DOI:10.3969/j.issn.1000-1298.2010.08.023

雪莲花产地鉴别的近红外光谱分析方法*

赵杰文　蒋　培　陈全胜
(江苏大学食品与生物工程学院, 镇江 212013)

【摘要】 以青海、西藏、云南和新疆 4 个不同产地的雪莲花为研究对象,利用 K-最近邻域(KNN)模式识别方法建立雪莲花产地鉴别模型,模型参数 K 和主成分因子数(PCs)通过交互验证的方法优化;同时比较了标准正态变量变换(SNV)、多元散射校正(MSC)、一阶导数和二阶导数 4 种预处理方法对模型结果的影响。试验结果显示,通过 SNV 光谱预处理后,在 K 为 3 和 PCs 为 5 时,所得到的模型最佳,模型交互验证识别率和预测识别率均为 100%。研究表明,近红外光谱技术结合 KNN 方法可以成功鉴别雪莲花产地。

关键词: 雪莲花　产地鉴别　近红外光谱　K-最近邻域
中图分类号: O657.33; R282.5　　**文献标识码**: A　　**文章编号**: 1000-1298(2010)08-0111-04

Discrimination of Snow Lotus from Different Geographical Origins by Near Infrared Spectroscopy

Zhao Jiewen　Jiang Pei　Chen Quansheng
(*School of Food and Biological Engineering*, *Jiangsu University*, *Zhenjiang* 212013, *China*)

Abstract

Snow lotus samples from four different geographical origins (Qinghai, Tibet, Yunnan and Sinkiang) were studied. K - nearest neighbors (KNN) algorithm was applied to build discriminating model as a pattern recognition method. The parameter K of the KNN model and the number of principal component factors (PCs) were optimized. The spectra were preprocessed by four different spectral preprocessing methods of standard normal variate (SNV), multiplieative scatter correction(MSC), first derivative and second derivative, and their effects on results of KNN models were compared. Experimental results showed that the optimal model was obtained with PCs 5 and $K = 3$ after SNV spectral preprocessing, and the discriminating rates were all 100% in cross-validation and prediction. The work demonstrated that NIR spectroscopy technique with KNN method could be successfully applied to discriminate snow lotus from different geographical origins.

Key words　Snow lotus, Geographical origins discrimination, Near infrared spectroscopy, K - nearest neighbors

引言

不同产地的雪莲花由于受到海拔高度、岩层条件等外界因素影响,其内部有效成分含量也不尽相同,所以不同产地的雪莲花,其市场价格也随之各异。但在目前的中草药市场上,不同产地的雪莲花产品存在鱼目混珠的现象。

虽然原子吸收光谱[1]、电感耦合等离子体发射光谱法[1]和高效液相色谱[2]等方法能准确地鉴别不同产地的雪莲花,但这些方法均属于有损分析方法,其分析步骤繁琐,费用昂贵。近红外光谱(NIR)是一种快速无损的分析方法,近红外光谱分析技术结合模式识别方法已在食品、农产品品质的鉴别和分类中得到成功应用[3~7]。本文尝试利用红外光谱

收稿日期: 2009-09-26　修回日期: 2009-12-18
* 国家自然科学基金资助项目(30800666)和江苏省自然基金资助项目(BK2009216)
作者简介: 赵杰文,教授,博士生导师,主要从事农产品、食品无损检测及其融合技术研究, E-mail: zhao_jiewen@ujs.edu.cn

结合 K-最邻近距离(KNN)模式识别方法快速鉴别雪莲花产地。

1 材料与方法

1.1 试验材料

试验所用的材料为来自青海、西藏、云南和新疆4个产地的雪莲花。试验前,所有试验材料在105℃条件下干燥5 h,为使取样均匀,先将每株雪莲花用咖啡粉碎机粉碎,并过60目筛,然后按照四分法原则,随机称取2 g左右的粉末作为一个样本。每个产地取雪莲花25个样本,4个产地总共有100个样本。

