第31卷, 第4期
2011年4月

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

Vol. 31, No. 4, pp932-936
April, 2011

# 基于近红外光谱和支持向量机的子宫内膜癌早期诊断研究

翟 玮[1], 相玉红[1], 代荫梅[2], 张家进[1], 张卓勇[1*]

1. 首都师范大学化学系, 北京 100048
2. 首都医科大学北京妇产医院, 北京 100006

摘 要 近红外光谱结合化学计量学方法对癌症的辅助诊断已有了文献报道。该文测定了77例不同生理阶段的子宫内膜组织病理切片的近红外光谱, 对其分别进行多元散射校正(MSC)、正交信号校正(OSC)以及二者联用的预处理方法, 采用拉丁配分法选择3/4样本作为训练集, 1/4样本作测试集, 建立支持向量机(SVM)模型进行分类, 并与基于同样预处理方法建立的偏最小二乘(PLS)模型分类结果进行了比较。SVM对正常、增生和癌变三类不同的组织样品分类结果较好, 总分类正确率约92%, 好于PLS模型的结果(最高正确率90%)。研究结果表明, 光谱数据的预处理和建模方法对分类结果有重要影响, SVM结合子宫内膜组织的近红外光谱有望发展成为一种新型的肿瘤诊断方法。

关键词 近红外光谱; 子宫内膜癌; 支持向量机
中图分类号: O657.3 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2011)04-0932-05

## 引 言

子宫内膜癌(endometrial cancer, EC)是常见的女性生殖系统恶性肿瘤, 近年来发病率有明显上升的趋势[1]。引起子宫内膜癌的病因尚不是很清楚, 根据目前的研究, 一般认为与生育、激素、代谢、生理行为以及遗传等因素相关[2, 3]。子宫内膜增生(endometrial hyperplasia)属于子宫内膜癌癌前病变, 具有恶变潜能。不典型增生与高分化子宫内膜癌形态相似, 在病理诊断中易混淆[3]。基于子宫内膜增生的潜在危险性和临床诊断中的实际困难, 寻找一种能够准确分辨正常、增生以及癌变组织的方法具有重要意义。

近红外光谱是一种适用于组织病理学研究的准确、快速和经济的方法。近红外光谱(780 n~2 526 nm)主要由C—H, N—H, O—H基团分子振动的倍频和合频吸收峰组成, 可以提供组织化学成分的定性和定量信息。相比较于正常细胞, 癌变组织的血红蛋白、细胞色素、血氧饱和度等成分发生改变, 因此可以由近红外光谱检测出来, 并通过有效的模式识别技术进行聚类分析[4, 5]。因其具有无损的特点, 而得到越来越广泛的重视, 目前已被应用于直肠[4]、结肠[5]、大肠[6]、乳腺[7]、胃等多种癌症的辅助检测。近年来我们课题组开展了利用近红外光谱技术诊断子宫内膜癌的研究[8]。

支持向量机(support vector machine, SVM)是Vapnik等[9]于1995年首先提出的, 它在解决小样本、非线性及高维模式识别中表现出许多特有的优势, 并能够推广应用到其他机器学习问题中[10]。本文应用SVM建模, 对正常、增生及癌变的子宫内膜组织近红外光谱进行分类, 为子宫内膜癌的诊断提供新的方法。

## 1 理 论

### 1.1 NIR光谱的预处理方法

#### 1.1.1 多元散射校正(multiplicative scatter correction, MSC)

多元散射校正由Geladi等[11]提出, 可以去除近红外漫反射光谱中样品的镜面反射及不均匀性造成的噪声, 消除基线及光谱的不重复性。算法如下。

(1) 计算所需校正光谱的平均光谱

$$\overline{A_j} = \frac{\sum_{i=1}^{n} A_{ij}}{n} \quad (1)$$

(2) 对平均光谱作回归

$$A_i = m_i \overline{A_j} + b_i \quad (2)$$

(3) 对每一条光谱做校正

收稿日期: 2010-06-18, 修订日期: 2010-09-22
基金项目: 国家自然科学基金项目(20875065; 30772322)和北京市属高等学校人才强教计划项目(PHR20100718)资助
作者简介: 翟 玮, 女, 1986年生, 首都师范大学化学系硕士研究生 e-mail: irene_adler@sina.com
* 通讯联系人 e-mail: gusto2008@vip.sina.com

$$A_{i(\mathrm{MSC})} = \frac{A_i - b_i}{m_i} \quad (3)$$

其中 $A_i$ 为第 $i$ 个样品的光谱; $n$ 为样品数; $j$ 为波长点数; $m_i$ 和 $b_i$ 分别是线性回归得到的斜率和截距[12]。

1.1.2　正交信号校正(orthogonal signal correction, OSC)

