# 玉米秸秆纤维素和半纤维素NIRS特征波长优选

刘金明[1,2]，初晓冬[1]，王　智[1]，许永花[3]，李文哲[1]，孙　勇[1*]

1. 东北农业大学工程学院，黑龙江 哈尔滨　150030
2. 黑龙江八一农垦大学电气与信息学院，黑龙江 大庆　163319
3. 东北农业大学电气与信息学院，黑龙江 哈尔滨　150030

**摘　要**　预处理是提高玉米秸秆生物转化利用效率的有效途径。玉米秸秆经生物炼制转化为生物燃料时，转化率与其原料内的纤维素和半纤维素含量直接相关。为了实现对预处理后玉米秸秆的生物炼制过程的有效调控，提出使用近红外光谱(NIRS)对玉米秸秆的纤维素和半纤维素含量进行快速检测，解决传统化学方法测试速度慢、成本高的问题。为了提高NIRS检测的效率和精度，将遗传算法与模拟退火算法相结合构建遗传模拟退火算法(GSA)用于预处理后玉米秸秆纤维素和半纤维素含量NIRS特征波长优选。GSA算法以NIRS波长点数为码长进行二进制编码，以偏最小二乘法(PLS)回归模型的交叉验证均方根误差为目标函数，结合温度参数设计适应度函数，基于Metropolis判别准则实现扰动解的选择复制，能够在避免早熟的同时有效提高进化后期的搜索效率。采用碱预处理、生物预处理及其相结合的方法对采集的玉米秸秆进行预处理后制备样品120个，并测定其纤维素和半纤维素含量及NIRS。使用7点Savitzky-Golay平滑结合多元散射校正和标准正则变换对光谱进行预处理后，利用Kennard-Stone法按3∶1比例划分校正集和验证集。然后，使用GSA算法对NIRS全谱进行特征波长优选(记为Full-GSA)、对协同区间偏最小二乘法(SiPLS)优选后谱区进行特征波长优选(记为SiPLS-GSA)、对反向区间偏最小二乘法(BiPLS)优选后谱区进行特征波长优选(记为BiPLS-GSA)，并使用PLS回归模型和验证集对特征波长优选结果进行评测。Full-GSA以全谱1 557个波长点为基因，执行16次算法，优选出118个纤维素特征波长点和164个半纤维素特征波长点。SiPLS-GSA经SiPLS优选的纤维素和半纤维素谱区波长点数分别为388个和160个，再经GSA进一步优选后得到157个纤维素特征波长点和148个半纤维素特征波长点。BiPLS-GSA经BiPLS优选的纤维素和半纤维素谱区波长点数分别为358个和180个，再经GSA进一步优选后得到130个纤维素特征波长点和153个半纤维素特征波长点。结果表明，通过波长优选，不仅参与建模的波长点数量显著减少，而且回归模型的性能显著优于全谱建模。其中，采用Full-GSA优选的纤维素特征光谱回归性能最佳，采用SiPLS-GSA优选的半纤维素特征光谱回归性能最佳。回归模型验证集的平均相对误差(MRE)分别为1.752 4%和2.020 8%，较全谱建模分别降低了13.636 6%和25.368 4%。基于结合温度参数设计适应度函数的策略构建的GSA具有良好的全局搜索性能，适用于玉米秸秆纤维素和半纤维素含量NIRS特征波长优选。GSA以全谱每个波长点为染色体基因的编码方案适用于NIRS全谱的特征波长优选。GSA同样适用于SiPLS和BiPLS优选后谱区的特征波长优选，能够有效实现优选后谱区的波长点优选。

**关键词**　玉米秸秆；近红外光谱；遗传模拟退火算法；协同区间偏最小二乘法；反向区间偏最小二乘法；特征波长

**中图分类号**：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2019)03-0743-08

**收稿日期**：2018-03-10，**修订日期**：2018-07-28

**基金项目**：国家科技支撑计划课题(2015BAD21B03)，哈尔滨市科技创新人才专项(2016RAXXJ009)，黑龙江省青年科学基金项目(QC2016033)，黑龙江八一农垦大学校内培育课题(XZR2017-09)资助

**作者简介**：刘金明，1981年生，东北农业大学工程学院博士研究生，黑龙江八一农垦大学电气与信息学院副教授
e-mail：jinmingliu2008@126.com　　*通讯联系人　　e-mail：sunyong740731@163.com

## 引 言

玉米秸秆是我国主要的农作物秸秆之一，是一种亟待处理和利用的可再生资源。厌氧发酵产沼气和酶解糖化产乙醇等生物转化技术是将作物秸秆变废为宝的有效手段[1]。但玉米秸秆因其自身紧密的木质纤维素结构而具有强大的耐酶解特性，导致其生物转化效率偏低[2]。玉米秸秆主要由纤维素、半纤维素和木质素三种木质纤维素成分组成。通过预处理打破这三种物质在细胞壁中紧密结合在一起形成的坚硬、稳定的木质纤维素结构，是提高玉米秸秆生物转化利用率的有效途径[3]。玉米秸秆经生物炼制转化为生物燃料时，转化率与其原料内的纤维素和半纤维素两种碳水化合物含量直接相关[4]。因此，为了对预处理后玉米秸秆的生物炼制过程进行有效调控，有必要对秸秆中纤维素和半纤维素含量进行快速、准确测定[5]，但采用传统化学方法测定其含量时存在测试速度慢、成本高的问题[6]。

