王　涛,白铁成,喻彩丽,等. SPA – PLS 和 GA – PLS 算法预测胡杨叶片含水量的对比[J]. 江苏农业科学,2018,46(19):269 – 272.
doi:10.15889/j.issn.1002 – 1302.2018.19.070

SPA – PLS 和 GA – PLS 算法预测胡杨叶片含水量的对比

王　涛[1], 白铁成[1], 喻彩丽[1], 张楠楠[1], 王莎莎[2]
(1.塔里木大学信息工程学院/新疆南疆农业信息化研究中心,新疆阿拉尔 843300; 2.西北大学现代教育技术中心,陕西西安 710127)

摘要:采用 SavitZky – Golay 一阶导数法,分析叶片含水量对近红外光谱吸收谱的影响特征,建立综合利用多波段信息的作物叶片含水量预测模型。通过使用一阶导数法对胡杨叶片近红外光谱信息进行预处理,然后分别采用连续投影算法(SPA)和遗传算法(GA)筛选特征波长,建立并比较偏最小二乘回归(PLS)模型对含水量的预测效果,研究胡杨叶片含水量与叶片光谱信息的关系。试验结果发现,经过一阶导数预处理的光谱数据建模预测结果要优于原始光谱,并且 SPA – PLS 算法的回归预测结果要优于 GA – PLS 算法,其中基于一阶导数光谱使用 SPA – PLS 和 GA – PLS 算法的建模预测评价指标 RMSPCV、RMSEP、Precision、r 分别是 0.026 633、0.014 391、0.981 23、0.793 63 和 0.033 348、0.019 726、0.975 13、0.758 38,预测变量数分别是 18、29 个。说明基于一阶导数光谱使用 SPA – PLS 算法可实现胡杨叶片含水量信息的准确估测,数据优化筛选是可行的,有效提高了测量精度,减少了建模变量。

关键词:一阶导数光谱;遗传算法;连续投影算法;偏最小二乘法;胡杨;叶片;含水量
中图分类号: S127　　**文献标志码**: A　　**文章编号**:1002 – 1302(2018)19 – 0269 – 04

塔里木河流域的胡杨林对阻挡塔克拉玛干沙漠的风沙侵袭、维护区域生态平衡和保障绿洲农业起着重要作用。但近年来由于受干旱和虫害的影响,沿河两岸天然植被大幅削减和破坏,我国塔里木河流域的珍贵树种胡杨面临着生存危机[1],因此须要对胡杨林的健康状况进行及时有效的监测,胡杨叶片水分状况为胡杨干旱胁迫提供了指示作用,对胡杨林实施有效的保护具有重要的现实意义。

近红外光谱技术是一种高效率、稳定、低成本的检测方法。近年来,使用近红外方法对农产品品质进行测定主要以漫反射和透射光谱检测为主,包括蔬菜、小麦、玉米、水稻等主要农产品中水分、淀粉、蛋白质等成分含量的测定[2-6]。方美红等利用高光谱数据反演作物叶片含水量,采用小波分析方法,综合利用多波段信息的作物叶片含水量反演模型,提高了预测精度[7]。刘明博等基于连续投影算法结合主成分回归与偏最小二乘法(partial least squares regression,PLS)预测水稻叶片含氮量,证明了连续投影算法进行有效波长的选取是可行的[8]。Li 等基于遗传算法结合偏最小二乘法在植物水分近红外光谱分析模型中进行谱区选择,优化了预测模型,增强了模型的稳定性[9]。王加华等采用遗传算法用于 PLS 建立西洋梨糖度校正模型前的数据优化筛选是可行的,有效提高测量精度,减少建模变量[10]。前人利用各种光谱预处理方法,主要包括多元散射校正,矢量归一化、一阶导数、二阶导数等[11-13],分析了农产品关键成分与光谱的关系,证实了使用连续投影算法[14-16]与遗传算法[17-18]选取特征波长的可行性,而采用近红外波段的光谱信息检测胡杨叶片含水量研究鲜有报道。

本试验选用 SavitZky – Golay 一阶导数对样本的原始光谱进行预处理,然后分别使用连续投影算法(successive projection algorithm, SPA)和遗传算法(genetic algorithm, GA)[19]筛选特征波长,并结合偏最小二乘法[20]建立胡杨叶片含水量光谱预测模型,通过试验验证,该方法有效地剔除了噪声的影响,增加了特征波长的选择能力,提高了胡杨叶片含水量估测精度,从而为基于高光谱技术检测胡杨叶片含水量提供依据。