由 Wold 等[13]提出 OSC 的思想。其基本原理是在建立定量校正模型前, 将光谱阵用浓度阵正交, 滤除与浓度阵无关的信号, 减少建立模型所用的主因子数, 达到简化模型及提高模型预测能力和稳健性的目的[14]。具体算法如下。

(1) 在原始训练集光谱阵 $\boldsymbol{X}$ 和浓度阵 $\boldsymbol{Y}$ 间建立最小二乘模型

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{B} \quad (4)$$

$$\boldsymbol{B} = (\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X} \quad (5)$$

(2) 计算残差

$$\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{Y}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X} \quad (6)$$

(3) 对 $\boldsymbol{X}_0$ 进行主成分分析(PCA)

$$[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = \mathrm{svd}(\boldsymbol{X}_0) \quad (7)$$

得到主成分矩阵 $\boldsymbol{V}[v_1, v_2, \cdots, v_n]$

(4) 对测试集光谱进行校正

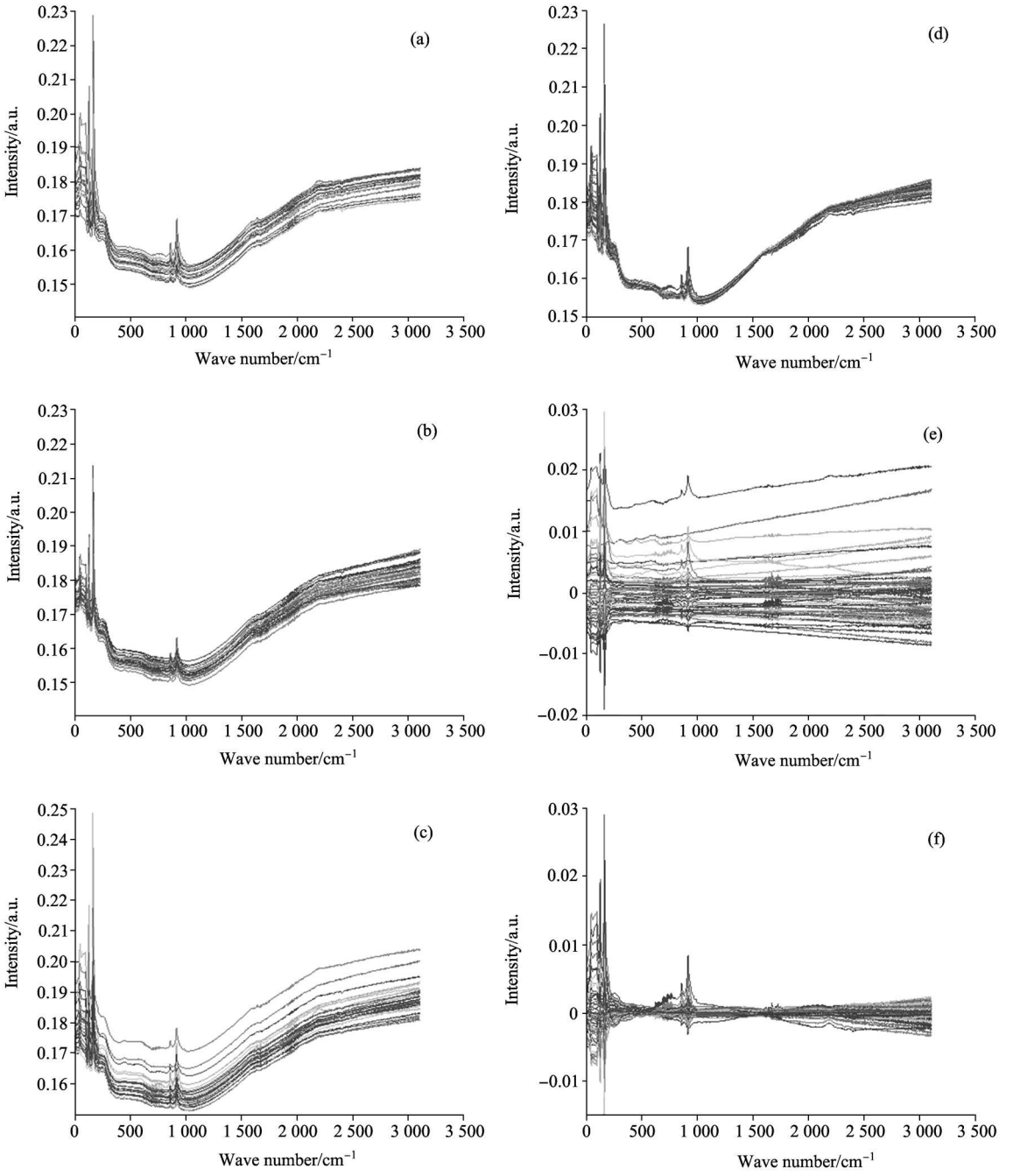


**Fig 1　Original NIR spectra of normal endometrium(a), hyperplasia endometrium(b) and malignant endometrium(c) and three pretreatment methods: MSC(d), OSC(e) and MSC+ OSC(f)**

$$\boldsymbol{X}_{(OSC)} = \boldsymbol{X}_{pre} - (\boldsymbol{X}_{pre}\boldsymbol{V})\boldsymbol{V}^{T} \quad (8)$$

**1.2** 支持向量机

给定一个训练集 $\boldsymbol{D}_i = \{(x_i, y_i),\ i = 1, 2, 3, \cdots, n\}$，其中 $x_i$ 为样本向量，$y_i \in \{-1, 1\}$ 为样本分类标识。如果训练集线性可分，SVM 的目的就是寻找一个超平面使正负两类样本可分，且最近的点到该超平面的几何间隔最大。如果训练集线性不可分，则 SVM 将样本点通过核函数(kernel function)投影到高维空间以使其线性可分。常见的核函数包括线性、径向基(RBF)、多项式和 sigmoid 核等多种形式。一般来讲，径向基核函数较稳定，通常使用较多。