近红外光谱(near infrared spectroscopy，NIRS)分析技术具有简便、快速、无损、低成本等众多优点[7]，已广泛用于农产品及农牧业废弃物的定性分析[8]和定量检测[9]。近年来，相关学者开始研究应用 NIRS 对植物细胞壁中的纤维素和半纤维素含量进行测定[10]。Niu 等[11]提出使用 NIRS 技术对不同成熟期的小麦秸秆进行纤维素和半纤维素等碳水化合物的测定。Xue 等[5]提出了一种 NIRS 在线检测方法用于玉米秸秆木质纤维素成分的实时检测。上述研究主要以采集的秸秆为研究对象，尚未见专门以预处理后的玉米秸秆为研究对象，使用 NIRS 对其进行纤维素和半纤维素测定。

随着近红外光谱仪采集精度的提高，采集的数据量越来越大。以全谱波长点建模时，计算量大，波长冗余严重。通过特征波长优选，可以有效消除光谱中的不相干和非线性波长点对模型预测精度的影响。相关学者提出应用遗传算法(genetic algorithm，GA)[12]、粒子群优化算法[13]、蚁群算法[14]等智能优化算法进行 NIRS 变量组合优化，从而筛选出有效的波长变量。其中 GA 具有较强的鲁棒性和全局搜索能力，其随机搜索特性能够有效解决光谱波长点之间的共线性问题[12]，还可以与其他光谱谱区优化算法相结合进行特征波长点的优选[12, 14]。但 GA 存在早熟问题，且进化后期搜索效率较低。

因此，将 GA 与模拟退火算法(simulated annealing algorithm，SA)相结合构建遗传模拟退火算法(genetic simulated annealing algorithm，GSA)，通过结合温度参数设计适应度函数，在避免算法早熟的同时有效提高算法进化后期的搜索效率。然后，以偏最小二乘(partial least squares，PLS)回归模型的交叉验证均方根误差(root mean square error of cross-validation，RMSECV)为目标函数，使用构建的 GSA 算法对预处理后玉米秸秆纤维素和半纤维素含量 NIRS 的全谱及协同区间偏最小二乘法(synergy interval partial least squares，SiPLS)和反向区间偏最小二乘(backward interval partial least squares，BiPLS)优选后的谱区进行特征波长优选，从而有效提高纤维素和半纤维素含量 NIRS 回归模型的性能。

## 1 实验部分

### 1.1 样品制备

实验用玉米秸秆取自东北农业大学向阳农场，自然风干后粉碎成 10 mm 的秸秆段备用。依据秸秆在生物炼制过程中碱性预处理[15]效果好、工艺简单，以及生物预处理方法在环境友好性方面的优势，采用地衣芽孢杆菌(生物方法)、NaOH 溶液(碱性试剂)、猪粪沼液(富含微生物的弱碱性溶液)、沼液加 NaOH(富含微生物的强碱性溶液)共四种方法对玉米秸秆进行了预处理，并按不同处理试剂浓度、不同处理时间采样，共计采样 120 个。将预处理后玉米秸秆样品进行烘干、粉碎，过 40 目筛后装袋备用。

### 1.2 纤维素和半纤维素含量测定

纤维素和半纤维素的测定按照 Van Soest 法的原理，采用 Ankom 200i 半自动纤维分析仪依次对秸秆粉末进行中性洗涤纤维、酸性洗涤纤维和酸性洗涤木质素的测定，通过计算得到样品的纤维素含量(酸性洗涤纤维含量－酸性洗涤木质素含量)和半纤维素含量(中性洗涤纤维含量－酸性洗涤纤维含量)。每个样品测试 3 次，取 3 次的平均值作为待测含量值。

### 1.3 光谱数据采集

对预处理后的玉米秸秆样品粉末使用 Nicolet 公司的 Antaris Ⅱ型傅里叶近红外光谱仪进行积分球漫反射光谱扫描，光谱采集范围 10 000～4 000 $cm^{-1}$，分辨率为 8.0 $cm^{-1}$，样品扫描 32 次，装样方式为带透明塑料袋扫描，扫描时去除透明塑料袋背景，三次扫描取平均值作为样品的原始光谱。原始光谱的波长点为 1 557 个，数据点间距为 3.86 $cm^{-1}$，起始波长点为 10 001.03 $cm^{-1}$，结束波长点为 3 999.64 $cm^{-1}$。

## 2 算法构建及模型评价

### 2.1 GSA 算法构建

#### 2.1.1 算法初始化

算法的初始化包括编码、种群初始化、初温设定、降温操作、进化代数等。编码方式采用二进制编码，码长为待优选光谱波长点个数。“1”和“0”分别表示该波长点对应的数据“是”、“否”选中参与运算。种群初式化时随机产生一个 $M\times L$ 的二元矩阵，其中 $M$ 为种群规模，$L$ 为码长。初温的确定依据 $t_0=K(f_{0\_max}-f_{0\_min})$，其中 $K$ 是一个正整数，$f_{0\_max}$ 和 $f_{0\_min}$ 为初始种群中的最大和最小目标函数值。退温操作依据 $t_{n+1}=\alpha t_n$，其中 $0<\alpha<1$。

