第33卷，第5期
2013年5月

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

Vol. 33, No. 5, pp1211-1214
May, 2013

样品表面近红外光谱结合多类支持向量机快速鉴别枸杞子产地

杜 敏[1]，巩 颖[2]，林兆洲[1]，史新元[1]，华国栋[2]*，乔延江[1]*

1. 北京中医药大学中药学院，北京 100102
2. 北京中医药大学东方医院药学部，北京 100078

摘 要 采用便携式近红外光谱仪采集枸杞子表面不同部位的近红外漫反射光谱，结合多类支持向量机算法对枸杞子产地进行快速无损辨识。以识别率为评价指标进行光谱预处理方法的选择，为了消除样本划分偏性对结果的影响，本研究通过重复划分样本集多次建模与预测，利用识别率的统计结果考察各个光谱采集部位的建模结果。实验结果表明，原始数据经二阶导数加SG平滑处理后，所建模型具有良好的产地预测性能。除了枸杞子顶端部位外，其他部位模型的稳定性及准确性均较好，其外部验证识别率的中位数与平均值均大于97%。这表明利用枸杞子样品表面近红外光谱可实现产地的准确鉴别，便携式近红外光谱技术可作为中药材流通环节中的有效监控手段。

关键词 枸杞子；产地鉴别；近红外；采集部位；识别率；多类支持向量机

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2013)05-1211-04

引 言

枸杞子为茄科植物宁夏枸杞 *Lyciumbarbarum* L. 的干燥成熟果实，具有滋补肝肾，益精明目等功效，其分布广泛，以宁夏枸杞质量较优。药材质量与其产地密切相关[1-3]，因此为保证枸杞子的质量与疗效，本工作考察近红外光谱法这种方便快速、绿色无损的分析技术[4]对枸杞子产地进行鉴别。近几年，近红外光谱法在太子参[5]、黄芩[6]、灵芝[7]等多种中药材产地鉴别中多有应用，所采用的算法主要有判别分析、判别偏最小二乘、人工神经网络和支持向量机等，其中支持向量机法[8]由于能够较好地解决小样本、非线性、高维数等实际问题而得到人们的关注与进一步开发，多类支持向量机算法已在多类分类中得到大量应用。目前尚未有研究多类支持向量机对枸杞子产地进行鉴别。同时，为拓宽近红外光谱技术在中药材现场分析中的应用，本文考察利用便携式近红外光谱仪[9]采集枸杞子表面光谱用以进行产地鉴别，实现真正快速无损的质量评价。

1 实验部分

1.1 样品

枸杞子样品分别产自内蒙古、宁夏和青海，所有药材经北京中医药大学刘春生教授鉴定为茄科植物宁夏枸杞 *Lycium barbarum* L. 的干燥成熟果实。为增加模型的适用性，从每个产地中随机抽取了不同颜色深浅、不同大小的枸杞子样本。其中内蒙古枸杞子有29个，宁夏枸杞子45个，青海枸杞子45个，具体见表1。在进行光谱采集前，用载玻片将枸杞子表面压平。

Table 1 Experiment samples

产地	DB	DM	DS	MB	MM	MS	LB	LM	LS	总计
内蒙古	5	0	5	4	0	5	5	0	5	29
宁夏	5	5	5	5	5	5	5	5	5	45
青海	5	5	5	5	5	5	5	5	5	45

Note: The first letter of row headings represent sample color(Deep, Middle, Light); the second letter represent sample size(Big, Middle, Small)

收稿日期：2012-09-18，修订日期：2012-12-12

基金项目：国家科技部“十一五”重大新药创制专项项目(2010ZX09502-002)北京中医药大学中青年教师项目(JYBZZ-JS038)资助

作者简介：杜 敏，女，1988年生，北京中医药大学在读硕士研究生 e-mail：huagong0606dumin@163.com

*通讯联系人 e-mail：yjqiao@263.net；zhjhgd@tom.com

1.2 仪器与光谱采集

仪器：Ocean quest256-2.5 便携式近红外光谱仪，配备有 InGaAs 检测器和漫反射光纤探头(光纤长度为 2 m)。其波长范围是 870.18～2 533.55 nm，共 256 个变量。

