第26卷, 第12期
2006年12月

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

Vol. 26, No. 12, pp2186-2189
December, 2006

基于 FTIR-SVM 的西洋参与籽播参的分类研究

李丹婷[1], 程存归[1*], 杜正雄[2], 何佑秋[2], 孔黎春[1]

1. 浙江师范大学化学与生命科学学院, 浙江 金华 321004
2. 西南大学化学化工学院, 重庆 400715

摘 要 支持向量机(SVM)是根据统计理论提出的一种新的学习算法。文章以40个西洋参样品为实验材料, 通过FTIR-SVM建立了西洋参样品与籽播参识别的模型。对学习训练集中的30个样品模型识别率为100%, 对10个预测样品的识别准确率为90%。研究结果表明, FTIR-SVM可以用于中药西洋参与籽播参的区别。

主题词 傅里叶变换红外光谱法; 支持向量机; 西洋参; 籽播参; 分类
中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2006)12-2186-04

引 言

西洋参为五加科植物西洋参 *Panax quinquefolium* L. 的干燥根, 其功能主治为"补气养阴, 清热生津。用于气虚阴亏, 内热, 咳喘痰血, 虚热烦倦, 消渴和口燥咽干"。人参为五加科植物人参 *Panax ginseng* C. A. Mey. 的干燥根, 为籽播参。籽播参为人参种子播种到参床上, 不移栽, 4年后挖起, 按西洋参的低温烘干法干燥而得。两者功能主治差别很大, 人参功能主治为"大补元气, 复脉固脱, 补脾益肺, 生津和安神。用于体虚欲脱, 肢冷脉微, 脾虚食少, 肺虚喘咳, 津伤口渴, 内热消渴, 久病虚羸, 惊悸失眠, 阳痿宫冷, 心力衰竭和心原性休克"。籽播参是市场上新出现的西洋参冒充品。因为籽播参实为人参的不同种植和加工方法的生晒参, 现在籽播参不经过苗移, 断面坚实无裂隙, 故很难判断。由于两者一般均为饮片, 其外观极其相像, 采用常规的药典法比较难鉴别[1]。借助现代科学中的各种手段鉴别成分复杂的中药质量, 对中药实现现代化、国际化具有十分重要的意义[2]。而针对中药材的FT IR鉴定也已有不少报道[3-6]。

支持向量机(support vector machine, SVM)是近几年产生的机器学习算法[7], 支持向量机的核心思想就是把数据非线性映射到高维特征空间, 在高维特征空间中构造具有低VC维的最优分类超平面, 使分类风险上界最小。本文在直接测定样品的FT IR基础上, 首次利用支持向量机(SVM)的多级分类器对植物中药材西洋参与籽播参进行了分类, 旨在提出一个性能优于其他的分类方法, 从而提高西洋参与籽播参的分类结果。

1 支持向量机的分类机理

人工神经网络等方法多是以经验风险最小化原则为前提, 该类方法只有在样本数趋向于无穷大时其性能才有理论上的保证。SVM方法的最大特点是包含了Vapnik结构风险最小化原则, 它不仅要求最优分类面将各类无错误的分开, 而且要使类间间隔最大, 从而保证真实风险最小。SVM方法最初是从线性可分情况下提出来的。由于实际中存在线性不可分的情况, 因此SVM方法可扩展到求解非线性分类的问题。基本思想是: 设包含 n 个样本的训练集$(x_i, y_i) \in R^d \times \{\pm 1\}$, 通过非线性映射$g: = (g_1, g_2, \cdots)$ 将输入向量x_i 变换到一个高维特征空间 Ω 的向量$g(x_i)$, 然后在这个新空间求取最优分类面, 这种非线性映射是通过定义适当的内积函数实现的。构建在特征空间 Ω 的最优超平面可以表达为