#### 2.1.2 适应度函数设计

适应度函数对算法的进化方向起指导作用，其设计的是否合理直接决定着算法的性能。选择 PLS 回归模型的 RMSECV 作为目标函数，结合温度参数设计适应度函数如下

$$fit(x)=\exp\left(-\frac{f(x)-f_{\min}}{t}\right)$$

式中，$f(x)$为当前染色体的目标函数值，$f_{\min}$为当前代种群中的最小目标函数值，$t$ 为当前代温度值。

采用此适应度函数设计方法，使得算法在初始阶段计算的适应度值差异较小，能够有效避免个别优良染色体充斥整个种群导致算法收敛到局部最优解；在进化后期优良染色体具有相对更大的适应度函数值，更容易遗传给下一代，进而加快算法的搜索速度。

2.1.3　进化过程设计

算法的进化过程分为 GA 的选择、交叉、变异操作和 SA 的 Metropolis 选择复制两部分。GA 的选择操作采用带最有保留策略的赌轮选择，交叉操作采用离散重组交叉，变异操作采用离散变异策略。SA 的 Metropolis 选择复制包括邻域解的构建和状态接受函数两部分。邻域解的构建采用多位变异策略，即在当前染色体中随机选择 $m$ 位进行位变异。状态接受函数基于 Metropolis 判别准则实现。

### 2.2　模型建立及评价

在进行 NIRS 波长优选前，先使用 Savitzky-Golay 平滑、一阶导数、二阶导数、多元散射校正(multivariate scattering correction，MSC)、标准正则变换(standard normal variate，SNV)及其组合方法进行光谱预处理，再使用 Kennard-Stone 法将预处理后光谱划分为 90 个校正集样本和 30 个验证集样本，然后计算全谱下的 PLS 回归模型性能，并基于预测均方根误差(root mean square error of prediction，RMSEP)来确定光谱预处理方法。经计算比较后最终确定纤维素和半纤维素含量预测模型的光谱预处理方法为 7 点 Savitzky-Golay 平滑、MSC 和 SNV。在选定预处理方法后，使用 GSA 算法分别对全谱(记为 Full-GSA)及 SiPLS 优选后谱区(记为 SiPLS-GA)和 BiPLS 优选后谱区(记为 BiPLS-GSA)进行特征波长优选。最后，分别以三种优选后的特征波长为输入，使用 PLS 建立秸秆纤维素和半纤维素含量定量分析模型，并采用校正决定系数($R_c^2$)、验证决定系数($R_p^2$)、校正均方根误差(root mean square error of calibration，RMSEC)、RMSEP、验证集的平均相对误差(mean relative error，MRE)和相对分析误差(residual predictive deviation，RPD)来评价预测模型的优劣。

算法全部在 Matlab R2012b 软件平台中实现，其中 BiPLS 和 SiPLS 基于 itoolbox 工具箱实现。

## 3　结果与讨论

### 3.1　采集数据分析

对 120 个样品的原始光谱经 Savitzky-Golay 平滑、MSC 和 SNV 预处理后，使用 Kennard-Stone 法按 3∶1 的比例划分样本，得到校正集样本 90 个、验证集样本 30 个，对应的纤维素和半纤维素含量如表 1 所示。

对预处理后的近红外光谱进行主成分分析，第一、第二和第三主成分的贡献率分别为 69.73%、13.87%和 7.38%，前三个主成分的累积贡献率达 90.98%。校正集和验证集的三维主成分空间分布情况如图 1 所示。

由表 1 和图 1 可知，校正集和验证集样本的分布比较均匀，校正集纤维素和半纤维素含量基本涵盖了验证集，可以使用该样本划分方法进行 NIRS 分析。

**表 1　纤维素和半纤维素含量**

**Table 1　Contents of cellulose and hemicellulose**

| 样本 | 成分 | 平均值/% | 最大值/% | 最小值/% | 标准偏差 SD/% |
|---|---|---|---|---|---|
| 校正集 | 纤维素 | 42.935 9 | 51.526 8 | 36.227 0 | 3.310 3 |
| | 半纤维素 | 20.033 2 | 33.252 9 | 9.484 2 | 8.256 8 |
| 验证集 | 纤维素 | 43.389 7 | 50.982 2 | 36.477 3 | 2.891 5 |
| | 半纤维素 | 17.227 3 | 33.801 3 | 9.512 2 | 7.233 5 |