光谱采集参数：分辨率为 9.5 nm；积分时间为 100 ms；累积扫描次数 50 次；平滑度为 1。检测器温度设为－16 ℃。

样品光谱采集：将光纤探头垂直于样品表面，每个样品由基部到顶端，等间隔采集 5 个部位的光谱。每个部位采集正反面两点，每点平行采集 3 条光谱，将每个部位正反两面共 6 条光谱取平均值用于后续分析。具体光谱采集部位见图 1。

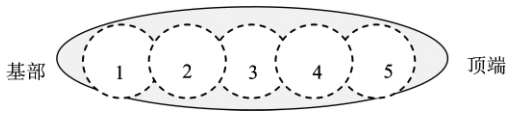

Fig. 1 Spectral acquisition sites of wolfberry fruit

1.3 数据处理

为消除枸杞子样品的光程差异，克服杂散光等外界环境的影响，首先对光谱数据进行了预处理[10]，并考察了多元散射校正(multiplicative signal correction，MSC)、标准正则变换(standard normal variate，SNV)方法、一阶导数(first derivative，1D)＋SG 平滑(Savitzky-Golay filter smoothing)、二阶导数(Second Derivative，2D)＋SG 平滑。光谱预处理采用 The Unscrambler 7.8 软件(CAMO 软件公司，挪威)。

光谱数据经优化后，本文采用基于“投票法”策略的一对一多类支持向量机算法[11](One versus-One Multi-class SVMs，1-v-1 SVMs)进行枸杞子产地的定性判别。该算法采用 Weka3.6.6 软件实现[12]，其在训练支持向量分类器时采用序列最小优化算法(sequential minimal optimization algorithm，SMO)。在建模时首先对核函数及其参数进行选择，并以识别率为评价指标进行模型筛选。

2 结果与讨论

2.1 枸杞子的原始光谱图

不同产地枸杞子的近红外光谱，以及同一样品不同部位的近红外原始光谱均严重重叠，并不能明显区分(见图 2)。由图可知，枸杞子近红外光谱比较粗糙，这是因为利用光纤探头进行光谱采集时，易受到外界杂散光影响，而且光谱两端波段噪音较大，因此在后续数据分析时采用 950～2 450 nm 这一波段。

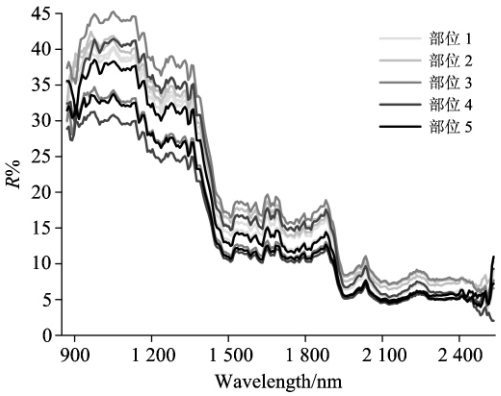

Fig. 2 Raw spectra of different sites of one wolfberry fruit sample

2.2 光谱预处理方法的选择

从表 1 中不同产地的每个类别中随机选出 1 个，共 24 个样本作为验证集(其中内蒙 6 个，宁夏与青海各 9 个)，剩余 95 个样本作为校正集。不同预处理方法下 10 折交叉验证及外部验证的结果见表 2。

Table 2 Predictive ability of SVM models with different data pre-processing methods

预处理方法	部位 1		部位 2		部位 3		部位 4		部位 5	
	10CV	验证	10CV	验证	10CV	验证	10CV	验证	10CV	验证
RAW	65.26	70.83	63.16	54.17	61.05	62.50	65.26	66.67	69.47	62.50
MSC	84.21	83.33	82.11	79.17	86.32	87.50	87.37	79.17	88.42	87.50
SNV	83.16	83.33	81.05	75.00	86.32	79.17	89.47	79.17	88.42	91.67
1D＋SG(9,2)	93.68	95.83	91.57	100.00	88.42	91.67	90.53	95.83	92.63	91.67
2D＋SG(9,2)	98.95	95.83	96.84	100.00	93.68	95.83	96.84	100.00	97.89	100.00

Note: Savitzky-Golay(data points, polynomial order), parameters of SG smoothing in table 2 were optimal

由表 2 可看出，数据处理后的建模结果明显优于原始数据，这表明近红外光谱严重受到光程差异、基线漂移和噪音等的影响。经光程校正后，模型的预测准确度明显提高，识别率由(61.05%～70.83%)增加到(81.05%～91.67%)；导数加平滑的方法结果最好，且二阶导数优于一阶导数，其 10 折交叉验证识别率均大于 93%，外部验证识别率均大于 95%。因此采用 2D＋ SG(9,2)平滑作为光谱预处理方法。