$$H(x) = \sum_{i=1}^{n} \alpha_i y_i < g(x), g(x_i) > + \alpha_0 \qquad (1)$$

特征空间 Ω 是*Hilbert* 空间。在该空间的变换中, 不必明确知道 $g(x)$ 是什么, 它只涉及核函数的内积运算

$$K(x, x_i) = < g(x), g(x_i) > \qquad (2)$$

通过适当选取满足 *Mercer* 条件的核函数 K, 就可设法将输入空间中线性不可分的样本在高维特征空间中线性分开或接近线性分开。(1) 式可改写为

$$H(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + \alpha_0 \qquad (3)$$

系数 α_i 可由求解下列优化问题得到

收稿日期: 2005-11-28, 修订日期: 2006-03-06
基金项目: 浙江省自然科学基金项目(301468)资助
作者简介: 李丹婷, 女, 1983年生, 浙江师范大学在读硕士研究生 * 通讯联系人

$$\max_{\alpha} W(\alpha) = \max_{\alpha}\left(-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{n}\alpha_i\right) \tag{4}$$

其约束为

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \tag{5}$$

$$0 \leqslant \alpha_i \leqslant C \qquad i = 1, \cdots, n \tag{6}$$

如果训练向量 x_i 对应 $\alpha_i > 0$, 那么它就是支持向量, α_0 由支持向量(x_s, y_s) 确定

$$\alpha_0 = y_s - \sum_{i=1}^{n}\alpha_i y_i K(x_i, x_s) \tag{7}$$

(6) 式中 C 为正实数, 考虑可能存在一些样本不能正确被分类而引入的松驰变量控制参数, 它起控制错分样本的惩罚程度[8]。

实现 SVM 的结构如图 1 所示。

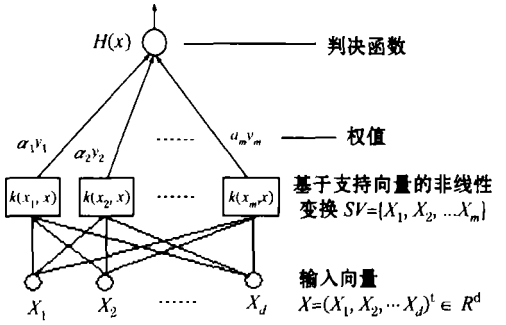

Fig 1　The structure of SVM

2　实验材料与方法

2.1　仪器和参数设置

美国 Nicolet 公司生产的 NEXUS 670 型傅里叶变换红外光谱仪, DTGS 检测器, OMNIC E. S. P. 5.1 智能操作软件, HATR, 光谱范围 4 000~ 650 cm^{-1}, 分辨率 4 cm^{-1}, 扫描累加次数 128 次。

2.2　材料

西洋参为五加科植物西洋参 *Panax quinquefolium* L. 的干燥根; 籽播参为五加科植物人参 *Panax ginseng* C. A. Mey. 的干燥根。所有样品均由国药集团杭州新亚有限公司提供, 并经过浙江省药品检验所及浙江师范大学植物教研室鉴定。

2.3　测定方法

在采集数据前, 根据仪器测试要求把 HATR 附件水平放置在傅里叶变换红外光谱仪的样品仓中, 采用单面刀分别切取样品不同部位分置于傅里叶变换红外光谱仪的HATR的金刚石与校正压力装置之间, 按照所给定的测试条件直接测定样品的 HATR-FTIR。

2.4　数据处理

采用 Matlab6.1 软件, 进行样本的回归估计, 把西洋参作为正品样本, 样本数训练选取 30 个, 试验样本数选择 10 个, 其所选特征吸收峰的 A 值作为特征信息进行 SVM 的学习。由于特殊核函数的要求, 分类器的输入样本需要规范化, 即使核函数无此要求, 规范化在大多数情况下也是有好处的, 它可以改善优化问题求解中 Hessian 矩阵的条件数, 从而保证训练方法具有更为一致的收敛性。

3　结果与分析

3.1　样品木质部及外表皮部的 FTIR 及取值标准

图 2 和图 3 分别为典型的西洋参与籽播参木质部及外表皮部的 FTIR。

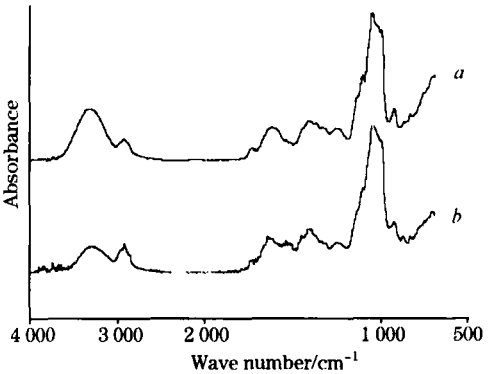

Fig 2　FTIR spectra of the xylem of *a* Panax quinquefolium L ; *b* Panax ginseng C. A. Mey