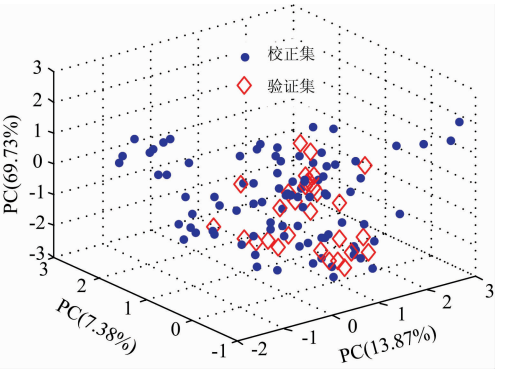


**图 1　样本主成分空间分布**

**Fig. 1　Distribution of the samples in principal components space**

### 3.2　NIRS 特征波长优选

3.2.1　Full-GSA 特征波长优选

Full-GSA 以全谱 1 557 个波长变量为基因，随机生成 300 个码长为 1557 的染色体构建初始种群；综合考虑搜索效率和性能，选取初温确定系数 K 取 100，退温系数取 0.8，进化代数取 50，交叉概率取 0.7，变异概率取 0.1，邻域解扰动位数 m 取 50。为消除 GSA 算法的随机性，分别执行算法 10 次(记为 Full-GSA-10)、12 次(记为 Full-GSA-12)和 16 次(记为 Full-GSA-16)对纤维素和半纤维素含量特征波长进行优选。多次执行时，每次都选中的波长点代表了染色体的优良基因，以这些特征波长点作为特征波长建立回归模型时，可以有效消除 GSA 算法的随机性，且能够得到较高的回归模型性能。测试发现，校正集和验证集回归性能参数 $R_c^2$ 和 $R_p^2$ 随选中次数的增加呈逐先升高后降低的趋势，但验证集要早于校正集降低。原因在于 GSA 以 RMSECV 为依据进行特征波长优选，验证集回归性能拐点出现时表明校正集发生了过拟合。

进行纤维素特征波长 GSA 优选时，测试发现，算法执行 10 次时，选中 9 次以上的波长点(73 个)回归性能最佳；执行 12 次时，选中 10 次以上的波长点(101 个)回归性能最佳；执行 16 次时，选中 13 次以上的波长点(118 个)回归性能最佳，其优选结果与预处理后的平均光谱对比如图 2 所示。

由图 2 可知，筛选出的特征波长中多数波长点在样本 NIRS 吸收峰附近，能真实的反应纤维素对应的 C—C，C—O，C—H，—OH 和—$CH_2$ 等官能团。选中 15 次以上的波长点(37 个)中，258(9 009.80 $cm^{-1}$)，400(8 462.11 $cm^{-1}$)，

406(8 438.97 $cm^{-1}$)，470(8 192.13 $cm^{-1}$)，630(7 575.02 $cm^{-1}$)，665(7 440.02 $cm^{-1}$)，667(7 432.31 $cm^{-1}$)，785(6 977.19 $cm^{-1}$)和 822(6 834.49 $cm^{-1}$)对应 C—H，$—CH_2$ 和—OH 的二级倍频，1 013(6 097.81 $cm^{-1}$)，1 037(6 005.25 $cm^{-1}$)，1 038(6 001.39 $cm^{-1}$)，1 048(5 962.82 $cm^{-1}$)，1 089(5 804.69 $cm^{-1}$)，1 090(5 800.83 $cm^{-1}$)，1 118(5 692.83 $cm^{-1}$)，1 179(5 457.56 $cm^{-1}$)，1 224(5 284.00 $cm^{-1}$)，1 269(5 110.44 $cm^{-1}$)，1 280(5 068.01 $cm^{-1}$)，1 283(5 056.44 $cm^{-1}$)和 1 288(5 037.16 $cm^{-1}$)对应着 C—H，$—CH_2$，C—O 和—OH 的一级倍频，1 374(4 705.46 $cm^{-1}$)，1 440(4 450.90 $cm^{-1}$)，1 468(4 342.91 $cm^{-1}$)和 1 486(4 273.48 $cm^{-1}$)对应 C—C，C—H，$—CH_2$ 和 C—O的组合频。

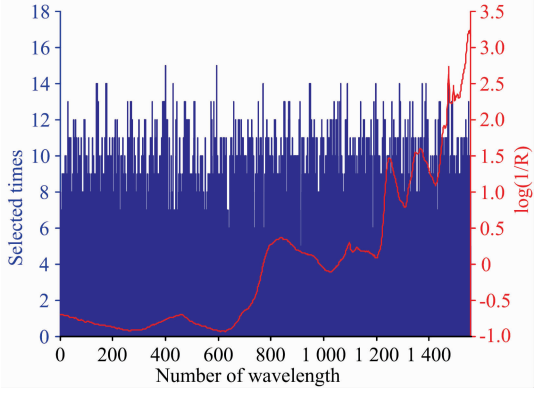


**图 2 纤维素 Full-GSA 波长优选结果**

**Fig. 2 Results of wavelength optimization for cellulose by Full-GSA**

在进行半纤维素特征波长优选时，执行 Full-GSA 算法 10 次，选中 8 次以上的波长点(83 个)回归模型性能最佳；执行 12 次时，选中 9 次以上的波长点(119 个)回归性能最佳；执行 16 次时，选中 11 次以上的波长点(164 个)回归性能最佳，其优选结果如图 3 所示。

由图 3 可知，筛选出的多数特征波长点在样本的 NIRS 吸收峰附近，能真实的反应半纤维素对应的 C—C，C—O，C—H，C═O，—OH，—CH，$—CH_2$，—CHO 和$—NH_2$ 等官能团。

3.2.2 SiPLS-GSA 特征波长优选

SiPLS-GSA 先使用 SiPLS 将全谱划分成多个均匀的子区间，再组合不同个数的子区间进行建模，选择模型 RMSECV 最小的组合区间作为 SiPLS 优选后谱区，然后再使用 GSA 进行特征波长点优选。为考察分割波长点个数对波长选择及模型预测性能的影响，分别按约 30，40，50，60，80，100，120 个波长点划分子区间，依次将光谱划分为 52，39，31，26，20，16 和 13 个子区间，并依据 RMSECV 选取 2～4 个子区间构建的组合区间作为 SiPLS 优选的特征谱区。不同子区间个数下优选的纤维素和半纤维素特征谱区如表 2 所示。