1　材料与方法

1.1　光谱采集

试验采用 Zolix Gaia Sorter 近红外成像高光谱仪,光谱测定范围 900 ~ 1 700 nm(实际测量到 1 750 nm),光谱分辨率 5 nm,光谱采样点 4 nm,共 256 个波段。样本在室内 20 ~ 25 ℃环境下进行扫描,获取一维影像和光谱信息,通过自带软件获取每张叶片的平均光谱值,每个样本数据测量 5 次取平均值,共采集 100 个样本,表 1 是根据 Kennard – Stone(K – S)算法[21]挑选出 30 份胡杨样品作为预测集,剩下的 70 份样品作为定标集。叶片水分采用烘干法进行测量,按如下公式计算:

$$\text{叶片含水量} = \frac{\text{叶片鲜质量} - \text{叶片干质量}}{\text{叶片鲜质量}} \times 100\%。$$

1.2　光谱变量选择与建模方法

1.2.1　SPA – PLS 方法　使用 SPA – PLS 方法进行特征波长选取和建立预测模型,其算法分为4个阶段:第一阶段,筛选

收稿日期:2017 – 05 – 16
基金项目:国家自然科学基金(编号:61362026);塔里木大学校长基金(编号:TDZKQN201614)。
作者简介:王　涛(1982—),男,陕西西安人,硕士,讲师,主要从事遥感与数字农业技术研究。E – mail:wujiang0156@163.com。
通信作者:白铁成,硕士,副教授,主要从事遥感与数字农业技术研究。E – mail:baitiecheng1983@163.com。

表 1　胡杨叶片校正集和预测集含水量统计

数据集	样本数（份）	含水量（%）			
		最小值	最大值	平均值	标准差
校正集	70	0.446 6	0.678 6	0.600 5	0.048 4
预测集	30	0.557 1	0.659 1	0.616 2	0.025 2

出共线性最小的若干组备选波长变量子集。第二阶段，分别使用各子集中的变量建立多元线性回归（MLR）模型，选出均方根误差（RMSE）最小的子集。第三阶段，对第二阶段选出的子集进行逐步回归建模，在尽量不损失预测准确度的前提下，得到 1 个变量数目较少的集合，该集合中的波长变量即是所选有效波长。第四阶段，对第三阶段中所选的有效波长作偏最小二乘法（PLS）的输入变量，叶片含水量作为输出变量进行预测模型的建立。SPA－PLS 具体算法过程可参阅文献[21－22]。

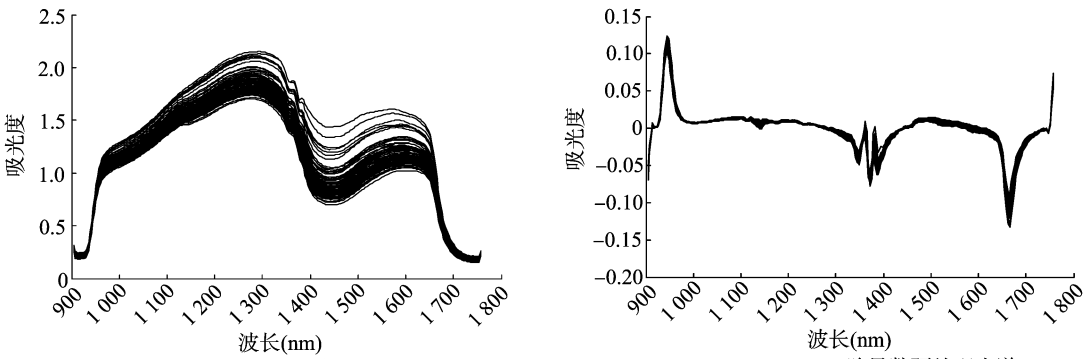
1.2.2　GA－PLS 方法　GA 算法引入染色体概念，将变量视为染色体内的基因。通过随机建立种群，利用适宜度（fitness）评价种群内个体优劣并繁衍后代，模拟自然界遗传选择规律，以优胜劣汰机制选择更适宜的基因。另外，引入交叉机制模拟种群间的基因交叉，生成新的个体保证了寻优过程的收敛，同时引入变异机制以避免结果终止于局部最优。GA－PLS 具体算法过程可参阅文献[23－24]。

