第29卷,第9期
2009年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 29, No. 9, pp2346-2350
September, 2009

基于支持向量机的中药太赫兹光谱鉴别

陈艳江, 刘艳艳, 赵国忠, 王卫宁, 李福利*

首都师范大学物理系, 北京 100048

摘 要 文章将支持向量机用于中药材太赫兹光谱识别。利用太赫兹光谱系统测得三组相似中药炙甘草和生甘草、南柴胡和北柴胡、山豆根和北豆根的太赫兹光谱, 傅里叶变换后得到它们的吸收系数作为分类鉴别的特征数据。用线内积函数、多项式内积函数和径向基内积函数分别构建三种不同的支持向量机, 并建立误差反传神经网络(BP 神经网络), 分别用支持向量机和 BP 神经网络对中药的特征数据进行鉴别。识别结果比较表明, 支持向量机在小样本情况下对中药两分类识别的效果明显超过 BP 神经网络。

关键词 太赫兹光谱识别; 光谱学; 支持向量机

中图分类号: O433.4 **文献标识码**: A **DOI**: 10.3964/j.issn.1000-0593(2009)09-2346-05

引 言

太赫兹(THz, 1 THz = 10^{12} Hz)辐射指的是波长在 3 mm ~30 μm 之间的电磁波, 其波段处于微波和红外之间, 属于远红外电磁辐射范畴。太赫兹时域光谱(THz-TDS)技术采用 THz 脉冲透射样品或在样品上产生反射, 测量由此产生的电磁场在时间上的变化, 通过傅里叶变换获得频域上幅值与位相的变化, 进而得到样品信息。THz 脉冲的典型脉宽在 ps 量级, 通过电光采样测量技术能够有效抑制背景噪声干扰, 信噪比远远高于傅里叶变换红外光谱技术。同时, THz 光子只有毫电子伏特能量而且对于非极性绝缘物质具有很高的穿透性, 可以做到对样品的无损检测。因此, THz 光谱技术可以成为傅里叶变换红外光谱技术和 X 射线技术的重要补充。近年来 THz 技术发展迅速, 在光谱分析和成像方面有了广泛应用[1-5]。

中药有效成分复杂。通常有机分子内化学键的振动吸收频率主要在普通红外波段, 中药在普通红外波段的光谱特性国内已有研究[6,7]。但有机分子之间较弱的相互作用(如氢键)及大分子的骨架振动(构型弯曲)、偶极子的旋转和振动跃迁以及晶体中晶格的低频振动吸收频率对应于 THz 波段范围, 这些振动所反映的分子结构及相关环境信息都与 THz 波段的频谱相对应, 这就为利用太赫兹时域光谱技术鉴别中药中化合物结构、构型与状态成为可能。在大量的中药 THz 光谱实验过程中, 很多中药在 THz 频段并没有出现明确的吸收峰, 不能用指纹谱的方法进行识别, 这很可能是中药复杂化学成分之间相互作用的结果。

如何利用 THz 频谱鉴别没有特征吸收峰的中药, 并深入研究它们的性质, 需要合适的方法。20 世纪 90 年代中期出现了一种新的机器学习方法, 即支持向量机(Support vector machine)。它是 Vapnik 等在统计学习理论上提出来的。它遵循结构风险最小的原则, 能较好地解决神经网络难以解决的小样本、非线性、高维数和存在局部极小点等实际问题[8], 因此很快成为了继神经网络研究之后新的研究热点。目前, 支持向量机在食品和药品的检测和鉴定领域取得了很多应用成果[9]。鉴于此, 本文将支持向量机与中药的太赫兹光谱技术相结合, 对没有特征峰的中药进行鉴别。

1 THz 光谱实验

本实验使用了首都师范大学 THz 实验室的 THz 光谱系统(图 1)。实验采用反射式产生太赫兹脉冲的装置。所用的发射极是 <100> InAs 晶体, 探测极是 <110> ZnTe 晶体。使用的激光器是 Spectra-Physics 公司的锁模钛宝石激光器。激光器平均功率为 0.66 W, 脉冲的中心波长为 810 nm, 脉冲宽度为 100 fs, 激光器重复率为 80 MHz。激光射出后经分束器 CBS 分为泵浦光与探测光两路。泵浦光通过斩波器 Chopper, 反射镜 M3 和 M4, 透镜 L1 聚焦到 InAs 晶体上, 从而在晶体内部沿光照方向上形成了一个内建电场, 飞秒激光脉冲激发的瞬态载流子在扩散过程中受内建电场的作用向