2.3 核函数的选择

对于支持向量机算法，通过适当选取核函数，可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，核函数及其参数的选择直接影响分类结果。按照 2.2 中的样本集划分方法和数据预处理方法，以部位 1 为例，对多项式核函数(Polynomial Kernel)，归一化多项式核函数(Normalized Polynomial Kernel)，PUK 核函数(Precomputed Kernel Matrix Kernel)[13]，RBF 核函数(Radial Basis Function Kernel)四种核函数进行对比分析。对每种核函数首先进行了参数优化，最优参数下不同核函数支持向量机的建模结果见表 3。

Table 3　Effects of kernel function type on the performance of SVM

核函数	参数	10 CV 识别率/%	外部验证识别率/%
多项式核函数	C=1，Exponent=3	96.84	100.0
归一化多项式核函数	C=15，Exponent=2	91.58	95.83
RBF 核函数	C=3，Gamma=0.1	90.52	95.83
PUK 核函数	C=3，Omega=1，sigma=4.0	86.32	95.83

结果表明四种核函数的建模结果均较好，其中多项式核函数的识别率最高，对外部验证集的识别率达到 100%，因此本文选用具有多项式核函数的支持向量机建立枸杞子产地的鉴别模型。

2.4　不同光谱采集部位的建模结果

为对不同部位建模结果进行综合评价，将样本集多次重复划分进行多次建模与预测。采用 Matlab 7.9 自编程序进行样本集的划分，从每个产地的样本中随机抽取三分之一作为验证集，剩余样本作为校正集。按照此法每个部位分别建立 50 个判别模型，计算每个模型十折交叉验证和外部验证的识别率，并利用箱式图对其进行描述性分析(图 3，图 4)。

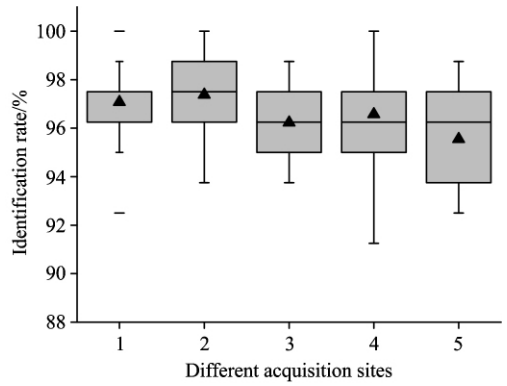

Fig. 3　Classification results of 10-fold cross validation (Note: the black triangle represents the average value)

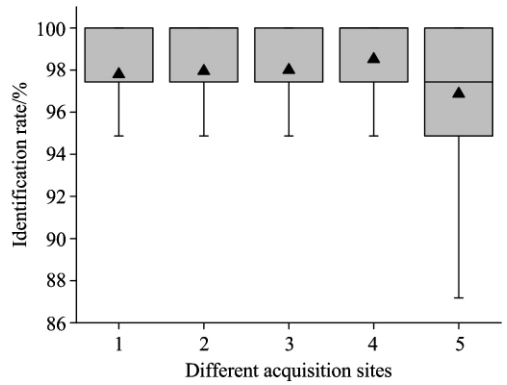

Fig. 4　Classification results of external validation (Note: the black triangle represents the avesage value)

由图 3 和图 4 可看出，各个部位十折交叉验证及外部验证识别率的分布有所差异，其中部位 1，2，3，4 的分布较集中，说明这些部位通过随机划分样本集进行建模所得结果较稳定；而部位 5 的稳定性稍差些，这可能因为枸杞子顶端在光谱采集时受操作偏差的影响相对较大。各个部位十折交叉验证识别率的中位数均大于 96%，而部位 1，2，3，4 的平均值略高于部位 5。由图 4 可看出，各个部位外部验证识别率的中位数均大于 97%，但部位 1，2，3，4 的平均值高于部位 5。在实际光谱采集中，很难准确控制光谱采集部位，因此为避免样品顶端部位以保证模型的稳定性和准确性，应尽可能采集样品中间区域的光谱。实验结果也进一步表明了采用便携式近红外光谱仪采集样品表面光谱用于产地鉴别的可行性。

