

近红外漫反射光谱法和模式识别技术鉴别中药材产地

刘沐华¹, 张学工^{1*}, 周群², 孙素琴^{2*}

1. 清华大学自动化系, 北京 100084
2. 清华大学化学系, 北京 100084

摘 要 采用近红外漫反射光谱法获得了来自不同产地的中药材的红外光谱, 结合近邻法和多类支持向量机等模式识别技术, 对来自四个不同产地的 269 个白芷样本和六个不同产地的 350 个野生和栽培丹参样本进行了产域鉴别, 得到的交叉验证准确率分别达到 99% 和 95%, 为中药材产地的快速无损鉴别探索了一条有效的途径。

主题词 近红外光谱; 模式识别; 中药材鉴别
中图分类号: O657.3 **文献标识码:** A **文章编号:** 1000-0593(2006)04-0629-04

引 言

中药的真伪优劣鉴别自古以来为医家所重视。中药材的伪品、混淆品直接影响了中药的质量。药材的滥用、误用、混用, 大大降低了医疗效果。因此, 中药质量的可控性研究是中医药界关注的热门课题。我们已经利用中红外光谱与模式识别技术相结合成功地实现了中药材产地的自动鉴别^[1], 本文尝试将该方法推广到近红外漫反射光谱分析领域, 从而达到无损、快速鉴别中药材的目的。

近红外漫反射光谱分析与经典的红外光谱法分析相比, 倍频、合频吸收系数小, 因此样品无需稀释即可直接测定, 并可排除微量杂质的干扰。另外, 近红外区波长短, 不被玻璃吸收, 可用玻璃样品池, 样品大多不需预处理, 所以操作费用低; 而且不需压片直接检测, 测定速度极快^[2, 3]。但同时, 近红外光谱又存在谱峰重叠、信噪比低、光谱稳定性差等技术难点。

本文对采集自河南、河北、四川、浙江四个产地的 269 个中药材白芷(ADR)样本和山东、山西、河南、四川、浙江、河北六个产地的 350 个野生、栽培中药材丹参(SMR)样本进行了近红外漫反射光谱测量, 采用近邻法和支持向量机对其一阶导数谱进行了鉴别分析。结果表明, 中红外光谱与模式识别技术相结合鉴别中药材的经验可以成功地推广到近红外漫反射光谱分析领域, 成为无损、快速、准确鉴别中药材的有效方法。

1 实验内容

1.1 样品来源
所有样品均由中国中医研究院中医药研究所提供并鉴定。样品来源如表 1 和表 2 所示。

Table 1 (a) Geographical origins of ADR samples

省份	河南	河北	四川	浙江	共计
样本数	64	60	75	70	269

Table 1 (b) Geographical origins of SMR samples

	野生/ 栽培	省份						共计
		山东	山西	河南	四川	浙江	河北	
样本数	野生	50	30	30	0	0	0	110
	栽培	70	40	25	30	35	40	240
共计		120	70	55	30	35	40	350

1.2 仪器设备与样品制备
白芷和丹参药材的近红外漫反射光谱采用 Perkin Elmer 公司的 Spectrum One NTS 光谱仪, 将粒度为 200 目的样品粉末直接放在样品台上进行测定, 获得各样品的近红外光谱图。

1.3 数据预处理
为减小实验误差, 每个样品测定 2 至 5 次, 将各次测定得到的光谱曲线加和取平均得到各样品的数据曲线。为了突

收稿日期: 2004-07-08, 修订日期: 2004-12-18
基金项目: 国家自然科学基金项目(60275014)和国家中医药管理局科技重大项目(国中医药科 2001ZDZX01)资助
作者简介: 刘沐华, 1979 年生, 清华大学自动化系硕士研究生 * 通讯联系人

©1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

出谱线的变化部分、消除各谱线基线不同带来的影响,对所有曲线都求了一阶导数,作为最后用来研究的数据曲线。

1.4 留一法交叉验证

本文用留一法来评价和比较分类结果。所谓留一法就是从 N 个样本中取出一个样本后,用剩下的 $N-1$ 个样本来设计分类器,然后用取出的那个样本作检验。这样重复 N 次,检验 N 次,并统计被错分的样本总数 K ,最后 K/N 作为错误率的估计值,则正确率的估计值为 $1-K/N$ 。其优点是有效地利用了 N 个样本,比较适用于样本数 N 较小的情况^[4]。