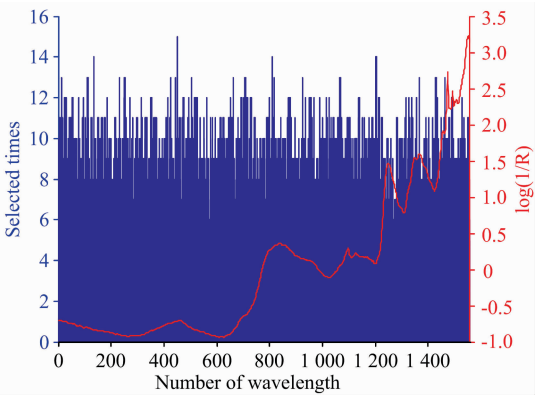


**图 3 半纤维素 Full-GSA 波长优选结果**

**Fig. 3 Results of wavelength optimization for hemicellulose by Full-GSA**

**表 2 纤维素和半纤维维素 SiPLS 优选谱区结果**

**Table 2 Results of optimized spectral regions for cellulose and hemicellulose by SiPLS**

| 成分 | 划分区间数 | 主成分数 | 最佳子区间编号 | 波长点数 | RMSECV |
|---|---|---|---|---|---|
| 纤维素 | 13 | 5 | [8 9 10 13] | 479 | 1.361 0 |
| 纤维素 | 16 | 10 | [11 12 13 15] | 388 | 1.228 0 |
| 纤维素 | 20 | 10 | [12 14 16 19] | 311 | 1.299 0 |
| 纤维素 | 26 | 10 | [7 18 19 13] | 241 | 1.311 0 |
| 纤维素 | 31 | 9 | [8 9 23] | 150 | 1.316 0 |
| 纤维素 | 39 | 9 | [10 11 28] | 120 | 1.233 0 |
| 纤维素 | 52 | 9 | [13 14 16 28] | 120 | 1.246 0 |
| 半纤维素 | 13 | 10 | [2 5 7 12] | 479 | 1.062 0 |
| 半纤维素 | 16 | 10 | [6 9 12 14] | 388 | 1.079 0 |
| 半纤维素 | 20 | 10 | [6 11 13 18] | 311 | 1.048 0 |
| 半纤维素 | 26 | 10 | [8 14 15 22] | 240 | 1.045 0 |
| 半纤维素 | 31 | 10 | [14 17 22 27] | 200 | 1.066 0 |
| 半纤维素 | 39 | 10 | [17 21 27 33] | 160 | 1.032 0 |
| 半纤维素 | 52 | 10 | [24 28 36 44] | 120 | 1.044 0 |

依据 RMSECV 选取划分 16 个子区间的最佳组合区间[11 12 13 15]作为 SiPLS 优选后的纤维素特征谱区。SiPLS-GSA 将该组合区间对应的 976～1 266 和 1 364～1 460 波长点(388 个)作为 GSA 的输入波长进行再次寻优，GSA 算法参数设置如下：码长为 388，种群规模为 80，退温系数取 0.9，进化代数取 100，邻域解扰动位数 $m$ 取 15，其他参数与 Full-GSA 一致。算法连续执行 10 次后，经计算后确定选中 6 次以上的波长点(157 个)作为 SiPLS-GSA 优选的纤维素特征波长。

选取划分 39 个子区间的组合区间[17 21 27 33]作为 SiPLS 优选后的半纤维素特征谱区(波长点为 641～680，801～840，1 041～1 080 和 1 281～1 320，共计 160 个)，并使用 GSA 对其进行波长点优选，算法码长为 160，种群规模为 50，邻域解扰动位数 $m$ 取 10，其他参数与纤维素波长优选时一致。算法连续执行 10 次后，经计算后确定选中 3 次以上的波长点(148 个)为 SiPLS-GSA 优选的半纤维素特征波长。

3.2.3　BiPLS-GSA 特征波长优选

BiPLS-GSA 采用与 SiPLS-GSA 相同的区间划分方案，先将光谱划分为 13，16，20，26，31，39 和 52 个子区间，再使用 BiPLS 选取每种子区间数下 RMSECV 最小的组合区间作为 BiPLS 优选后谱区，然后再使用 GSA 对优选后的谱区进行特征波长点优选。不同子区间数下优选的纤维素和半纤维素特征谱区如表 3 所示。

