文章编号：0258-7025(2009)03-752-06

用支持向量机识别毒品的太赫兹吸收光谱

赵树森 陈思嘉 沈京玲

（首都师范大学物理系，北京市太赫兹波谱与成像重点实验室，太赫兹光电子学省部共建教育部重点实验室，北京 100048）

摘要 在采用太赫兹时域光谱(THz-TDS)技术对 9 种常见毒品纯品和 3 种混合物进行实验研究，并得到它们在 0.2～2 THz频率范围的特征吸收光谱的基础上，用支持向量机(SVM)对毒品纯品和混合物的太赫兹吸收光谱进行了识别分类。用归一化预处理后的 9 种毒品和面粉的太赫兹吸收光谱训练 libsvm 模型。选用与训练光谱不同时间测得的毒品和混合物的太赫兹吸收光谱作为检测光谱，经过归一化预处理之后分别输入到训练好的 libsvm 模型中进行识别，识别率达 100%。识别结果充分表明，用支持向量机可以实现对不同种类毒品的识别和鉴定，为太赫兹光谱技术用于毒品的检测和识别提供了另一种有效的方法。

关键词 光谱学；毒品识别；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/CJL20093603.0752

Identification of Terahertz Absorption Spectra of Illicit Drugs Using Support Vector Machines

Zhao Shusen Chen Sijia Shen Jingling

(*Beijing Key Laboratory for Terahertz Spectroscopy and Imaging, Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Department of Physics, Capital Normal University, Beijing* 100048, *China*)

Abstract On the base of absorption spectra in the range of 0.2～2 THz of nine illicit drugs and three mixed drugs obtained by using terahertz time-domain spectroscopy (THz-TDS) technique, the THz absorption spectra of different illicit drugs and mixed drugs were identified by support vector machines (SVM). Absorption spectra of the nine illicit drugs and flour, which were pretreated by normalized unit, were used to train libsvm program. Absorption spectra of the illicit drugs and the mixed drugs which were measured at different time and pretreated by normalized unit too, were identified by the libsvm and the identification rate was 100%. The results indicate that it is feasible to apply SVM to identification of illicit drugs, which provides an effective method in the secure inspection and identification for illicit drugs.

Key words spectroscopy; illicit drugs identification; terahertz absorption spectra; support vector machines

1 引 言

针对不同类型和不同数量的毒品，刑法中对毒贩的量刑也有所不同，所以从对毒贩的量刑来说，迫切需要一种快速有效的识别手段。国内外在毒品检测和毒品分析方面有很多种方法[1,2]，其中化学分析，X 射线及紫外光谱对毒品样品都有一定程度的破坏，属于有损检测。其中 X 射线及紫外光谱对人体还有一定的伤害，红外光谱和拉曼光谱都存在较强的吸收和散射问题。所以迫切需要一种快速有效的无损检测方法。研究结果都表明毒品分子在太赫兹(THz)波段存在特征吸收，使得应用 THz 技术对毒品进行检测成为可能[3～6]。

支持向量机(SVM) 是 V. Vapnik [7]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。由于支持向量机在高维小训

收稿日期：2008-04-28；**收到修改稿日期**：2008-08-12

基金项目：北京市自然科学基金、北京市教育委员会科技发展计划重点项目（KZ200610028016)资助项目。

作者简介：赵树森(1983 －)，男，黑龙江人，硕士研究生，主要从事太赫兹光谱检测方面的研究。

导师简介：沈京玲(1957 －)，女，北京人，教授，主要从事太赫兹光谱、非线性光学及混沌方面的研究。

E-mail：sjl-phy@mail.cnu.edu.cn（通信作者）

练样本情况下有着很好泛化能力，并且相比于神经网络，尤其是自组织神经网络和 BP 神经网络[8,9]，在参数设定和识别过程要节省很多时间。故本文用支持向量机以 9 种毒品纯品样品和面粉 THz 相对吸收光谱（以下称吸收光谱）作为训练数据，对 9 种毒品中的某些毒品纯品样品不同时间的吸收光谱和含有某种毒品样品的混合物 THz 吸收光谱分别进行识别，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（THz－TDS）技术的透射式光谱实验装置图如图 1 所示。HWP 为 λ/2 波片，BS 为分束镜，Chopper 为折波器，PM1～4 为离轴抛面镜，QWP 为 λ/4 波片，PBS 为偏振分束镜。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，脉冲功率为 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 ℃。