$$K(x, y) = \exp\left[-\frac{\|x - y\|^2}{2\sigma^2}\right] \quad (9)$$

式中 $x$ 和 $y$ 分别表示不同样本的测量数据，$\sigma$ 为径向基核函数的宽度。

## 2 实 验

**2.1** 子宫内膜癌组织样品

本实验选用的 77 例子宫内膜癌组织切片均由首都医科大学附属北京妇产医院提供。根据病理诊断结果，子宫内膜癌组织切片 29 例，增生组织切片 30 例，正常组织切片 18 例。所有组织切片厚度均为 4 μm，常规取材，4% 甲醛固定，分别经浸蜡、包埋、切片、二甲苯脱蜡、梯度乙醇脱水、粘片及中性树胶封固等一系列技术处理制成。

**2.2** 仪器与光谱采集

本实验采用 Thrmo Electron 公司生产的 Nicolet 6700 FTIR 扩展型傅里叶变换近红外光谱分析仪，漫反射积分球采样系统，扫谱范围 4 000~ 10 000 $cm^{-1}$，光谱分辨率 4 $cm^{-1}$，光谱采样间隔 1.928 $cm^{-1}$，扫描次数为 64 次，InGaAs 检测器。在室温下仪器以空气作为空白扫描近红外光谱。数据分析软件采用近红外光谱仪自带的 OMNIC V7.3 软件。每个样品选取 5 个不同位置进行平行扫描，再将 5 个光谱平均，得到目标光谱。

**2.3** 数据处理和建模

将采集到的 77 个光谱随机分成 4 等份，其中任意 3 份作训练集，另 1 份作预测集。为保证每个样品都参与到建模中，采用自助拉丁配分(bootstrap latin partition)方法，每次配分组合得到 4 个不同的训练集和测试集，保证每个样本在测试集中出现且仅出现一次。为比较不同的预处理方法对分类结果的影响，本文分别采用单独 MSC、单独 OSC 以及 MSC 和 OSC 联用的方法对光谱进行预处理，用预处理后的光谱建立 SVM 模型。为确保实验的稳健性，整个配分过程重复多次。采用相同的光谱预处理方法，用偏最小二乘法(PLS)建模，对两种方法的分类结果进行比较。数据的预处理和建模采用 Matlab V7.8 软件。

## 3 结果与讨论

三类组织样本原始近红外光谱和经过三种方法预处理后的光谱见图 1。从图中可以看出三种样本近红外吸收曲线非常相似，特征吸收峰及吸收强度都非常接近，无法直接加以区别。因此选择合适的化学计量学方法提取样本光谱的特征信息是非常必要的。