1.2.3　模型精度检验　采用预测集相关系数（r）、预测集均方根误差（RMSEP）、预测精度（precision）以及交叉验证均方根误差（RMSPCV），对胡杨叶片含水量进行精度评价。模型 r 和 Precision 越高，RMSEP 和 RMSEP 越小，则模型的预测性能越好。

2　结果与分析

2.1　光谱预处理

利用 Zolix Gaia Sorter 近红外成像高光谱仪采集了 100 组胡杨叶片样本的近红外光谱吸收谱图，结果发现，在 1 280、1 420、1 620 nm 附近有明显的吸收峰、吸收谷存在，其中 1 420 nm 附近对应 H—O 键的 1 倍频波长位置[16]（图 1）。光谱仪中得到的光谱信号既包括对建模有用的光谱信息，又包含不利于建模的噪声，会影响到特征波长的选取，因此对光谱信号进行消除噪声等预处理是十分必要的。试验中应用 SavitZky－Golay 一阶导数对原始光谱进行预处理，图 1 是原始光谱与一阶导数预处理后的光谱图。

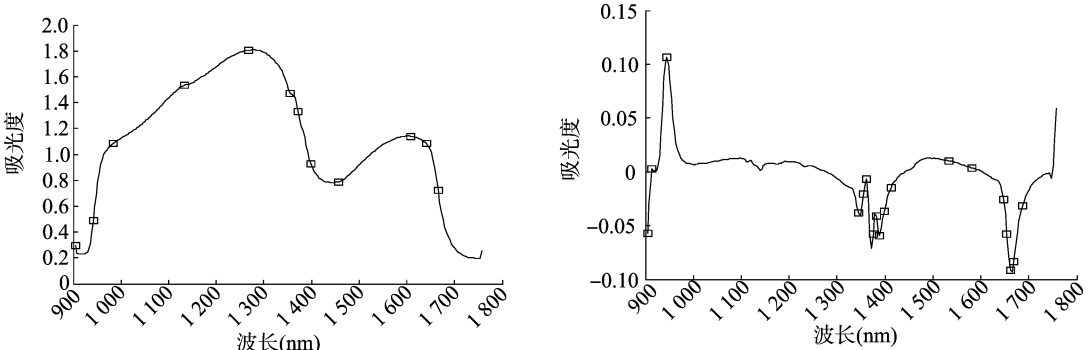


a.原始光谱　　b. SavitZky-Golay 一阶导数预处理光谱

图1　原始光谱与不同导数预处理优化后的光谱

2.2　特征波长选取

2.2.1　SPA 选取特征波长　使用连续投影算法（SPA）分别对胡杨叶片的原始光谱与一阶导数光谱数据的校正集与验证集进行 SPA 特征波长选择，SPA 选择变量数的最优区间是[2,50][15]，其中基于原始光谱选择的波长数为 21 个，且在 1 280、1 460、1 620 nm 附近集中了多数的波长，它们分布在平滑光谱中各个峰值的位置；基于 SavitZky－Golay 一阶导数选择的波长数为 16 个，且在 1 360、1 650 nm 附近集中了多数的波长，分布在一阶导数谱中各个峰值的位置，无信息的平缓区域没有波长被选取，这正是连续投影算法的优点（图 2）。