**表 3　纤维素和半纤维素 BiPLS 优选谱区结果**

**Table 3　Results of optimized spectral regions for cellulose and hemicellulose by BiPLS**

| 成分 | 划分区间数 | 主成分数 | 最佳子区间编号 | 波长点数 | RMSECV |
|---|---|---|---|---|---|
| 纤维素 | 13 | 8 | [4 7 8 9 10 12 13] | 838 | 1.343 8 |
| 纤维素 | 16 | 8 | [6 11 12 15] | 388 | 1.322 7 |
| 纤维素 | 20 | 10 | [12 14 16 19] | 311 | 1.298 9 |
| 纤维素 | 26 | 7 | [16 18 19 20 24 25] | 358 | 1.282 5 |
| 纤维素 | 31 | 8 | [15 18 21 22 23 28 29] | 350 | 1.291 8 |
| 纤维素 | 39 | 8 | [2 10 14 24 27 28 30 35 37] | 359 | 1.276 3 |
| 纤维素 | 52 | 8 | [18 23 30 35 37 40 47 49 50] | 269 | 1.288 5 |
| 半纤维素 | 13 | 10 | [3 5 7 8 9 12] | 719 | 1.060 5 |
| 半纤维素 | 16 | 9 | [5 6 9 14 16] | 486 | 1.087 2 |
| 半纤维素 | 20 | 10 | [1 3 7 8 10 11 14 17 20] | 701 | 1.028 0 |
| 半纤维素 | 26 | 10 | [7 11 12 14 15 18 22 23 24] | 539 | 1.030 6 |
| 半纤维素 | 31 | 10 | [5 10 16 17 22 26 27] | 351 | 1.029 6 |
| 半纤维素 | 39 | 9 | [1 13 15 18 20 21 32 34 36] | 360 | 1.038 8 |
| 半纤维素 | 52 | 10 | [17 20 24 28 36 44] | 180 | 0.983 9 |

选取划分 26 个子区间的最佳组合区间[16 18 19 20 24 25]作为 BiPLS 优选后的纤维素特征谱区。对应的波长点为 901～960，1 021～1 200 和 1 381～1 498，共计 358 个。分析发现，SiPLS 和 BiPLS 优选的纤维素特征谱区存在 260 个波长点的重复，占 BiPLS 优选谱区的 72.63%，说明 SiPLS 和 BiPLS 谱区优选具有一致性。将这 358 个波长点作为 GSA 的输入波长进行再次优选，算法参数设置如下：码长为 358，种群规模为 70，邻域解扰动位数 $m$ 取 10，其他参数与 SiPLS-GSA 一致。连续执行算法 10 次后，经计算确定选中 6 次以上的波长点(130 个)作为 BiPLS-GSA 优选的纤维素特征波长。

选取划分 52 个子区间的最佳组合区间[17 20 24 28 36 44]作为 BiPLS 优选后的半纤维素特征谱区，对应波长点为 481～510，571～600，691～720，811～840，1 051～1 080 和 1 291～1 320，共计 180 个。BiPLS 和 SiPLS 优选的半纤维素特征谱区同样具有一致性。使用 GSA 对 BiPLS 优选后谱区进行波长点优选时，算法码长为 160，种群规模为 50，其他参数与纤维素波长优选时一致。算法连续执行 10 次后，计算确定选中 3 次以上的波长点(153 个)作为 BiPLS-GSA 优选的半纤维素特征波长。

### 3.3　特征波长优选结果评价与分析

以 Full-GSA，SiPLS-GSA 和 BiPLS-GSA 优选后的特征波长点作为 PLS 回归模型的输入，建立秸秆纤维素和半纤维素含量定量分析模型，并与 GA 算法对全谱(记为 Full-GA)、SiPLS 优选后谱区(记为 SiPLS-GA)和 BiPLS 优选后谱区(记为 BiPLS-GA)进行特征波长优选的结果进行对比，其结果如表 4 所示。

由表 4 可知，各种波长优选方法下，GSA 优选结果的预测性能都高于 GA；GSA 优选的纤维素回归模型的 $R_c^2$ 和 $R_p^2$ 都大于 0.83，RPD 都大于 2.5，说明建模基本成功；GSA 优选的半纤维素回归模型的 $R_c^2$ 和 $R_p^2$ 都大于 0.97，RPD 都大于 6，说明模型非常成功。半纤维素回归模型的预测效果高于纤维素回归模型，这与文献[6]中的研究结果一致。但半纤维素含量值约为纤维素含量的一半，导致半纤维素回归模型的 MRE 值略高于纤维素。各种波长优选方法下，半纤维素 RPD 值都是纤维素的 2 倍多。半纤维素 RPD(RPD＝SD/RMSEP)值较大的原因在于预处理过程中半纤维素容易被破坏，导致样本中半纤维素含量值变化较大，进而 SD 值较大(见表 1)，而半纤维素与纤维素的 RMSEP 值相差不大。

由 SiPLS 和 BiPLS 回归模型的评价参数可知：SiPLS 和 BiPLS 依据 RMSECV 选择多个组合区间作为优选谱区，在一定程度上体现了纤维素和半纤维素特征波长点的分布特性，两种方法的回归模型性能都高于 PLS，其中半纤维素采用 SiPLS 优选的谱区回归性能最佳，而纤维素采用 SiPLS 和 BiPLS 优选的谱区回归性能相差不大。

由 Full-GSA，SiPLS-GSA 和 BiPLS-GSA 优选后的特征波长回归模型评价参数可知，通过 GSA 可以在现有波长优选方法的基础上进行有效的再次优化，去掉无关波长点，提高回归模型的性能。在使用 GSA 对全谱及 SiPLS 和 BiPLS 优选后谱区进行特征波长点优选时，设置算法参数非常重要。经反复评测发现，Full-GSA 中全谱的码长为 1 557，种群规模设置为码长的 1/5 时，能够有效兼顾搜索速度和寻优精度，整体性能较好。因种群规模较大，为节省运行时间，可适当减小遗传代数和降温系数。而在执行 SiPLS-GSA 和 BiPLS-GSA 时，因码长较小，种群规模也较少，可以通过增大遗传代数和降温系数的方式来提高算法的进化次数、减缓