收稿日期: 2008-05-06, **修订日期**: 2008-08-08

基金项目: 北京市教育委员会科技发展计划面上项目(KM200710028004)资助

作者简介: 陈艳江, 1972 年生, 首都师范大学物理系研究生 e-mail: sxchengyj@163.com

*通讯联系人 e-mail: lfl-phy@sohu.com

外辐射太赫兹脉冲。第一组金镜 PM1 和 PM2 将太赫兹波聚焦到被测样品上，第二组金镜 PM3 和 PM4 将携带有样品信息的太赫兹波聚焦到探测器上。探测光经过 M7，M8，Si 片反射后也照射到探测晶体上。探测器通过测定两路光强度的不同确定太赫兹信号的波形。对样品 THz 脉冲波形进行傅里叶变换，得到其频域谱，进而得到样品的吸收系数。公式为

$$\alpha = \frac{2}{d}\ln\left\{\frac{4n(\omega)}{\rho(\omega)[n(\omega)+1]^2}\right\} \tag{1}$$

其中 ω 是信号的频率，d 是样品的厚度，$n(\omega)$ 是样品折射率，$\rho(\omega)$ 是样品的透射系数。本次实验选择六种中药，炙甘草和生甘草(Zhigancao and Shenggancas)、南柴胡和北柴胡(Nanchahu and Beichaihu)、山豆根和北豆根，均购于北京同仁堂药店。六种中药分为三组进行两两鉴别。首先将六种中药粉碎，粉碎后的药材粉末经过 200 目的筛子过滤，此时粉末颗粒的直径大小在 5 μm 左右，然后用 3 t 左右的压力压成厚度在 1～2 mm 之间，直径约 13 mm，内部均匀、两表面互相平行的薄片，最后将中药薄片放在太赫兹时域光谱系统中进行测量。每种中药测量不同位置的 15 组时域信号作为实验数据。

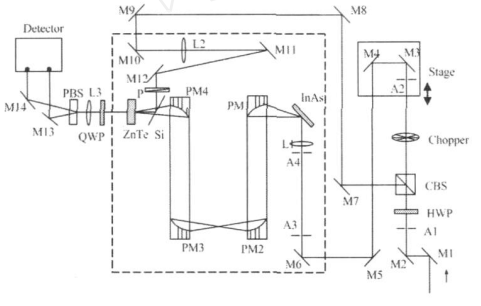

Fig. 1　THz-TDS system

2　支持向量机原理[10,11]

支持向量机是在统计学习理论和结构风险最小原理基础上发展起来的一种新的通用学习方法。它根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力。

在样本线性可分的情况下，将 N 个样本的训练集正确分类的判别函数形式为：$y(x) = w\cdot x + b$，分类面方程为：$w\cdot x + b = 0$。不仅能将两类样本正确分开，而且要求分类间隔最大的分类面就是最优分类面。最优分类面上的样本即为支持向量。最优分类面的判别函数为

$$f(x) = \mathrm{sgn}\{(w^*\cdot x) + b^*\} = \mathrm{sgn}\left\{\sum_{i=1}^{n}\partial_i^* y_i (x_i\cdot x) + b^*\right\} \tag{2}$$

式中的求和实际上只对支持向量进行。∂^* 为最优解，可以通过解下面的凸优化问题得到。在约束条件 $\sum_{i=1}^{n}\partial_i y_i = 0$，和 $\partial_i \geq 0$，$i = 1, 2, \cdots, n$ 之下对 ∂_i 求以下目标函数的最大值

$$w(\partial) = \sum_{i=1}^{n}\partial_i - \frac{1}{2}\sum_{i,j}^{n}\partial_i\partial_j y_i y_j (x_i\cdot x_j) \tag{3}$$

b^* 是分类阈值，可以通过两类中任意一对支持向量取中值求得。

对非线性问题，可以通过非线性变换转化为某个高维空间中的线性问题，在高维特征空间中求最优分类面。满足 Mercer 条件的核函数 $k(x_i, x_j) = \Phi(x_i)\cdot\Phi(x_j)$ 可以将样本变换到高维空间。这时的目标函数和相应的分类函数为

$$w(\partial) = \sum_{i=1}^{n}\partial_i - \frac{1}{2}\sum_{i,j}^{n}\partial_i\partial_j y_i y_j k(x_i\cdot x_j) \tag{4}$$

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{n}\partial_i^* y_i k(x_i\cdot x) + b^*\right\} \tag{5}$$