3　结　论

采用便携式近红外光谱仪采集枸杞子表面光谱，以多类支持向量机法对枸杞子产地鉴别进行了研究。首先以识别率为评价指标进行光谱预处理方法的选择，然后通过多次建模与预测，利用识别率的统计结果考察各个光谱采集部位的建模结果。

结果表明，尽管表面光谱严重受到基线漂移和噪音等的影响，但经二阶导数+SG(9，2，平滑处理后，所建模型具有良好的预测性能(外部验证识别率均大于 95%)。对比枸杞子样品表面各个部位的建模结果发现，除了枸杞子顶端部位外，其他部位模型的稳定性及准确性均较好。因此在后续应用中，应尽可能采集样品中间区域的近红外光谱，求取平均值后用于分析。本研究表明利用便携式近红外光谱仪采集样品表面光谱，结合适宜的化学计量学方法对中药材进行鉴定具有可行性，为中药材流通环节中的质量控制提供了新思路。

References

[1] Amagase H, Farnsworth N R. Food Research International, 2011, 44(7): 1702.
[2] Zheng G, Zheng Z, Xu X, et al. Biochemical Systematics and Ecology, 2010, 38(3): 275.
[3] LI Mei-lan(李梅兰). Journal of Qinghai University(Nature Science)(青海大学学报·自然科学版), 2010, 28(1): 83.
[4] Blanco M, Villarroya I. TrAC Trends in Analytical Chemistry, 2002, 21(4): 240.
[5] Lin H, Zhao J, Chen Q, et al. Spectrochim Acta A Mol. Biomol. Spectrosc., 2011, 79(5): 1381.
[6] Li W, Xing L, Cai Y, et al. Vibrational Spectroscopy, 2011, 55(1): 58.

[7] LIU Zhi-gang, LI De-ren, QIN Qian-qing, et al(刘志刚, 李德仁, 秦前清, 等). Computer Engineering and Applications(计算机工程与应用), 2004, (7): 10.
[8] AN Hong, SHI Xin-yuan, WEN Jian-sheng(安 红, 史新元, 温建省). Optical Technique(光学技术), 2007, 33(1): 343.
[9] Rinnan A, Berg F V D, Engelsen S B. Trac. Trends in Analytical Chemistry, 2009, 28(10): 1201.
[10] Devos O, Ruckebusch C, Durand A, et al. Chemometrics and Intelligent Laboratory Systems, 2009, 96(1): 27.
[11] Xu Z, Dai M, Meng D. IEEE Trans Syst. Man. Cybern B Cybern, 2009, 39(5): 1292.
[12] Platt J C. Microsoft Research, Technical Report MSR-TR-98-14, 1998.
[13] Üstün B, Melssen W J, Buydens L M C. Chemometrics and Intelligent Laboratory Systems, 2006, 81(1): 29.

Rapid Identification of Wolfberry Fruit of Different Geographic Regions with Sample Surface Near Infrared Spectra Combined with Multi-Class SVM

DU Min[1], GONG Ying[2], LIN Zhao-zhou[1], SHI Xin-yuan[1], HUA Guo-dong[2]*, QIAO Yan-jiang[1]*
1. Beijing University of Chinese Medicine, Beijing 100102, China
2. Dongfang Hospital, Beijing University of Chinese Medicine, Beijing 100078, China

Abstract Portable near infrared spectrometer combined with multi-class support vector machines was used to discriminate wolfberry fruit of different geographic regions. Data pre-processing methods were explored before modeling with the identification rate as indicator. To eliminate the influence of sample subset partitioning on model performance, multiple modeling and predicting were conducted and the statistical result of identification rate was utilized to assess model performance of different acquisition sites. The results showed that SVM model with raw spectra after pretreatment of second derivative and Savitzky-Golay filter smoothing showed the best predicative ability. And the model of every acquisition site except for site 5 exhibited good stability and prediction ability and its median and average of identification rate of external validation were all greater than 97%. It was suggested that surface NIR spectra of wolfberry fruit was applicable to accurate identification of geographic region, and portable near infrared spectrometer could act as an effective means of monitoring the quality of Chinese herbal medicine in circulation.

Keywords Wolfberry fruit; Discrimination of geographic region; Near infrared; Spectra acquisition site; Identification rate; Multi-class SVM

(Received Sep. 18, 2012; accepted Dec. 12, 2012)

* Corresponding author