**表 4 不同波长优选方法评价指标**

**Table 4 Evaluation indicators of wavelength optimization for different methods**

| 成分 | 方法 | 波长点数 | $R_c^2$ | $R_p^2$ | RMSEC | RMSEP | MRE/% | RPD |
|---|---|---|---|---|---|---|---|---|
| 纤维素 | PLS | 1 557 | 0.868 8 | 0.831 4 | 1.127 4 | 1.206 3 | 2.029 1 | 2.379 1 |
| 纤维素 | Full-GSA-10 | 73 | 0.877 3 | 0.854 0 | 1.094 4 | 1.111 4 | 1.798 1 | 2.601 7 |
| 纤维素 | Full-GSA-12 | 101 | 0.882 0 | 0.857 5 | 1.075 4 | 1.104 4 | 1.762 2 | 2.618 3 |
| 纤维素 | Full-GA-16 | 48 | 0.884 0 | 0.861 7 | 1.067 2 | 1.107 4 | 1.833 1 | 2.611 1 |
| 纤维素 | Full-GSA-16 | 118 | 0.882 2 | 0.858 9 | 1.074 6 | 1.085 5 | 1.752 4 | 2.663 7 |
| 纤维素 | SiPLS | 388 | 0.877 1 | 0.842 9 | 1.094 9 | 1.120 4 | 1.849 0 | 2.580 8 |
| 纤维素 | SiPLS-GA | 220 | 0.892 4 | 0.850 9 | 1.031 5 | 1.104 0 | 1.829 5 | 2.619 3 |
| 纤维素 | SiPLS-GSA | 157 | 0.898 9 | 0.853 3 | 1.003 1 | 1.103 8 | 1.821 6 | 2.619 7 |
| 纤维素 | BiPLS | 358 | 0.883 8 | 0.830 2 | 1.066 3 | 1.146 5 | 1.908 9 | 2.522 2 |
| 纤维素 | BiPLS-GA | 204 | 0.891 0 | 0.823 5 | 1.037 9 | 1.153 0 | 1.867 8 | 2.507 8 |
| 纤维素 | BiPLS-GSA | 130 | 0.892 1 | 0.830 9 | 1.033 0 | 1.144 2 | 1.845 0 | 2.527 2 |
| 半纤维素 | PLS | 1 557 | 0.982 8 | 0.972 4 | 1.075 0 | 1.171 3 | 2.707 7 | 6.175 8 |
| 半纤维素 | Full-GSA-10 | 83 | 0.983 8 | 0.976 5 | 1.041 3 | 1.083 7 | 2.628 7 | 6.674 8 |
| 半纤维素 | Full-GSA-12 | 119 | 0.985 0 | 0.977 6 | 1.003 4 | 1.066 3 | 2.512 8 | 6.783 8 |
| 半纤维素 | Full-GA-16 | 68 | 0.983 6 | 0.979 8 | 1.047 6 | 1.007 9 | 2.553 5 | 7.176 6 |
| 半纤维素 | Full-GSA-16 | 164 | 0.985 5 | 0.982 8 | 0.988 6 | 0.920 9 | 2.263 5 | 7.854 8 |
| 半纤维素 | SiPLS | 160 | 0.987 6 | 0.984 6 | 0.914 6 | 0.862 8 | 2.035 0 | 8.384 0 |
| 半纤维素 | SiPLS-GA | 147 | 0.988 0 | 0.984 6 | 0.900 8 | 0.863 7 | 2.100 2 | 8.375 3 |
| 半纤维素 | SiPLS-GSA | 148 | 0.987 5 | 0.985 0 | 0.918 5 | 0.852 1 | 2.020 8 | 8.488 6 |
| 半纤维素 | BiPLS | 180 | 0.985 4 | 0.978 8 | 0.991 6 | 1.005 8 | 2.486 9 | 7.191 6 |
| 半纤维素 | BiPLS-GA | 151 | 0.987 0 | 0.982 5 | 0.935 2 | 0.961 4 | 2.331 1 | 7.524 0 |
| 半纤维素 | BiPLS-GSA | 153 | 0.986 7 | 0.983 1 | 0.945 1 | 0.944 6 | 2.284 6 | 7.657 5 |

注：Full-GA-16 表示执行 16 次 Full-GA 算法

Note: Full-GA-16 represents that the Full-GA algorithm is executed 16 times

适应度函数值的变化速度，进而有效提高算法的寻优性能。

Full-GSA 能够充分发挥 GSA 算法全局寻优的特性，可以通过加大搜索次数的方式来解决随机性问题，适用于 NIRS 特征波长的优选。但以全谱波长点为码长执行 GSA 算法耗时较长，在选定参数下，采用 10 折交叉验证时执行一次算法约需 8.27 h(硬件配置：CPU 为 AMD A6-7310 2.0 GHZ，内存 4 GB)，而且还需要考虑码长太长引起解空间发散的问题。结合 SiPLS 和 BiPLS 在特征谱区优选方面的优势，采用 GSA 对优选后谱区进行特征波长点优选，能够在兼顾波长优选性能的同时有效减少搜索时间。采用相同硬件配置下，按本区间划分方法执行 SiPLS 和 BiPLS 优选特征谱区的时间分别为 5.35 和 0.14 h；SiPLS 和 BiPLS 优选的纤维素特征谱区波长点分别为 388 和 358，然后分别再执行一次 GSA 二次优选算法的时间约为 0.91 和 0.69 h；SiPLS 和 BiPLS 优选的半纤维素特征谱区波长点分别为 160 和 180，然后分别再执行一次 GSA 二次优选算法的时间约为 0.39 和 0.48 h。多次执行 GSA 优选算法时，SiPLS-GSA 和 BiPLS-GSA 的搜索时间明显少于 Full-GSA。但 SiPLS 和 BiPLS 却