最常用的核函数有以下三种。

1. 线性内积函数 $k(x, y) = x\cdot y$ (6)
2. 多项式内积函数 $k(x, y) = [(x\cdot y) + 1]^d$ (7)
3. 径向基内积函数 $k(x, y) = \exp\{-|x - y|^2/\sigma^2\}$ (8)

支持向量机是为了解决二分类问题而提出来的，是基于小样本学习的理论，不需要利用样本趋向于无穷大的渐进性条件，因而在小样本情况下，能够得到好的效果。本次实验使用 Matlab 中的 Libsvm 2.85 工具箱[12]，进行相应的修改后用于数据处理。

3　实验数据处理及分析

3.1　特征数据提取

将六种中药的太赫兹时域光谱进行傅里叶变换，得到其频域光谱，然后利用(1)式，计算得到样品的吸收系数。在每种中药的 15 个实验数据中，随机抽取 1 个，其太赫兹频域谱图和吸收系数曲线如图 2 所示。

从 3 组中药的频域光谱和吸收系数曲线可以看到，在 0.3T～1.8 THz 范围内，3 组中药都没有明显的吸收峰而且每组中两种中药的吸收系数曲线都非常相似。实验将中药在 0.3～1.8 THz 范围内的吸收系数数据作为识别的特征数据。

3.2　将特征数据输入支持向量机进行训练建模

支持向量机的模型和参数选择对其实际应用非常重要。为了比较全面的了解核函数对分类效果的影响，实验采用三种最常用的核函数，并且使用相同的核函数参数。函数形式为式(6)～(8)。两分类支持向量机的参数主要有两个 C 和 g。C 是二分类支持向量机的惩罚因子，它的作用是在确定的特征空间中调节学习机器的置信范围和经验风险的比例以调节其推广能力。C 取值太大或太小都会对分类效果产生影响。g 就是式(8) 中的 σ^2，当 σ^2 太小时会产生对测试样本不具有任何泛化能力的严重“过学习”现象。而当 σ^2 太大时则会产生把所有样本都划分为一类的“欠学习”现象。本次建模设置主要参数为：$C = 10$，$g = 0.5$。将六种中药的 10 组特征数据作为样本集分别输入用三种核函数构建的支持向量机进行训练，完成三种支持向量机的建模。

3.3　将特征数据输入神经网络完成建模

三层结构的误差反传神经网络(BP 神经网络)和支持向

量机有很多相似处。利用 Matlab 7 中的神经网络工具箱建立三层结构的 BP 神经网络。设置 BP 神经网络的输入层与隐含层的传递函数为 S 型的正切函数(Tansig)，隐层与输出层的传递函数为纯线性函数(Purelin)，BP 网络所选的学习方法为带动量的自适应学习速率梯度下降法(Traingdx)，其动量因子设为 0.95。选择隐含层神经元数目为 8。对于无噪声训练集合，期望的误差为 0.001，指定训练的最大次数为 2 000次。将训练支持向量机的样本特征数据输入 BP 神经网络进行训练，确定 BP 神经网络模型。

3.4 将检测数据输入支持向量机和 BP 神经网络进行鉴别

首先将六种中药的另外 5 组特征数据分别输入支持向量机和 BP 神经网络进行鉴别，得到两种方法的检测正确率和机器运行时间如表 1，然后将 5 组特征数据加入 5 %的噪声，分别输入三种支持向量机和 BP 神经网络再进行鉴别，得到含噪情况下的检测正确率和机器运行时间如表 2。两个表中的检测的正确率为 100 次检测的平均值，时间为 100 次运行的总时间。

从表 1 可以看出，支持向量机的检测正确率均高于 BP 神经网络的检测正确率，并且支持向量机的运行时间大多仅为 BP 神经网络运行的 1/2。在表 2 中，支持向量机在检测正确率和运行时间两个指标上都明显优于 BP 神经网络。从两表中可以看出，使用三种不同内积核函数构建的支持向量机对 3 组中药的检测结果不同，因此选择不同的核函数对样品的鉴别效果有一定的影响，但从结果可以看到这种影响并不