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。

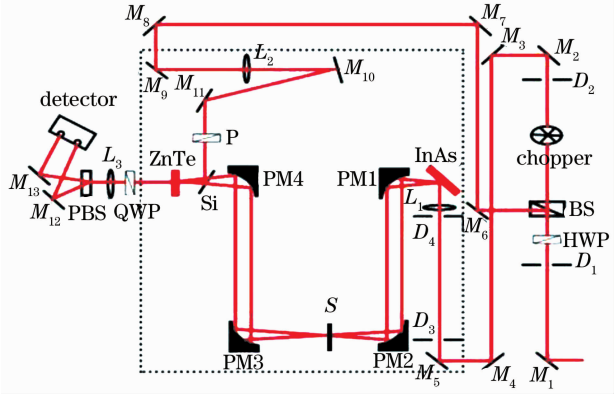

图 1　太赫兹产生和探测系统

Fig. 1　Schematic setup of the THz generation and detection system

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 $\alpha(\omega) = \ln\left[\frac{A_r(\omega)}{A_s(\omega)}\right]$ 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。以罂粟碱的吸收光谱为例，罂粟碱的相对吸收光谱如图 2 所示。

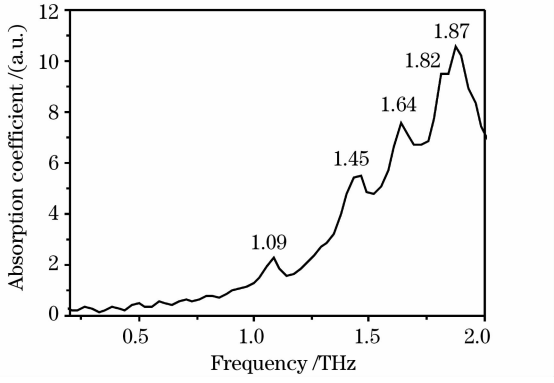

图 2　罂粟碱的相对吸收系数谱

Fig. 2　Absorption spectra of thenarceine

从图 2 可以看出，罂粟碱的吸收峰在 1.09 THz，1.45 THz，1.64 THz，1.82 THz，1.87 THz 这 5 个位置。实验表明，其他毒品在太赫兹波段也存在数量不等的吸收峰，并且各种毒品在 0.2～2.6 THz 波段内的吸收峰对应的频率不同[3～6]。这是由于毒品的不同分子结构引起的。可以用密度泛函理论计算出毒品的吸收峰，计算结果与实验结果进行比较获得了较好的一致性[3,10]。同一种毒品的太赫兹吸收谱既不受物质状态（粉末或片状）的影响，也不因时间的流逝而变化。因此称毒品的太赫兹吸收谱为“指纹谱”。“指纹谱”充分体现了毒品在太赫兹波段的吸收特性，也使利用吸收谱识别各种毒品成为可能。通过一定的识别算法（如支持向量机）能够有效地将不同毒品的太赫兹光谱识别出来，从而利用太赫兹技术进行毒品的探测和无损检测。

3　支持向量机

3.1　基本原理

支持向量机[11～15]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。其基本思想是对于非线性可分样本，将其输入向量经非线性变换映射到另一个高维空间组 Z 中。在变换后的空间中寻找一个最优的分界面（超平面），使其推广能力最好。以两类模式的分类为例说明其基本原理。

线性可分情况下的两类分类可以用图 3 的二维情况说明。图中，实心点和空心点代表两类样本，H 为分类线，H_1，H_2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫作分类间隔（margin）。所谓最优分类线就是要求

分类线不但能将两类正确分开(训练错误率为 0),而且使分类间隔最大。

设线性可分得样本集 n 有个样本(x_i,y_i),其中 $i=1,2,\ldots n$,$x\in R^N$,$y\in\{-1,1\}$ 是标识符号。在高维空间中,将两类样本无错分开的分类超平面满足

$$w\cdot x+b=0,\qquad w\in R^N,b\in R \tag{1}$$

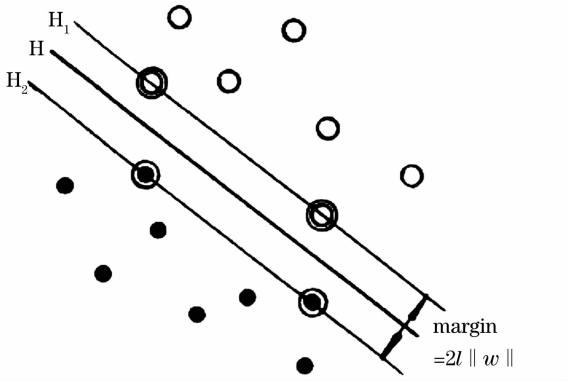

图 3 线性可分情况下的最优分类线

Fig. 3 Optimal separating line in linear classification situation

对向量系数 w 进行归一化,可以使所有样本满足 $|g(x_i)|\geqslant 1$。这样分类间隔就等于 $2/\|w\|$,因此使分类间隔最大实际上就是使 $\|w\|$ 最小;满足条件(1)且使 $2/\|w\|$ 最小的分类面就叫作最优分类面,H_1,H_2 上的训练样本点就称作支持向量。考虑到线性不可分的情况,引入了软边缘最优超平面的概念,即引入非负变量 ξ_i,最优分类面问题可表示成约束优化问题