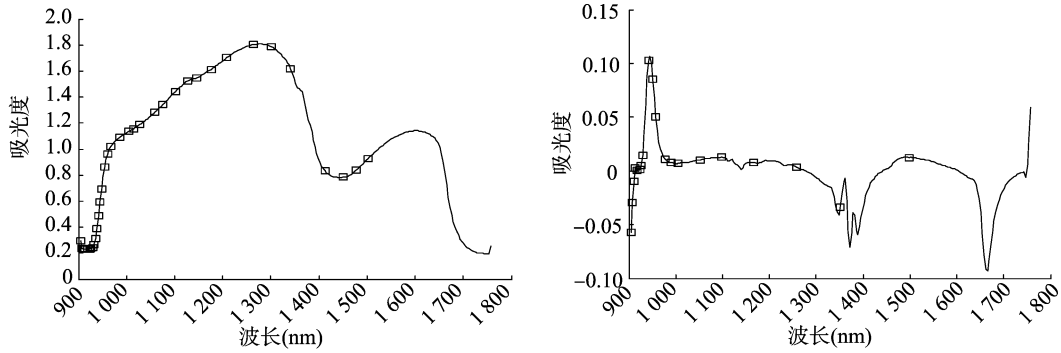


a.基于原始光谱进行波长选取　　b. 基于 SavitZky-Golay 一阶导数预处理波长选取

图2　基于原始光谱和预处理光谱 SPA 选择的特征波长

2.2.2　GA 选取特征波长　分别对原始光谱和一阶导数光谱使用 GA 方法进行特征波长的选取和对 256 个波段变量进行 GA 运算，设定遗传算法迭代次数为 100，种群大小为 30 个数据点，变异概率为 0.01、遗传概率为 0.6，依变量被选中的

频率对变量排序。为了防止算法运行过程中随机性对结果的影响,研究共进行 5 次运算,最后挑选出其中性能最好的模型所选用的频率变量作为最佳变量。每次迭代过程中,波段特征变量(优势基因)在所设定的竞争模式下保留。通过 GA 所选的特征波段主要集中在 900 ~ 1 600 nm 之间,并且在 900 ~ 1 300 nm 之前特别集中(图 3)。这是由于 GA 算法在寻优路径上的随机性造成特征波段选择数目的不确定性,即每次运行结果之间具有差异,甚至存在陷入局部最优的概率,所以基于每种预处理选择的最佳变量数存在差异,并且存在陷入 900 ~ 1 300 nm 局部最优波段的可能。

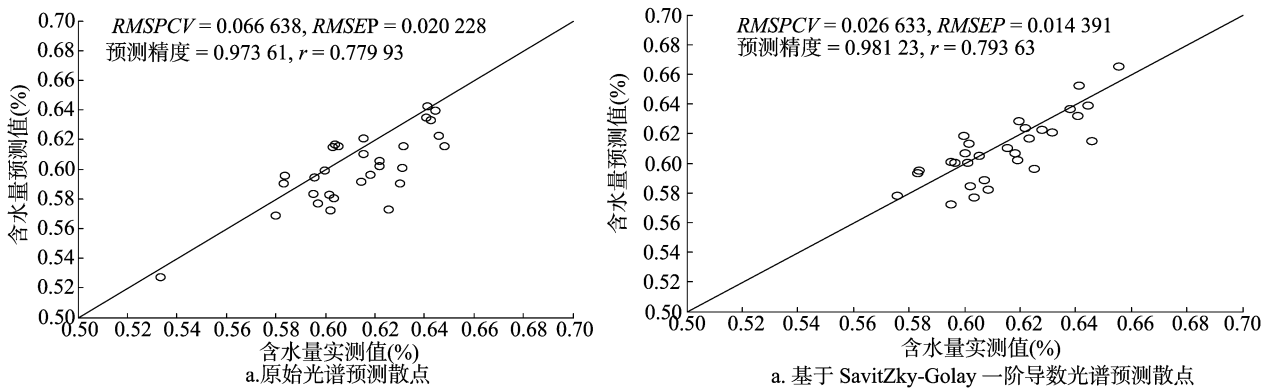


a.基于原始光谱进行波长选取　b. 基于 SavitZky-Golay 一阶导数预处理波长选取

图3　基于原始光谱和预处理光谱 GA 选择的特征波长

2.3　模型的建立和预测

2.3.1　SPA – PLS 模型建立与预测　通过 SPA 和 PLS 算法,分别对胡杨叶片原始光谱和一阶导数光谱进行建模,将 SPA 算法提取的特征波长,作为 PLS 的输入变量,叶片含水量作为输出变量。结果发现,基于一阶导数光谱与 SPA 算法提取的特征波长进行建模的精度、相关性均明显提高,交叉验证预测均方差(RMSPCV)由 0.666 38 降低到 0.026 633,预测均方根误差(RMSEP)由 0.020 228 降低到 0.014 391,预测精度由 0.973 61 提高到 0.981 23,相关系数(r)由 0.779 93 提高到 0.793 63(图 4)。试验结果表明,基于 SavitZky – Golay 一阶导数使用连续投影算法(SPA)能够有效地对光谱数据进行压缩,提取特征波长,消了散射影响,降低噪声干扰、提高建模精度。