2 结果

2.1 红外光谱

图 1 和图 2 分别为白芷和丹参的近红外光谱图。从图中可以看出,近红外谱图的峰比较少,而且多为大而宽的峰包,不同产地的样本很难直观区分开来。

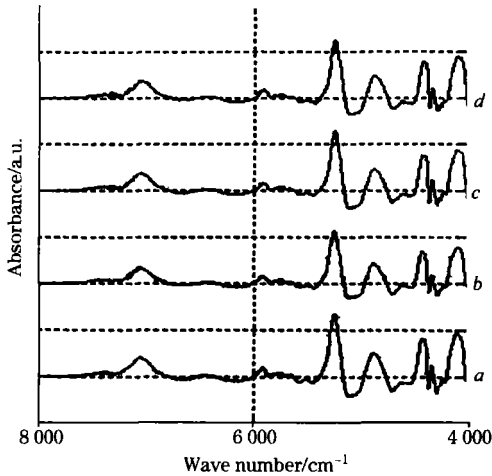


Fig 1 NIR spectra of ADR samples from different geographical origins

a: Henan; b: Hebei; c: Sichuan; d: Zhejiang

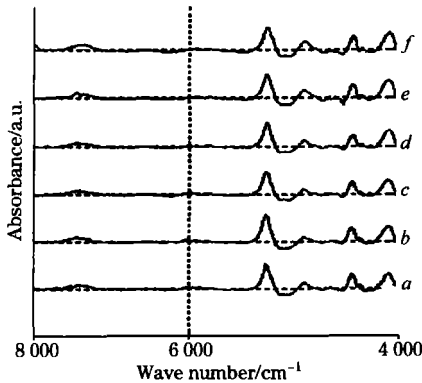


Fig 2 NIR spectra of SMR samples from different geographical origins

a: Shandong; b: Shanxi; c: Henan; d: Sichuan; e: Zhejiang; f: Hebei

2.2 近邻法

如同中红外光谱分析一样^[1],首先采用了近邻法(NNM)对白芷和丹参的近红外一阶导数谱进行了产地鉴别。近邻法^[5]是非参数模式识别方法中最重要的方法之一。它概念简单,易于实现。本文使用皮尔森相关系数评估中红外光谱的相似程度,其直观解释就是对任一未知样本,只要比较它与所有已知类别样本之间的皮尔森相关系数,并决策它与它最相似(皮尔森相关系数最大)的样本同类,具体的实验结果见表 2(表 2 是近邻法鉴别白芷(丹参)产地留一法结果)。

Table 2 (a) LOOCV results of discriminating origins of ADR samples by NNM

实际产地	样本数分类结果				小计	错判数	正确率
	河南	河北	四川	浙江			
河南	64	0	0	0	64	0	100%
河北	0	57	3	0	60	3	95%
四川	0	1	73	1	75	2	97.3%
浙江	0	0	2	68	70	2	97.1%
共计				26	9	7	97.4%

Table 2 (b) LOOCV results of discriminating wild/ cultivated SMR samples by NNM

	样本分类结果		小计	错判数	正确率
	野生	栽培			
实际野生	110	99	110	11	90%
实际栽培	240	7	240	7	97.1%
共计			350	18	94.9%

Table 2 (c) LOOCV results of discriminating origins of SMR samples by NNM

实际产地	样本分类结果						小计	错判数	正确率
	山东	山西	河南	四川	浙江	河北			
山东	117	2	0	0	0	0	120	19	7.5%
山西	1	60	8	1	0	0	70	10	85.7%
河南	0	7	48	0	0	0	55	7	87.3%
四川	0	2	0	28	0	0	30	2	93.3%
浙江	0	0	0	0	35	0	35	0	100%
河北	0	0	0	0	0	40	40	0	100%
共计							350	22	93.7%

从试验结果中可以看出,97%的白芷样本和94%的丹参样本在产地识别中被正确鉴别,95%的丹参样本在野生/栽培识别中被正确鉴别。这说明了近红外光谱技术与模式识别方法相结合进行中药产地识别的可行性。

2.3 SVM多类分类器

20世纪90年代以来,随着统计学习理论自身的不断发展和成熟,产生了支持向量机(support vector machine,简称SVM)方法。支持向量机是实现结构风险最小化准则的具体算法,是统计学习理论中最实用的部分,它在解决小样本以及高维模式识别问题中表现出许多特有的优势。支持向量机方法的思路是:首先设计函数集的有序结构使每个子集都能

使经验风险最小(比如使经验风险为 0),然后,从函数集中选择一个适当的子集,使风险的置信范围最小。对于两类样本线性可分的情况,SVM 的基本思想就是计算两类样本的最优分类面。所谓最优分类面,就是要求分类面不但能将两类无错误地分开,而且要使两类的分类空隙最大。可以证明,最优分类面不仅能保证分类器的经验风险最小,而且能使真实风险最小^[6]。对于多类问题,采用“一对多”的策略来设计分类器^[7]。本实验中 SVM 分类器的实现采用了 Ronan collobert 的 SVM 应用程序 SVMToch^[8]。

Table 3 (a) LOOCV results of discriminating the origins of ADR samples by SVM

实际产地	样本分类结果				小计	错判数	正确率
	河南	河北	四川	浙江			
河南	64	0	0	0	64	0	100%
河北	0	59	1	0	60	1	98.3%
四川	0	0	74	1	75	1	98.7%
浙江	0	0	1	69	70	1	98.6%
共计					269	3	98.9%