1.2 光谱采集

试验设备为 Antaris Ⅱ 型傅里叶变换近红外光谱仪(Thermo Fisher,美国),采用高灵敏度 InGaAs 检测器,谱数范围 10 000 ~ 4 000 cm^{-1},扫描次数为32;分辨率 8 cm^{-1},数据采样间隔为 3.856 cm^{-1}。由于雪莲花样本为不透明的粉末状颗粒,所以试验采用积分球的漫反射式采样方式。在试验过程中,室内温度保持在25℃左右,湿度基本保持不变。每个样本不同时间采集3次,取其平均光谱作为该样本的原始光谱。雪莲花原始近红外光谱数据如图1所示。

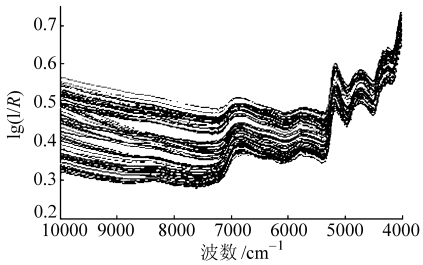

图1 雪莲花样本的原始光谱

Fig. 1 Raw NIR spectra of snow lotus samples

1.3 光谱预处理方法

试验中,雪莲花样本颗粒的粒径和样本的密实度不可能完全一致,这些必然影响到光在固体颗粒内的漫反射[8]。因此,需要对样本的原始光谱数据进行预处理。不同的预处理方法对所建立的模型有一定的影响,本研究采用标准正态变量变换(SNV)、多元散射校正(MSC)、一阶导数和二阶导数4种光谱预处理方法,并对它们的结果进行比较。SNV 首先从原光谱中减去该条光谱的平均值,再除以标准偏差,主要用于消除由于样品颗粒不均匀和密实度不一致对光谱的影响;MSC 通过对每条光谱进行移位、旋转等变化,使其尽可能的与平均光谱呈线性关系;一阶导数和二阶导数分别用于消除光谱中基线的平移和漂移散射,可有效消除其他背景干扰,分辨重叠峰,提高分辨率和灵敏度[9]。原始雪莲花光谱经过4种方法预处理后的结果如图2所示。

1.4 数据分析方法

研究最终目的是利用近红外光谱来鉴别不同产地的雪莲花,所以选择一种合适的模式识别方法来建立判别模型至关重要。本文采用 K-最近邻域法(K-nearest neighbors,简称 KNN)模式识别方法来建立模型。KNN 是一种有管理的方法,不需要对机器进行已知试样的训练,具有很好的鲁棒性,简单易用,分类准确率较高,对于大规模数据非常有效。