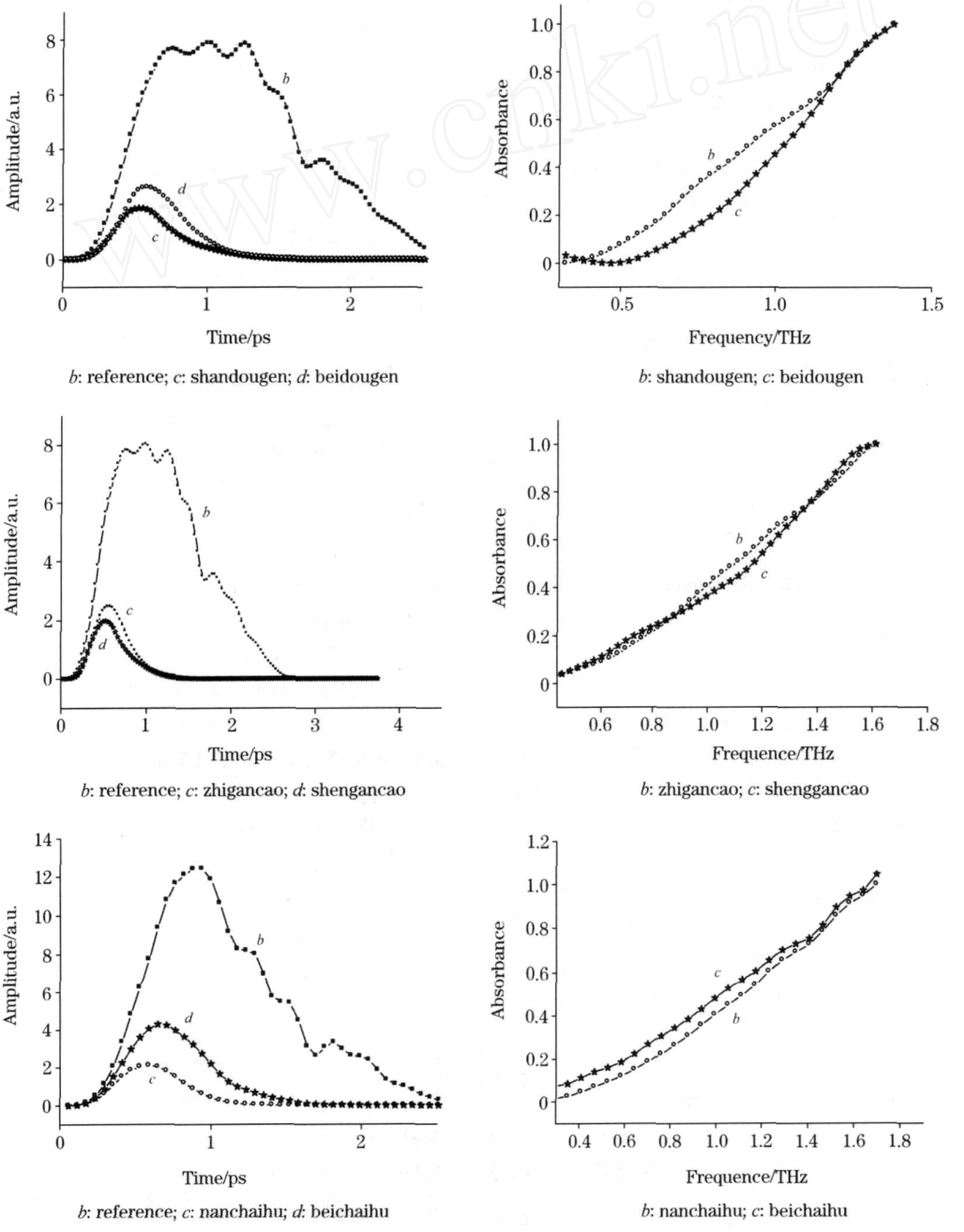

Fig. 2 THz frequency spectrum and Characteristic absorption of three groups

Table 1　Result of classification with three SVM and BP neural network analysis

	炙甘草和生甘草		南柴胡和北柴胡		山豆根和北豆根	
	正确率/ %	运行时间/ s	正确率/ %	运行时间/ s	正确率/ %	运行时间/ s
线性内积核函数 SVM	100. 0	0. 316 9	100. 0	0. 317 8	100. 0	0. 451 4
多项式内积核函数 SVM	100. 0	0. 384 1	100. 0	0. 289 4	100. 0	0. 282 0
径向基内积核函数 SVM	100. 0	0. 327 1	100. 0	0. 542 7	100. 0	0. 542 9
三层 BP 神经网络	90. 0	0. 989 6	100. 0	1. 018 5	90. 0	1. 1 243

Table 2　Result of classification with three SVM and BP neural network analysis(with noise)