$$y_i(x_i\cdot w+b)-1+\xi_i\geqslant 0,\quad i=1,2,\ldots,n \tag{2}$$

在(2)式约束下求函数

$$\Phi(w,\xi)=\frac{1}{2}\|w\|^2+c\Big[\sum_{i=1}^{n}\xi_i\Big] \tag{3}$$

的最小值。为此,可利用 Lagrange 函数把原问题转化为较简单的 Wolfe 对偶问题:

在约束条件

$$\sum_{i=1}^{n}y_i\alpha_i=0,\quad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\ldots,n \tag{4}$$

之下对 α_i 求解函数的最大值

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_j(x_i\cdot x_j) \tag{5}$$

求解上述问题后得到的最优分类函数为

$$f(x)=\operatorname{sgn}[(w^*\cdot x)+b^*]=\operatorname{sgn}\Big[\sum_{i=1}^{n}\alpha_i^*y_i(x_i\cdot x)+b^*\Big]。 \tag{6}$$

通过上述讨论可以看出,(6)式求和实际上只对支持向量进行。而且 b^* 为分类阈值,可以用任意一个支持向量求得,或通过两类中任意一对支持向量取中值求得。因此对于非线性可分的特征空间,考虑通过一个非线性映射 T 将特征 x 映射到高维线性特征空间 F 中。高维线性空间中的内积可以定义为

$$K(x_i,x_j)=T(x_i)\cdot T(x_j), \tag{7}$$

式中 $K(x_i,x_j)$ 称为核函数,它的选择需要满足 Mercer 条件。

此时目标函数(5) 变为

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_jy_iy_jK(x_i,x_j), \tag{8}$$

而相应的分类函数(6)式也变为

$$f(x)=\operatorname{sgn}\Big[\sum_{i=1}^{n}\alpha_i^*y_iK(x_i,x)+b^*\Big], \tag{9}$$

这就是支持向量机。

支持向量机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。SVM 分类函数形式上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 4 所示。

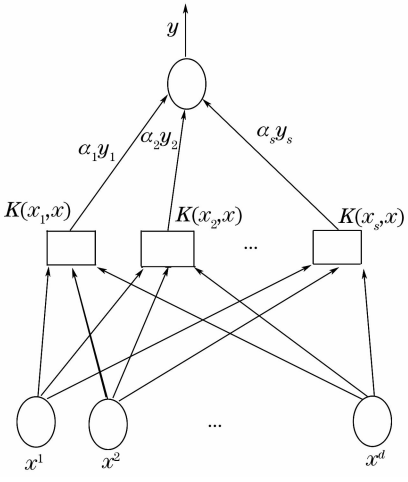

图 4 支持向量机算法图

Fig. 4 Algorithm of SVM

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。

支持向量机方法的优点在于没有必要知道映射 T 的具体形式,而只需定义高维空间中的内积运算 $K(x_i,x_j)$ 即可。即使变换后空间维数增加很多,计算的复杂度也没有太大的变化。

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毒品的吸收光谱，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，故取几种样品有效频宽重合的频率段，即0.2～2 THz。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，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，故在识别过程中取的是默认值径向基核函数。所用分类器与传统RBF方法的重要区别是，这里每个基函数中心对应一个支持向量，它们及输出权值都是由算法自动确定的。由于有多种样品，对应的是多类分类。而所谓的多类分类，是对于 N 类问题，构造 N 个两类分类器，第 i 个分类器用第 i 类训练样本作为正的训练样本，将其他类的训练样本作为负的训练样本，此时最后的输出是 N 个两类分类器输出中最大的那一类。

4　实验结果及分析

用氯胺酮(1)、甲基苯丙胺(2)、甲基麻黄碱(3)、可卡因(4)、面粉(5)、麻黄碱(6)、伪麻黄碱(7)、罂粟碱(8)、安眠酮(9)、杜冷丁(10)这10种纯品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甲基麻黄碱、可卡因、麻黄碱、伪麻黄碱、罂粟碱的5种纯品吸收光谱和3种含有氯胺酮、甲基苯丙胺的各个混合物吸收光谱作为识别样本。