Table 3 (b) LOOCV results of discriminating wild/ cultivated SMR samples by SVM

	样本分类结果			小计	错判数	正确率
	野生	栽培				
实际野生	110	101	9	110	9	91.8%
实际栽培	240	10	230	240	10	95.8%
共计				350	19	94.6%

在样本数比较少的情况下,SVM 被认为是推广性最好的模式识别方法;而在多类问题中应用“一对多”的策略,也提高了分类的正确率,因为即使有两类分类器分类出错,系统也有可能根据权值得到正确的分类结果。将此方法应用于白芷和丹参的样本,产地鉴别的准确率分别达到 99%和

95%,鉴别野生和栽培丹参的准确率达到 95%(见表 3,该表是白芷(丹参)产地分类留一法结果)。

Table 3 (c) LOOCV results of discriminating origins of SMR samples by SVM

实际产地	样本分类结果						小计	错判数	正确率
	山东	山西	河南	四川	浙江	河北			
山东	118	2	0	0	0	0	120	2	99.2%
山西	0	63	7	0	0	0	70	7	97.1%
河南	0	6	49	0	0	0	55	6	89.1%
四川	0	1	0	29	0	0	30	1	96.7%
浙江	0	0	0	0	35	0	35	0	100%
河北	0	0	0	0	0	40	40	0	100%
共计							350	16	95.4%

3 结 论

本研究对光谱曲线求一阶导数,减少谱线基线漂移的影响。然后采用近邻法和支持向量机对一阶导数谱进行鉴别分析,其中支持向量机对白芷和丹参进行产地鉴别的留一法准确率分别达到 99%和 95%。

该结果与采用中红外光谱法和模式识别技术相结合鉴别中药材的结果具有可比性^[8](白芷鉴别产地正确率 99.6%,丹参鉴别产地正确率 99.2%),这表明虽然近红外漫反射光谱法较之中红外光谱法存在谱峰重叠、信噪比低、光谱稳定性差等缺点,但是通过进行合理的光谱预处理,选择合适的鉴别分析方法,近红外漫反射光谱法和模式识别技术相结合仍然可以准确鉴别中药材产地。

近红外漫反射光谱法具有不破坏样品、检测速度快的特点,本方法可以成为一种快速无损鉴别中药材的切实可行的方法。

参 考 文 献

- [1] LIU Shu-hua, ZHANG Xue-gong, ZHOU Qun, SUN Su-qin(刘沐华, 张学工, 周 群, 孙素琴). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(6): 878.
- [2] SUN Su-qin, ZHOU Qun, YU Jian-yuan(孙素琴, 周 群, 郁鉴源). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 28(2): 211.
- [3] Young-AhWoo, Hyo-Jin Kim, Jung-Hwan Cho, et al. J. Pharm. Biomed. Anal., 1999, 21: 407.
- [4] Richard O. Duda. Pattern Classification(2nd ed.). New York, USA: John Wiley and Sons, 2000.
- [5] BIAN Zhao-qi, ZHANG Xue-gong(边肇祺, 张学工). Pattern Recognition(模式识别)(2nd ed.). Beijing: Tsinghua University Press(北京: 清华大学出版社), 2000. 136, 296.
- [6] Vladimir Vapnik. The Nature of Statistical Learning Theory(2nd ed.). New York, USA: Springer-Verlag, 2000.
- [7] Ramaswamy S, Tamayo P, Rifkin R, et al. PNAS, 2001, 98: 15149.
- [8] Collobert R, Bengio S. JMLR, 2001, 1: 143.

Determination of Geographical Origins of Chinese Medical Herbs by NIR and Pattern Recognition

LIU Shu-hua¹, ZHANG Xue-gong¹, ZHOU Qun², SUN Su-qin^{2*}

1. Department of Automation, Tsinghua University, Beijing 100084, China

2. Department of Chemistry, Tsinghua University, Beijing 100084, China

Abstract Geographical origin of medical herbs is an important factor of the quality of many traditional Chinese herbal medicines. The objective of the present study is to investigate whether NIR spectroscopy coupled with pattern recognition techniques could effectively discriminate geographical origins of medical herbs. Nearest neighbor method (NNM) and a SVM-based multiclass classifier were employed to discriminate 269 Angelicae Dahuricae Radix (ADR) samples from 4 provinces and 380 Salviae Miltiorrhizae Radix (SMR) samples from 6 provinces in China. The multiclass classifier achieves leave-one-out cross-validation accuracy of 99% for (ADR) and 95% (SMR). This classification scheme can be a highly accurate approach to the rapid and nondestructive discrimination of medical herbs of different origins.

Keywords NIR; Pattern recognition; Chinese medical herb

(Received Jul. 8, 2004; accepted Dec. 18, 2004)

* Corresponding author