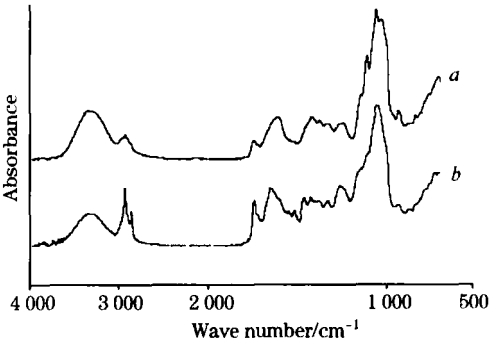

Fig 3　FTIR spectra of the primary cuticula of (*a*) Panax quinquefolium L ; (*b*) Panax ginseng C. A. Mey.

从上面 2 张谱图中可以比较发现, 西洋参的木质部及外表皮部与籽播参均有明显的差异, 但从差异的大小来看, 还是以外表皮的差异更为明显。故本文采用外表皮部进行区别鉴定。

从西洋参外表皮部与籽播参的典型红外光谱图中可看出, 除了 C—H 键的伸缩振动外, 在 1 800~ 700 cm^{-1}范围内, 其峰位置和峰强度均有较明显的差异, 这一差异为西洋参的识别奠定了一定的数学基础, 因此本文在此区间, 由以下标准取值。

(1) 纵坐标进行 T% 变换后再经过二阶导数转换来确定吸收峰位;

(2) 以每隔 8 cm^{-1}为取值标准;

(3) 二者有差异的特征吸收及所对应的红外吸光度值 A_i。

按照此取值标准共选取 30 个数据点, 这 30 个吸光度值分别可反映出二者的特征吸收的区别, 作为 SVM 学习的输

入值。

3.2 核函数的选择及分类结果

支持向量机采用核函数来将非线性分类问题转化为高维空间中的线性问题，而其常用的核函数有 3 种：多项式函数、径向基函数和 Sigmoid 函数，本文分别采用这 3 种核函数对训练样本进行学习分类。同时为了进行比较，也采用了线性可分的支持向量机函数。4 种不同函数的分类识别结果见表 1。

Table 1 The results of SVMs using different kernel(%)

样本	核函数			
	线性	多项式	径向基函数	S 函数
Train	100	100	96.7	56.7
Test	70.0	90.0	80.0	60.0

从表 1 可以看出，除了 Sigmoid 核函数的识别结果较差外，其他 3 种核函数的识别结果都较高。其中多项式核函数和径向基函数的识别准确率最高，而线性函数的识别结果相对较差，这说明对于 FTIR 鉴定问题来说，其本质还是非线性的分类问题。2 种非线性的核函数识别结果基本一致，这也与 Vapnik 和 Scholkopf 等所得的结果一致，即非线性的 SVM 会表现出大致相同的性能[9, 10]。Sigmoid 核函数的结果却出现异常结果，甚至比线性 SVM 的识别结果都差，这可能是原始问题中的数据的分布导致了 Sigmoid 核矩阵的非正常，从而引起二次优化问题解的偏差。

在 FTIR 基础上进行 SVM 对西洋参与籽播参的外表皮进行分类，结果与植物分类学相一致，表明了 FTIR-SVM 方法具有较好的分类结果。

4 结 论

FTIR 直接测定法具有方便、快速的特点，采用支持向量机进行西洋参与籽播参的 FTIR 分类具有较高的识别率。在采用 SVM 进行训练及检验时，采用非线性的分类器如多项式函数或径向基函数等核函数具有较高的准确率，线性的分类器结果稍差，而 Sigmoid 核函数由于其具有不稳定性，导致分类结果较不理想。

参 考 文 献

[1] The Pharmacopoeia Committee of People's Republic of China(国家药典委员会). Chinese Pharmacopoeia(中国药典)(2005. Vol. Ⅰ). Beijing: Chemical Industral Publishing House(北京：化学工业出版社), 2005. 87, 7.