	炙甘草和生甘草		南柴胡和北柴胡		山豆根和北豆根	
	正确率/ %	运行时间/ s	正确率/ %	运行时间/ s	正确率/ %	运行时间/ s
线性内积核函数 SVM	90. 0	0. 223 1	100. 0	0. 375 7	97. 8	0. 126 1
多项式内积核函数 SVM	98. 0	0. 296 2	100. 0	0. 562 0	95. 2	0. 377 8
径向基内积核函数 SVM	90. 0	0. 335 4	100. 0	0. 327 6	99. 3	0. 4 736
三层 BP 神经网络	71. 7	0. 979 2	99. 9	1. 001 8	86. 9	1. 251 1

大。造成 BP 神经网络鉴别效果不佳的原因很多，主要有训练样本少，训练时的期望误差较大，存在过学习和过拟和现象等。提高 BP 神经网络的训练精度，虽然可以提高鉴别效果，但却会大大增加网络的训练时间。相对而言，支持向量机不存在这样的问题。由于样本数目小，所以计算次数少，支持向量机可以在短时间内完成建模和检测。在含噪情况下的数据表明，支持向量机有非常强的泛化能力。

4　结　论

本文将支持向量机用于炙甘草和生甘草、南柴胡和北柴胡、山豆根和北豆根的太赫兹光谱分析。对三组样品的鉴别正确率都达到了 100 %。在含噪 5 %情况下，鉴别正确率也在 90 %以上，表明支持向量机理论结合 THz 光谱是实现中药快速、高效、准确两分类的有效方法，该方法也可在其它无特征吸收峰的小样本、两分类样品的光谱识别研究中推广应用。支持向量机对于中药多分类的情况还需要进一步的研究。

参　考　文　献

[1] Cai Y, Brener I, et al. Appl. Phys. Lett., 1998, 73: 444.

[2] Liu Haibo, Chen Yunqing, Glenn J Bastiaans, et al. Optics Express, 2006, 14: 415.

[3] Hua Zhong, Albert Redo-Sanchez, Zhang X C. Optics Express, 2006, 14, 20: 913.

[4] Wang S, Ferguson B, Abbott D, et al. Journal of Biological Physics, 2003, 29: 247.

[5] Jiang Zhiping, Zhang X C. Appl. Phys. Lett., 1998, 72: 16.

[6] SUN Su-qin, YUAN Zhi-ming, et al(孙素琴，袁子民，等). Computers and Applied Chemistry(计算机与应用化学), 2002, 19(1): 77.

[7] MA Shu-min, LIU Si-dong, ZHANG Zhuo-yong(马书民，刘思东，张卓勇). Computers and Applied Chemistry(计算机与应用化学), 2007, 24(1): 121.

[8] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[9] CHEN Quan-sheng, et al(陈全胜，等). Acta Optica Sinica(光学学报), 2006, 6(26): 933.

[10] BIAN Zhao-qi, ZHANG Xue-gong(边肇祺，张学工). Pattern Recongnition(模式识别). Beijing: Tsinghua University Press(北京：清华大学出版社), 2002.

[11] Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods(支持向量机导论). Translated by LI Guo-zheng, WANG Meng, ZENG Hua-jun(李国正，王　猛，曾华军，译). Beijing: Publishing House of Electronic Industry(北京：电子工业出版社), 2004.

[12] http://www.csie.ntu.edu.tw/~cjlin/

Chinese Traditional Medicine Recognition by Support Vector Machine (SVM) Terahertz Spectrum

CHEN Yan-jiang, LIU Yan-yan, ZHAO Guo-zhong, WANG Wei-ning, LI Fu-li *
Department of Physics, Capital Normal University, Beijing 100048, China

Abstract Identification is very important for the development of Chinese traditional medicines. In recent years, rapid progress in ultrafast laser technology provides a steady and available source for terahertz pulses generation, which greatly promotes the development of THz spectroscopy and imaging technique. SVM is a method for recognition of two kinds of samples. Appling SVM to the identification of Chinese traditional medicines through THz spectrum is a new way. The experiment on three groups of Chinese traditional medicines (zhigancao and shengancao, nanchaihu and beichaihu, shandougen and beidougen) was studied. The THz frequency spectrum and absorptivity were obtained and used to construct the feature space of Chinese traditional medicines. Three kinds of SVM were build, which used three kinds of kernel functions. By comparison, a model of BP artificial neural network was constructed. The result of using three kinds of SVM and BP artificial neural network to identify the Chinese traditional medicines showed that both methods have good prediction ability, but obviously the effect of SVM is better than BP artificial neural network for small samples. Using SVM in terahertz spectrum is a efficacious way for classification of Chinese traditional medicines.

Keywords Terahertz spectrum; Spectroscopy; SVM

(Received May 6, 2008; accepted Aug. 8, 2008)

* Corresponding author