原始输出数据见表1，第一列为待识别数据期望输出的结果。对于纯品即为训练时样品的标签，第二列为实际输出结果。由识别结果表2可以看出，用9种纯品和面粉的吸收光谱作为训练集建立的模型去识别不同时间测得的某些纯品本身，是完全可以识别出来的。而同样用9种纯品和面粉的吸收光谱作为训练集建立的模型去识别含有某些纯品的混合物，如果某种纯品在该混合物中的质量分数超过70%，都是可以识别出混合物含有这种纯品的。根据朗伯-比尔定律[16]可以推知，如果混合物中某种纯品的含量较高，混合物的吸收光谱就与这种纯品的吸收光谱比较相似。实验测量的结果也是如此，以甲基苯丙胺纯品和含甲基苯丙胺70%(质量分数)面粉30%(质量分数)的混合物的吸收光谱为例说明，如图5所示。

表1　原始输出数据

Table 1　Orignal output data

Expected output	Actual output
3	3
3	3
4	4
4	4
1	1
1	1
6	6
6	6
7	7
7	7
8	8
8	8
2	2
2	2
5	5
5	5

表2　识别结果

Table 2　Classification result

Sample	Classification result
Methylephedrine	Methylephedrine
Cocaine	Cocaine
Ketamine 90%＋ Polyethylene 10%	Ketamine
Ephedrine	Ephedrine
Pseudoephedrine	Pseudoephedrine
Papaverine	Papaverine
Methamphetamine 70%＋Flour 30%	Methamphetamine
Methamphetamine 20%＋Flour 80%	Flour

由图5可以看到，本次实验结果甲基苯丙胺的吸收光谱的主要吸收峰为1.26 THz，1.63 THz，

1.84 THz,以甲基苯丙胺为主的混合物的吸收峰为1.24 THz,1.60 THz,1.85 THz,各对应吸收峰的位置比较接近。以上结果说明,支持向量机不但可以识别纯品物质的 THz 吸收光谱,也可以识别在混合物中有一定含量的某种物质,所以说支持向量机本身的识别分类效果比较优秀。

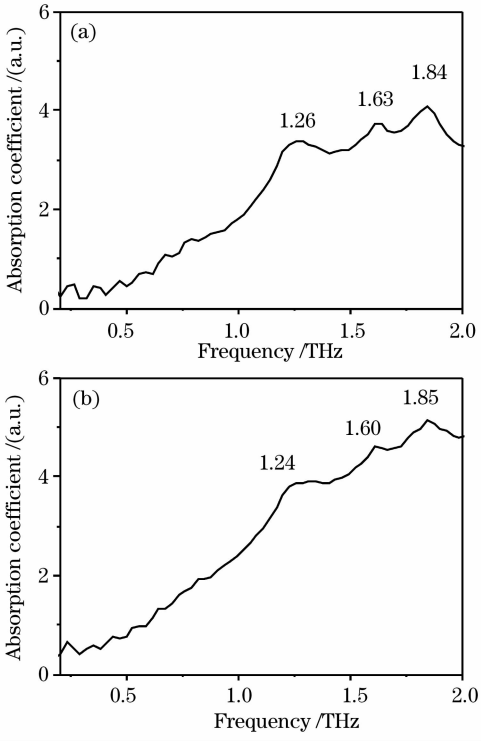

图 5 甲基苯丙胺纯品(a)和含甲基苯丙胺 70%面粉 30%的混合物(b)的吸收光谱

Fig. 5 Absorption spectra of Methamphetamine (a) and Methamphetamine 70% +flour 30% (b)

对比 BP 神经网络的识别结果[8],BP 神经网络对毒品纯品的识别率为 89%,而本文中被识别的样品中既含有纯品也含有混合物,识别率为 100%。由此可以看出在准确率方面 SVM 要优于 BP。在训练速度方面,对比 SOM 神经网络聚类方法[9],同样的训练数据,SOM 聚类训练时间为 24 h ,而 SVM 分类训练时间为 1 h,由此可以看出 SVM 在训练速度上远远优于 SOM 神经网络。此外,支持向量机用网格搜索法来确定参数,比 SOM 和 BP 需要人工操作要少。总体说来,支持向量机在识别率方面优于 BP 神经网络,在训练速度方面优于 SOM 神经网络,并且需要人工操作少。在进行分类实验的过程中,从 2 类分类开始,然后推广到 3 类分类,4 类分类直至 8 类分类,原则上可以将十几种毒品样品都作为训练样本来建立一个模型,用这个模型去识别未知毒品的类别,进而将支持向量机应用到实际的检测中。