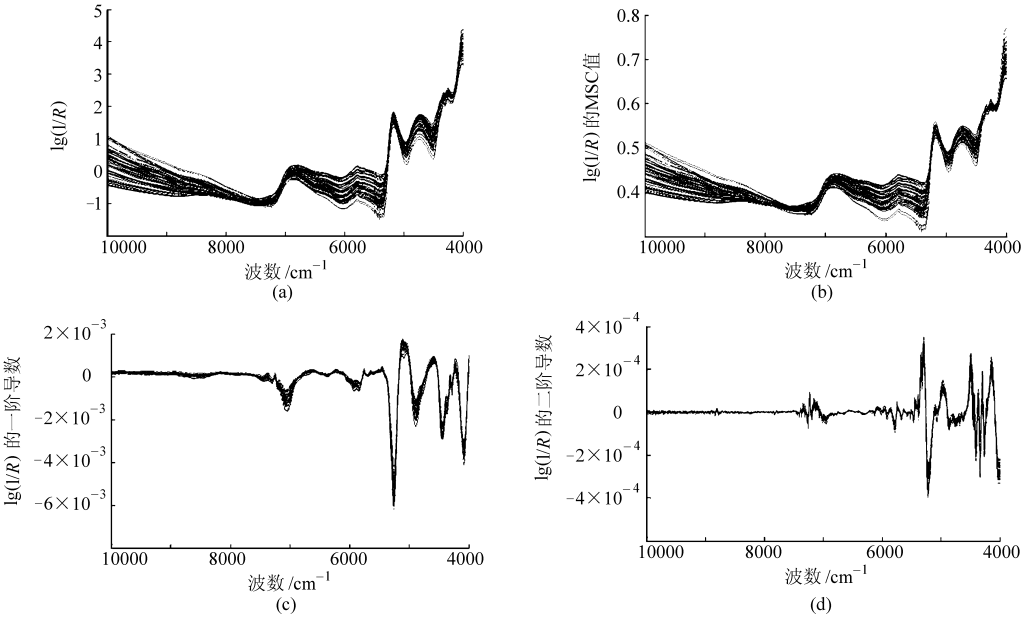

图2 经4种不同预处理后的雪莲花光谱

Fig. 2 NIR spectra of snow lotus samples after four different spectral preprocessing

(a) SNV (b) MSC (c) 一阶导数 (d) 二阶导数

KNN 基本思想是通过计算最近邻域中 K 个已知样本到未知样本的距离，对每个待判别的未知样本，逐一计算与各训练样本之间的距离，找出其中最近的 K 个进行判别[10]。在实际应用中，一般 K 取奇数，在最邻近的 K 个样本中，将未知样本归属于最多者那一类。原始试验数据通过 Results 软件（Antaris Ⅱ近红外系统自带）获取，并基于 Matlab V 6.5 软件平台分析。

2　结果与讨论

2.1　主成分分析

雪莲花中含有生物碱、黄酮、甾醇和挥发油等大量有机成分，这些有机成分的含氢基团都能在近红外区域产生倍频与合频吸收。因此，雪莲花样本的近红外光谱数据间存在大量的相关性，造成一定量的信息冗余。在建立分类识别模型时，这些冗余信息的介入会降低模型的预测性能。主成分分析是把多个指标转换为几个综合指标的一种统计方法，沿着协方差最大的方向由高维数据空间向低维数据空间投影，所得的各主成分向量相互正交，它可以将样本在高维空间的分布通过低维空间来展现。

经主成分分析处理后，取前 3 个主成分因子得分向量作图，结果如图 3 所示。其中第一主成分

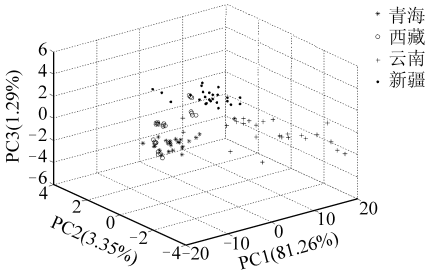

图 3　不同产地的雪莲花样本在三维主成分得分空间中的分布

Fig. 3　Distributions of snow lotus samples from four different geographical origins in the 3-D principal components space

（PC1）的方差贡献率为 81.26%，第二主成分（PC2）的方差贡献率为 3.35%，第三主成分（PC3）的方差贡献率为 1.29%；积累方差贡献率为 85.9%。从图 3 可看出，4 个产地的雪莲花样本在该得分图中有明显的分类趋势。来自青海和西藏的样本群体在空间中分布比较集中，且两类样本之间的距离较近；来自云南的样本在空间分布最为离散，且与其他样本群体空间距离较大；其次是新疆的样本群体。从中可以推断，来自西藏和青海的样本属性最为相似，其次是新疆的样本，而云南的样本与其他样本差异较大。这与青海和西藏的雪莲生长条件（海拔高度、气候和土壤等）比较相似，而云南雪莲本身的生长条件和与其他产地雪莲之间的差异性较大，这就造成了它们在图中分布比较离散，且相对其他群体的距离较远。

