicillin
Cefotaxime	Cefotaxime
Ospeneff	Ospeneff
Cefadroxil	Cefadroxil

以上结果表明,支持向量机可以识别不同种类的抗生素的 THz 吸收光谱,并且识别率为 100%,因此,支持向量机的识别分类效果是比较优秀的。

5　结　　论

本文将支持向量机引入到 β-内酰胺类抗生素吸收光谱识别中,达到了预期的效果。研究表明,支持向量机在实验样品的 THz 吸收光谱识别分类中识别率能够达到 100%。下一步的工作将研究不同形态的同种抗生素吸收光谱的异同及其识别,进而使 THz 技术可以应用于抗生素等药物的鉴别。

参　考　文　献

1 M. B. Devani, I. T. Patel, T. M. Patel. Rapid determination of ampicillin with fluorescence detection[J]. *Talanta*, 1992, **39**(10): 1391

2 Wang Chao, Li Shujuan. Determination of 5 penicillins in milk by HPLC with pre-column dervivation[J]. *J. Instrumental Aalysis*, 2000, **19**(6): 72～74
王　超, 李淑娟. 测定牛奶中五种青霉素残留量的高效液相色谱柱前衍生法[J]. 分析测试学报, 2000, **19**(6): 72～74

3 Cheng Cungui. Direct indentification of perilla frutescens seeds from its confusable varieties by FTIR[J]. *Spectroscopy and Spectral Analysis*, 2003, **23**(2): 282～284
程存归. FTIR 直接鉴定紫苏子及其伪品的研究[J]. 光谱学与光谱分析, 2003, **23**(2): 282～284

4 Xiong Heyu, Wang Jingwu. Progress of methods for thedetermination of cepholosporins [J]. *Jiangxi Chemical Industry*, 2003, **3**: 16～21
熊和玉,汪敬武. 头孢菌素类抗生素分析方法研究进展[J]. 江西化工, 2003, **3**: 16～21

5 Ji Te, Zhao Hongwei, Zhang Zengyan. Terahetz time-domain spectrscopy of D-, L-, and DL-penicillamines[J]. *Acta Phys.-Chem. Sin.*, 2006, **22**(9): 1159～1162
吉　特, 赵卫红 张增艳. D-,L-和 DL-青霉胺的太赫兹时域光谱[J]. 物理化学学报, 2006, **22**(9): 1159～1162

6 V. Vapnik. The Nature of Statistical Learning Theory [M]. New York: Springer Verlag, 1995

7 Bai Peng, Li Yan, Zhang Bin *et al*.. Key technologies research of mixture gas infrared spectrum analysis based on SVM[J]. *Acta Photonica Sinica*, 2008, **37**(3): 566～572
白　鹏,李　彦,张　斌 等. 基于 SVM 的混合气体红外光谱分析关键技术研究[J]. 光子学报, 2008, **37**(3): 566～572

8 Bai Peng, Wang Jianhua, Wang Hongke *et al*.. A method of mixed gas component infrared spectrum recognition based on SVM regression model [J]. *Acta Photonica Sinica*, 2008, **37**(4): 754～757
白　鹏,王建华,王宏柯 等. 基于 SVM 回归模型的混合气体组分种类光谱识别方法[J]. 光子学报, 2008, **37**(4): 754～757

9 Xiaoxia Yin, B. W.-H. Ng, B. M. Fischer *et al*.. Support vector machine applications in terahertz pulsed signals feature sets [J]. *Sensors J.*, *IEEE*, 2007, **12**(7): 1597～1608

10 Zhang Xuegong. Introduction to statistical learning theory and support vectormachines [J]. *Acta Automatica Sinica*, 2000, **26**(1): 32～42
张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, **26**(1): 32～42

11 John Shawe-Taylor, Nello Cristianini. Support Vector Machines and Other Kernel-Based Learning Methods [M]. Cambridge University Press, 2000