5 结 论

将支持向量机引入到毒品吸收光谱识别中,达到了预期的效果。用纯品样品吸收系数进行训练,能够至少在混合物中某纯品占 70%以上的情况下识别出混合样品的主要成分。由于样品的限制,目前还没有低于 70%的混合物的识别结果。上述研究表明,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%,并且识别过程比起 BP 和 SOM 神经网络能节省很多时间,识别效果非常优秀。以后的工作将一方面增加训练样本的数据类型对这种方法进行更多的验证;另一方面用这种方法识别出主要成分之后,接着可以用朗伯-比尔定律计算混合物的各种纯品的百分含量。本文的实验结果充分说明支持向量机可以与 THz 时域光谱技术相结合应用于毒品的检测和识别。

参 考 文 献

1 John F. Federici, Brian Schulkin, Feng Huang *et al*.. THz imaging and sensing for security applications-explosives, weapons and drugs [J]. *Semicond. Sci. Technol.*, 2005, **20**: S266～S280

2 Hu Xuzhou, Kan Jiade, Yuan Bo. The X-ray spectrum of the heroin [J]. *Spectroscopy and Spectral Analysis*, 1999, **19**(3), 434～436
胡绪洲,阚家德,袁 波. 海洛因的 X 射线衍射谱[J]. 光谱学与光谱分析, 1999,**19**(3), 434～436

3 Li Ning, Jingling Shen, Sun Jinhai *et al*.. Study on the THz spectrum of methamphetamine [J]. *Opt. Express*, 2005, **13**(18):6750～6755

4 Meihong Lu, Jingling Shen, Ning Li *et al*.. The detection and identification of illicit drugs using terahertz imaging [J]. *J. Appl. Phys.*, 2006, **100**(10), 1031042

5 Jia Yan, Li Ning, Lu Meihong *et al*.. Inspection and identification of illicit drugs THz spectra and imaging [J]. *Modern Scientific Instruments*, 2006, **2**:41～44
贾 燕,李 宁,逮美红 等. 太赫兹光谱和成像技术在毒品识别和检测方面的应用[J]. 现代科学仪器, 2006, **2**:41～44

6 Sun Jinhai, Shen Jingling, Liang Laishun *et al*.. Experimental investigation on terahertz spectra of amphetamine type stimulants[J]. *Chin. Phys. Lett.*, 2005, **22**(12):3176～3178

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

8 Jia Yan, Chen Sijia, Li Ning *et al*.. Identification of terahertz absorption spectra of illicit drugs using back propagation neural networks [J]. *Chinese J. Lasers*, 2007, **34**(5): 719～722
贾 燕,陈思嘉,李 宁 等. 利用误差逆传播神经网络法识别几种毒品的太赫兹光谱[J]. 中国激光, 2007, **34**(5):719～722

9 Meiyan Liang, Jingling Shen, Guangqin Wang. Identification of illicit drugs by using SOM neural networks, [J]. *Phys. D: Appl. Phys.*, 2008, **41**:35306 (6pp)

10 Guangqin Wang, Jingling Shen, Yan Jia. Vibrational spectra of ketamine hydrochloride and 3,4-methylenedioxymethamphetaminein terahertz range[J]. *J. Appl. Phys.*, 2007, **102**:01310

11 Zhang Xuegong, Introduction to statistical learning theory and support vector machines[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

12 Xing Fei, Guo Ping. Stellar spectral recognition based on wavelet de－noising and svm[J]. *Spectroscopy and Spectral Analysis*, 2000, **26**(7): 1368～1372
邢　飞, 郭　平. 基于小波降噪与支持向量机的恒星光谱识别研究[J]. 光谱学与光谱分析, 2000, **26**(7), 1368～1372

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

14 Zhang Xiang, Zhang Jianqi, Jin Wei *et al*.. Method for removing sun glint from hyperspectral image [J]. *Acta Optica Sinica*, 2008, **28**(4): 664～668
张　翔, 张建奇, 靳 薇 等. 一种新的高光谱图像中太阳耀斑去除方法[J]. 光学学报, 2008, **28**(4): 664～668

15 Liu Fei, He Yong, Wang Li. Methods for the prediction of sugar content of rice wine using visible-near infrared spectroscopy [J]. *Acta Optica Sinica*, 2007, **27**(11): 2054～2058
刘　飞, 何　勇, 王　莉. 黄酒糖度预测的可见-近红外光谱方法研究[J]. 光学学报, 2007, **27**(11): 2054～2058

16 A. G. Markelz, A. Roitberg, E. J. Heilweil. Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz [J]. *Chem. Phys. Lett.*, 2000, **320**: 42～48