a.原始光谱预测散点　a. 基于 SavitZky-Golay 一阶导数光谱预测散点

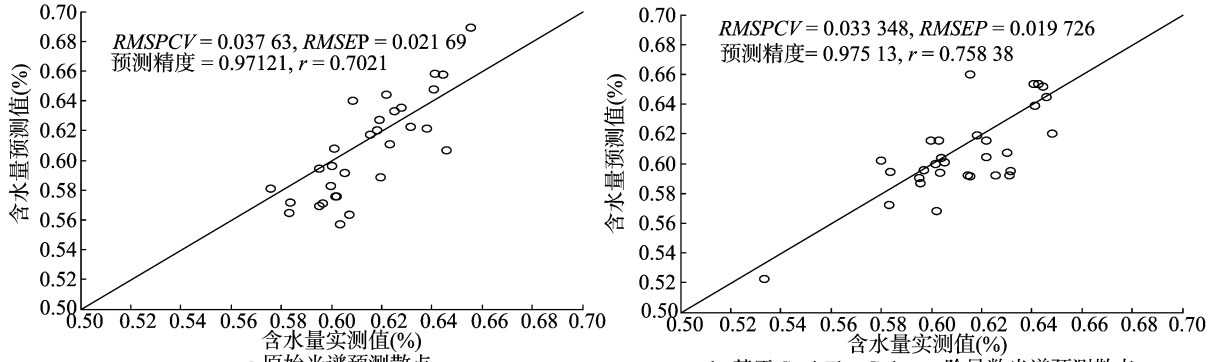
图4　基于原始光谱和预处理光谱的 SPA–PLS 预测模型

2.3.2　GA – PLS 模型建立与预测　通过 GA 和 PLS 算法,分别对胡杨叶片原始光谱和一阶导数光谱进行建模,在 PLS 方法交叉验证计算过程中,依变量负载值对变量排序,通过逐一计算误差值 RMSPCV,选取最小 RMSPCV 所对应的特征变量数即是最优拟合特征数。结果发现,基于一阶导数光谱与 GA 算法提取的特征波长进行建模的精度、相关性均明显提高,交叉验证预测均方差(RMSPCV)由 0.037 63 降低到 0.033 348,预测均方根误差(RMSEP)由 0.021 69 降低到 0.019 726,预测精度由 0.971 21 提高到 0.975 13,相关系数(r)由 0.702 1 提高到 0.758 38(图 5)。试验结果表明,基于 SavitZky – Golay 一阶导数使用遗传算法(GA)能够有效地对光谱数据进行压缩,提取特征波长,消了散射影响,降低噪声干扰、提高建模精度。

综合比较 SPA – PLS 和 GA – PLS 算法在同一预处理结果上的建模指数,SPA – PLS 总体要优于 GA – PLS。SPA – PLS 选择的变量只用了 18 个,而 GA – PLS 用了 29 个,并且评价指数均优于 GA – PLS,较少的波段能够提高运算速度,同时减少成本。因此,选择 SPA – PLS 算法为胡杨叶片含水量最佳预测模型。

3　结论

在胡杨叶片含水量近红外光谱监测中使用连续投影算法(SPA)与遗传算法(GA)进行有效波长的选取是可行的。对 SavitZky – Golay 一阶导数光谱数据使用 SPA 选取的有效波长基本上都分布在 1 360、1 650 nm 附近,并且所选波长与含水量有较好的相关性。利用SPA可以有效地降低光谱矩阵



a.原始光谱预测散点　　b. 基于 SavitZky-Golay 一阶导数光谱预测散点

图5　基于原始光谱和预处理光谱的 GA–PLS 预测模型

的维数。基于相同预处理光谱采用 SPA – PLS 的结果要优于 GA – PLS，预测精度达到了 0.981 23，相关系数达到了 0.793 63，为研制胡杨叶片水分便携式检测仪提供了理论依据。

参考文献：

[1]韩　路，王家强，王海珍，等. 塔里木河上游胡杨种群结构与动态[J]. 生态学报，2014，34(16)：4640 – 4651.

[2]王纪华，赵春江，郭晓维，等. 用光谱反射率诊断小麦叶片水分状况的研究[J]. 中国农业科学，2001，34(1)：104 – 107.

[3]Huan K，Liu X，Zheng F，et al. Variable selection of near – infrared spectroscopy for measuring wheat protein based on MC – LPG[J]. Transactions of the Chinese Society of Agricultural Engineering，2013，29(4)：266 – 271.