SVM 建模预测的结果列于表 1。从表 1 可以看出，OSC 的结果好于 MSC。MSC 的一个重要和必要的假设是理想光谱和单独光谱之间的关系是波长独立的，如果这一假设不成立，则 MSC 可能只在理想光谱和单独光谱成线性关系的区域内适用。由于只有一个理想光谱用来标准化，所以 MSC 的应用可能会有问题[15]。OSC 的目标是消除 $\boldsymbol{X}$ 中与 $\boldsymbol{Y}$ 正交并在 $\boldsymbol{X}$ 中占据最大差异的一个或多个方向[16]。一般当光谱阵与浓度阵相关性不大或光谱阵背景噪音太大时，前几个主因子对应的光谱载荷往往是与浓度阵无关的光谱信号。因此在建模前通过正交法将与浓度阵无关的光谱信号滤除，可减少建立模型所用的主因子数，进一步提高模型的预测能力和

**Table 1 Classification accuracy rate of each bootstrap of three pretreatment methods**

| Method | Times | Normal | Hyperplasia | Malignant | Total |
|---|---|---|---|---|---|
| MSC | 1 | 0.824 0 | 0.942 5 | 0.933 6 | 0.912 6 |
| | 2 | 0.818 0 | 0.945 1 | 0.935 1 | 0.914 4 |
| | 3 | 0.832 5 | 0.949 8 | 0.934 8 | 0.914 4 |
| | 4 | 0.830 8 | 0.944 1 | 0.932 0 | 0.912 3 |
| | 5 | 0.827 9 | 0.947 2 | 0.934 6 | 0.916 0 |
| | mean | **0.826 6** | **0.945 7** | **0.934 0** | **0.913 9** |
| OSC | 1 | 0.841 9 | 0.919 3 | 0.949 6 | 0.932 6 |
| | 2 | 0.847 7 | 0.919 1 | 0.950 8 | 0.923 8 |
| | 3 | 0.847 2 | 0.919 0 | 0.951 4 | 0.928 9 |
| | 4 | 0.846 1 | 0.917 8 | 0.947 8 | 0.925 2 |
| | 5 | 0.840 5 | 0.922 2 | 0.956 3 | 0.926 7 |
| | mean | **0.843 5** | **0.919 5** | **0.951 2** | **0.927 4** |
| MSC+ OSC | 1 | 0.810 8 | 0.946 2 | 0.961 5 | 0.920 3 |
| | 2 | 0.813 2 | 0.948 3 | 0.961 2 | 0.921 6 |
| | 3 | 0.812 0 | 0.950 0 | 0.963 6 | 0.922 9 |
| | 4 | 0.825 8 | 0.948 1 | 0.961 7 | 0.924 6 |
| | 5 | 0.811 9 | 0.947 2 | 0.958 9 | 0.920 0 |
| | mean | **0.814 7** | **0.948 0** | **0.959 1** | **0.921 9** |

**Table 2　Classification accuracy rate of PLS and SVM models of three pretreatment methods**

| Method | Times | PLS | SVM |
|---|---|---|---|
| MSC | 1 | 0.789 5 | 0.912 6 |
| | 2 | 0.842 1 | 0.914 4 |
| | 3 | 0.947 4 | 0.914 4 |
| | 4 | 0.947 4 | 0.912 3 |
| | 5 | 0.894 7 | 0.916 0 |
| | mean | **0.884 2** | **0.913 9** |
| OSC | 1 | 0.947 4 | 0.932 6 |
| | 2 | 0.894 7 | 0.923 8 |
| | 3 | 0.894 7 | 0.928 9 |
| | 4 | 0.894 7 | 0.925 2 |
| | 5 | 0.842 1 | 0.926 7 |
| | mean | **0.894 7** | **0.927 4** |
| MSC+ OSC | 1 | 0.842 1 | 0.920 3 |
| | 2 | 0.894 7 | 0.921 6 |
| | 3 | 0.947 4 | 0.922 9 |
| | 4 | 0.947 4 | 0.924 6 |
| | 5 | 0.894 7 | 0.920 0 |
| | mean | **0.905 2** | **0.921 9** |

稳健性[14]。可能是基于以上原因，得出OSC方法的结果好于MSC。由于MSC对增生类样本分类较好，而OSC对癌变样本分类较好，故联用的方法对这两类样本都取得了比较好的分类正确率。