2.2　模式识别结果

KNN 模式识别的方法是“有监督的学习”模式识别方法[5]，基本思路是用一组已知类别的样本作为训练集，即用已知样本进行训练，得到判别模型；为检测模型的判别能力，用另外一组已知类别的样本作为预测集来验证模型。在本研究中，从每个产地中随机选取 15 个样本作为训练集，余下的 10 个作为预测集，这样训练集中总共有 60 个样本，预测集中总共有 40 个样本。

2.2.1　主成分因子数和 K 值的优化

在 KNN 模型建立过程中，通过选择合理的主成分既可以避免建模中的信息冗余，又不会过多地丢失原始特征信息，同时在分析数据时达到简化的目的。因此，有必要对这些特征变量进行主成分分析，提取主成分特征作为判别模型的输入。主成分因子数对模型的稳定性有一定影响，所以在 KNN 模型建立过程中，需要通过交互验证的方法对模型的主成分因子数进行优化。另外，在 KNN 方法中，K 值大小对判别结果有一定的影响，一般情况下还是靠经验确定，也可以通过交互验证方法优化 K 值。在参数优化的过程中，首先以 SNV 预处理的光谱来分析。通过交互验证的方法同时优化主成分因子数和 K 值 2 个参数。考虑到 K 值仅取奇数，而主成分数超过 7 后累计方差贡献率已经接近 100%，因此，2 个参数的优化区间设 K 为[1,3,5,7,9]和 PCs 为[1,2,…,7]。优化的结果如图 4 所示。从图中可看出，当 K 为 3，PCs 为 5 时，所建立的 KNN 模型识别率最高，即取得的模型最佳。此时，在校正集，模型的交互验证识别率为 100%，用该模型验证预测集中的 40 个样本时，其识别率为 100%，即所有样本都被正确识别。

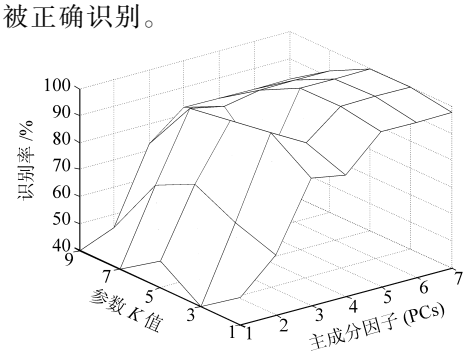

图 4　不同主成分因子和 K 水平下 KNN 模型的交互验证识别率

Fig. 4　Cross-validation recognition rate of KNN model under different principal component factors and K level

2.2.2 光谱预处理方法

光谱预处理方法对模型结果也有影响。有比较地运用了 SNV、MSC、一阶导数和二阶导数 4 种预处理数据来建立 KNN 鉴别模型,模型建立过程中,主成分因子数和 K 值的优化同 2.2.1 节。表 1 为不同的预处理条件下,模型所取得的识别结果及其所对应的主成分数和 K 值。从总体上看,4 种不同的光谱预处理方法对模型的结果影响不大,SNV 略好于 MSC、一阶导数和二阶导数,模型在校正时的交互验证识别率和预测时的识别率都达到了 100%。SNV 可以消除由于颗粒不均匀和密实度不一致而产生的散射影响[9]。因此,SNV 所取得的结果相对较好。

表 1 不同光谱预处理后 KNN 模型识别结果比较

Tab.1 Comparison of KNN models by preprocessing and raw spectra

预处理方法	主成分数	K 值	交互验证识别率/%	预测识别率/%
SNV	5	3	100	100
MSC	5	4	96.7	100
一阶导数	3	4	95.0	95.0
二阶导数	5	1	100	97.5
无	5	1	95.0	94.7