[4]Eisenstecken D，Panarese A，Robatscher P，et al. A near infrared spectroscopy (NIRS) and chemometric approach to improve apple fruit quality management：a case study on the cultivars "cripps pink" and "braeburn"[J]. Molecules，2015，20(8)：13603 – 13619.

[5]Olarewaju O O，Bertling I，Magwaza L S. Non – destructive evaluation of avocado fruit maturity using near infrared spectroscopy and PLS regression models[J]. Scientia Horticulturae，2016(199)：229 – 236.

[6]刘小军，田永超，姚　霞，等. 基于高光谱的水稻叶片含水量监测研究[J]. 中国农业科学，2012，45(3)：435 – 442.

[7]方美红，居为民. 基于叶片光学属性的作物叶片水分含量反演模型研究[J]. 光谱学与光谱分析，2015，35(1)：167 – 171.

[8]刘明博，唐延林，李晓利，等. 水稻叶片氮含量光谱监测中使用连续投影算法的可行性[J]. 红外与激光工程，2014，43(4)：1265 – 1271.

[9]Li L，Cheng Y B，Ustin S，et al. Retrieval of vegetation equivalent water thickness from reflectance using genetic algorithm (GA) – partial least squares (PLS) regression[J]. Advances in Space Research，2008，41(11)：1755 – 1763.

[10]王加华，潘　璐，孙　谦，等. 遗传算法结合偏最小二乘法无损评价西洋梨糖度[J]. 光谱学与光谱分析，2009，29(3)：678 – 681.

[11]吴静珠，李　慧，王克栋，等. 光谱预处理在农产品近红外模型优化中的应用研究[J]. 农机化研究，2011，33(3)：178 – 181.

[12]邓小蕾，李民赞，郑立华，等. 基于反射光谱预处理的苹果叶片叶绿素含量预测[J]. 农业工程学报，2014，30(14)：140 – 147.

[13]王伟明，董大明，郑文刚，等. 梨果糖浓度近红外漫反射光谱检测的预处理方法研究[J]. 光谱学与光谱分析，2013，33(2)：359 – 362.

[14]高洪智，卢启鹏，丁海泉，等. 基于连续投影算法的土壤总氮近红外特征波长的选取[J]. 光谱学与光谱分析，2009，29(11)：2951 – 2954.

[15]Qu F，Ren D，Wang J，et al. An ensemble successive project algorithm for liquor detection using near infrared sensor[J]. Sensors，2016，16(1)：89.

[16]Galvao R K，Ugulino Araujo M C，Silva E C，et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm[J]. Chemometrics and Intelligent Laboratory Systems，2008，92(1)：83 – 91.

[17]屠振华，籍保平，孟超英，等. 基于遗传算法和间隔偏最小二乘的苹果硬度特征波长分析研究[J]. 光谱学与光谱分析，2009，29(10)：2760 – 2764.

[18]Song K，Li L，Tedesco L P，et al. Hyperspectral determination of eutrophication for a water supply source via genetic algorithm – partial least squares (GA – PLS) modeling[J]. Science of the Total Environment，2012，426(2)：220 – 232.

[19]杭艳红，杨　林. 基于 GA – BP 神经网络的耕地自然质量计算模型研究[J]. 江苏农业科学，2017，45(8)：183 – 186.

[20]张　兵，范泽华，姚江河，等. 基于近红外光谱与多元模型的小麦氮含量估算[J]. 江苏农业科学，2016，44(9)：374 – 378.

[21]吴　迪，金春华，何　勇. 基于连续投影算法的光谱主成分组合优化方法研究[J]. 光谱学与光谱分析，2009，29(10)：2734 – 2737.

[22]成　忠，张立庆，刘赫扬，等. 连续投影算法及其在小麦近红外光谱波长选择中的应用[J]. 光谱学与光谱分析，2010，30(4)：949 – 952.

[23]邹小波，赵杰文. 用遗传算法快速提取近红外光谱特征区域和特征波长[J]. 光学学报，2007，27(7)：1316 – 1321.

[24]潘　璐，王加华，李鹏飞，等. 砂梨糖度近红外光谱波段遗传算法优化[J]. 光谱学与光谱分析，2009，29(5)：1246 – 1250.