[2] CHENG Cun-gui, YING Tao-kai(程存归，应桃开). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(1): 36.

[3] CHENG Cun-gui, RUAN Yong-ming, LI Bing-lan(程存归，阮永明，李冰岚). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(11): 1355.

[4] CHENG Cun-gui(程存归). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(2): 282.

[5] CHENG Cun-gui, SUN Cui-rong, PAN Yuan-jiang(程存归，孙翠荣，潘远江). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(9): 1055.

[6] WANG Hao, SUN Su-qin, XU Jin-wen, et al(王 昊，孙素琴，许锦文，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(2): 253.

[7] ZHANG Lu-da, SU Shi-guang, WANG Lai-sheng, et al(张录达，苏时光，王来生，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(1): 33.

[8] WANG Hao-jun, ZHENG Chong-xun, LI Ying, et al(王浩军，郑崇勋，李 映，等). J. Biomedical Engineering(生物医学工程学杂志), 2003, 20(3): 484.

[9] Vapnik V N. IEEE Trans. on Neural Networks, 1999, 10(5): 988.

[10] Scholkopf B, Smola A J. Learning with Kernels: Support Vector Machines, Regularization- Optimization and Beyond. Cambridge: MIT Press, 2002.

Classification of *Panax Quinquefolium* L. and *Panax Ginseng* C. A. Mey. Based on FTIR Analysis with SVM

LI Dan-ting[1], CHENG Cun-gui[1*], DU Zheng-xiong[2], HE You-qiu[2], KONG Li-chun[1]

1. College of Chemistry and Life Science, Zhejiang Normal University, Jinhua 321004, China
2. School of Chemistry and Chemical Engineering, Southwest University, Chongqing 400715, China

Abstract The support vector machine (SVM) is a new learning technique based on the statistical learning theory. In the present paper, forty *Panax quinquefolium* L. samples were used as experimental materials. The classification models were established using Fourier transform infrared spectra(FTIR)-SVM training method with the intention of identifying whether the *Panax quinquefolium* L. samples are genuine or they are just *Panax ginseng* C. A. Mey. samples. The thirty samples in training set were identified by the classifying models with an accurate rate of 100%, while the ten estimate samples had an accurate rate of 90%. The research result shows the feasibility of establishing the models with FTIR-SVM method to identify *Panax quinquefolium* L. samples and *Panax ginseng* C. A. Mey.

Keywords Fourier transform infrared spectra; Support vector machines; *Panax quinquefolium* L. samples; *Panax ginseng* C. A. Mey.; Classification

(Received Nov. 28, 2005; accepted Mar. 6, 2006)

* Corresponding author

《光谱学与光谱分析》对来稿英文摘要的要求

来稿英文摘要不符合下列要求者，本刊要求作者重写，这可能要推迟论文发表的时间。

1. 请用符合语法的英文，要求言简意明、确切地论述文章的主要内容，英文摘要应与中文摘要一致，且不加评论和补充解释。

2. 应拥有与论文同等量的主要信息，包括四个要素，即研究目的、方法、结果、结论。其中后两个要素最重要。有时一个句子即可包含前两个要素，例如“用某种改进的 ICP-AES 测量了鱼池水样的痕量铅”。但有些情况下，英文摘要可包括研究工作的主要对象和范围，以及具有情报价值的其他重要信息。在结果部分最好有定量数据，如检测限、相对标准偏差等；结论部分最好指出方法或结果的优点和意义。

3. 句型力求简单，尽量采用被动式，通常应有 10 个左右意义完整、语句顺畅的句子。英语词数以 150 至 200 个为宜，不能太短；也不要太长。用计算机单面隔行打印。

4. 摘要不应有引言中出现的内容，换言之，摘要中必须写进的内容应尽量避免在引言中出现。摘要也不要对论文内容作解释和评论，不得简单重复题名中已有的信息；不用非公知公用的符号和术语；不用引文，除非该论文证实或否定了他人已发表的论文。缩略语、略称、代号，除相邻专业的读者也能清楚地理解外，在首次出现时必须加以说明，例如用括号写出全称。