3 结束语

以青海、西藏、云南、新疆 4 个产地的雪莲花为研究对象,对获取的近红外光谱数据首先进行 SNV 光谱预处理,然后在主成分分析的基础上利用 KNN 模式识别方法建立识别模型。KNN 模式识别方法在 K 为 3、主成分因子数为 5 时所建立的识别模型最佳,模型对校正集与预测集中样本的识别率均达到 100 %。试验结果充分表明了近红外光谱结合合适的模式识别方法鉴别雪莲花产地是可行的。

参 考 文 献

1 杨若明,蓝叶芬,蓝翁驰. 两种藏药雪莲花的元素测定[J]. 中央民族大学学报:自然科学版,2005,14(2):121~123.
Yang Ruoming, Lan Yefen, Lan Wengchi, et al. The analysis of elements in flowers from two kinds of snow lotus herb of the tibetan drug [J]. Central University for Nationalities: Natural Sciences, 2005, 14 (2): 121~123. (in Chinese)

2 翟科峰,邢建国,杨伟俊. HPLC 法同时测定天山雪莲中紫丁香苷、绿原酸和芦丁的含量[J]. 药物分析杂志,2008, 28 (5):762~763.
Zhai Kefeng, Xing Jianguo, Yang Weijun. HPLC determination of syringin, chlorogenic acid and rutin in Saussurea involucrate Kar. et Kir[J]. Chinese Journal of Pharmaceutical Analysis, 2008, 28 (5):762~763. (in Chinese)

3 陈永明,林萍,何勇. 基于遗传算法的近红外光谱橄榄油产地鉴别方法研究[J]. 光谱学与光谱分析,2009, 29 (3): 671~674.
Chen Yongming, Lin Ping, He Yong. Study on discrimination of producing area of olive oil using near infrared spectra based on genetic algorithms[J]. Spectroscopy and Spectral Analysis, 2009, 29 (3): 671~674. (in Chinese)

4 刘飞,王莉,何勇. 应用可见/近红外光谱进行黄酒品种的判别[J]. 光谱学与光谱分析,2008, 28 (3): 586~589.
Liu Fei, Wang Li, He Yong. Discrimination of varieties of yellow wines using Vis/ NIR spectroscopy[J]. Spectroscopy and Spectral Analysis, 2008, 28 (3): 586~589. (in Chinese)

5 Chen Q S, Zhao J W, Liu M H, et al. Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms[J]. Journal of Pharmaceutical and Biomedical Analysis, 2008, 46(3): 1 321~1 324.

6 赵杰文,陈全胜,张海东. 近红外光谱分析技术在茶叶鉴别中的应用研究[J]. 光谱学与光谱分析,2009,26(7):1 601~1 604.
Zhao Jiewen, Chen Quansheng, Zhang Haidong. Study on the identification of tea using near infrared reflectance spectroscopy [J]. Spectroscopy and Spectral Analysis, 2009,26(7):1 601~1 604. (in Chinese)

7 吕强,汤明杰,赵杰文,等. 近红外光谱预测猕猴桃硬度模型的简化研究[J]. 光谱学与光谱分析,2009,29(7): 1 768~1 771.
Lü Qiang, Tang Mingjie, Zhao Jiewen, et al. Study of simplification of prediction model for kiwifruit firmness using near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2009 ,29(7): 1 768~1 771. (in Chinese)

8 刘翠玲,隋淑霞,吴静珠,等. 近红外光谱技术检测溶液中毒死蜱含量试验[J]. 农业机械学报,2009,40(1):129~131.
Liu Cuiling, Sui Shuxia, Wu Jingzhu, et al. Experimentation of detecting the chlorpyrifos content in solution by near infrared spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009,40(1):129~131. (in Chinese)

9 褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展,2004,16(4): 528~532.
Zhu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004, 16(4): 528~532. (in Chinese)

10 Chen Q S, Zhao J W, Lin H. Study on discrimination of roast green tea (*Camellia sinensis* L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2009, 72(4): 845~850.