基于相同的预处理方法建立的PLS模型分类结果见表2。PLS模型在两种处理方法联用的情况下取得了比较好的分类正确率。通过两种模型的比较可以看出，SVM模型分类的正确率较高，模型的稳定性也比较好。

## 4　结　论

本文对正常、增生和癌变子宫内膜组织病理切片的近红外光谱进行了三种方法的预处理，建立了SVM模型对光谱数据进行分类。对三类不同的组织样品获得了较好的分类结果。并与PLS模型的正确率进行了比较。本实验结果表明，光谱数据的预处理和建模方法对分类结果有重要影响。SVM是一种适合于子宫内膜癌组织近红外光谱分辨的方法。近红外光谱分析技术结合化学计量学方法可以实现对子宫内膜癌的鉴别诊断，有望发展成为一种新型的癌症早期无创诊断方法。

## References


[1] SUN Wen-chao(孙文超). J. Int. Obstet. Gyneco.(国际妇产科学杂志), 2009, 36(4): 275.
[2] Zhang Yan, Liu Zhiwei, Yu Xinchun, et al. Gynecologic Oncology, 2010, 117: 41.
[3] YU Mei, ZHEN Jing-ran(俞　梅, 甄璟然). Journal of China Prescription Drug (中国处方药), 2009, 3: 72.
[4] Kondepati V R, Keese M, Mueller R, et al. Vibrational Spectroscopy, 2007, 44: 236.
[5] Kondepati V R, Keese M, Mueller R, et al. Vibrational Spectroscopy, 2007, 44: 56.
[6] Kondepati V R, Oszinda T, Heise H M, et al. Anal. Bioanal. Chem., 2007, 387: 1633.
[7] Bhushan K R, Misra P, Liu F B. J. Am. Chem. Soc., 2008, 130(52): 17648.
[8] ZHAO Li-ting, XIANG Yu-hong, DAI Yin-mei, et al(赵立婷, 相玉红, 代荫梅, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2010, 30(4): 901.
[9] Vapnik V. Statistical Learning Theory. New York: John Wiley, 1998.
[10] DENG Nai-yang, TIAN Ying-jie(邓乃扬, 田英杰). Support Vector Machine——Theory, Algorithm and Expanding(支持向量机——理论、算法与拓展). Beijing: Science Press(北京: 科学出版社), 2009.
[11] Geladi P, Macdou G D, Martens H. Applied Spectrosc., 1985, 39(3): 491.
[12] LU Wan-zhen(陆婉珍). Modern Near Infrared Spectroscopy Analytical Technology(Second Edition)(现代近红外光谱分析技术, 第2版). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2007.
[13] Wold S, Antti H, Lindgren F. Chemom. Intell. Lab. Syst., 1998, 44: 175.
[14] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.
[15] Andersson C A. Chemometrics and Intelligent Laboratory Systems, 1999, 47: 51.
[16] Westerhuis J A, Jong S, Smilde A K. Chemometrics and Intelligent Laboratory Systems, 2001, 56: 13.

# Early Stage Diagnosis of Endometrial Cancer Based on Near Infrared Spectroscopy and Support Vector Machine

ZHAI Wei[1], XIANG Yu-hong[1], DAI Yin-mei[2], ZHANG Jia-jin[1], ZHANG Zhuo-yong[1*]

1. Department of Chemistry, Capital Normal University, Beijing 100048, China
2. Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100006, China

**Abstract** Near-infrared spectroscopy combined with chemometrics methods for diagnosis of cancer has been reported in literatures. In our study, the NIR spectra of 77 specimens of different physiological stages of endometrium were collected. Spectral data were pretreated firstly by multiplicative scatter correction (MSC), orthogonal signal correction (OSC), and both of them, respectively, and then by SG smoothing. Latin partition method was used to select 3/4 samples as a training set, and the other 1/4 samples for test set. Support vector machine (SVM) model was built for classification, and the classification results was compared with that of partial least squares (PLS) model based on the same pretreatment methods. Samples of malignant, hyperplasia and normal endometrium were classified better by SVM (classification accuracy was 92%) than PLS (classification accuracy was 90%). The results suggested that classification accuracy was affected by pretreatment methods and models. SVM combined with endometrial tissue near infrared spectroscopy is expected to develop into a new approach to tumor diagnosis.

**Keywords** Near infrared spectroscopy; Endometrial cancer; Support vector machine

(Received Jun. 18, 2010; accepted Sep. 22, 2010)

* Corresponding author