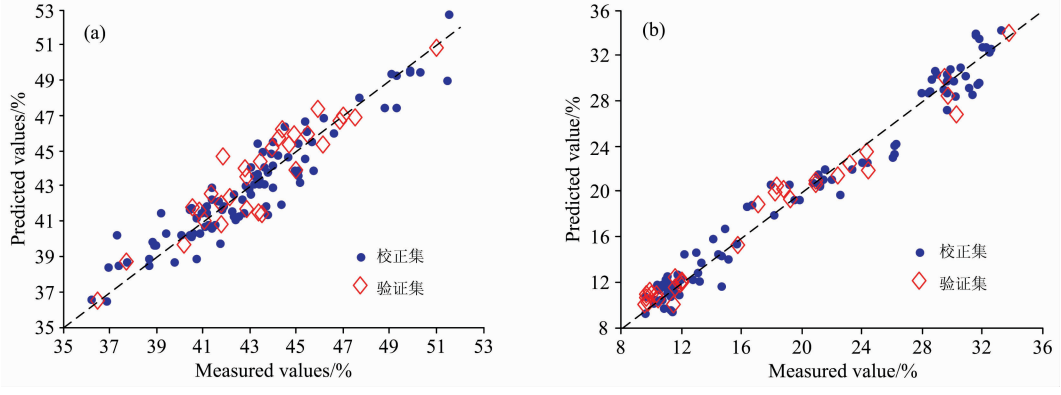


**图 4 纤维素和半纤维素实测值与预测值分布**

(a)：纤维素；(b)：半纤维素

**Fig. 4 Distribution of measured and predicted values for cellulose and hemicellulose**

(a): Cellulose; (b): Hemicellulose

限定了 GSA 搜索的波长点范围，在一定程度上影响了 GSA 在光谱空间上的全局寻优能力。因此，在解决实际问题时，需综合评定 Full-GSA，SiPLS-GSA 和 BiPLS-GSA 三种方法的性能，以确定最佳特征波长优选方案。

以 Full-GSA-16 作为纤维素的波长优选方案，以 SiPLS-GSA 作为半纤维素的波长优选方案，以优选后的特征波长进行 PLS 回归模型性能评测，其结果如图 4 所示。

由图 4 可知，纤维素和半纤维素含量的实测值与预测值点基本呈对角线分布，经检验发现各参数的预测值与实测值无显著性差异。纤维素和半纤维素回归模型的 $R_p^2$ 分别为 0.858 9 和 0.985 0，RMSEP 分别为 1.085 5 和 0.852 1，MRE 分别为 1.752 4%和 2.035 0%，RDP 分别为 2.663 7 和 8.488 6，可以用于预处理后玉米秸秆纤维素和半纤维素含量的 NIRS 快速定量检测。

## 4　结　论

基于结合温度参数设计适应度函数的策略构建的 GSA 具有良好的全局搜索性能，适用于玉米秸秆纤维素和半纤维素含量 NIRS 特征波长优选。GSA 以光谱波长点为染色体基因的编码方案适用于 NIRS 全谱的特征波长优选，更适用于 SiPLS 和 BiPLS 优选后谱区的特征波长优选，能够有效实现优选后谱区的波长点优选。通过波长优选，不仅参与建模的波长点数量显著减少，而且 PLS 回归模型的性能得到显著提升，预测精度更高。提出的 NIRS 波长优选方法，能够有效实现预处理后玉米秸秆纤维素和半纤维素含量的快速、准确检测。

## References


[1] Katsimpouras C, Zacharopoulou M, Matsakas L, et al. Bioresource Technology, 2017, 244: 1129.
[2] Yan X, Wang Z R, Zhang K J, et al. Bioresource Technology, 2017, 245: 419.
[3] Liu C M, Wachemo A C, Yuan H R, et al. Renewable Energy, 2018, 116: 224.
[4] Mourtzinis S, Cantrell K B, Arriaga F J, et al. Bioenergy Research, 2014, 7(2): 551.
[5] Xue J J, Yang Z L, Han L J, et al. Applied Energy, 2015, 137: 18.
[6] Jin X L, Chen X L, Shi C H, et al. Bioresource Technology, 2017, 241: 603.
[7] WANG Wen-xiu, PENG Yan-kun, XU Tian-feng, et al(王文秀，彭彦昆，徐田锋，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(12): 4001.
[8] Shen G H, Han L J, Fan X, et al. Journal of near Infrared Spectroscopy, 2017, 25(1): 63.
[9] Xie L J, Wang A C, Xu H R, et al. Transactions of the Asabe, 2016, 59(2): 399.
[10] Li X L, Sun C J, Zhou B X, et al. Scientific Reports, 2015, 5: 17210.
[11] Niu W J, Huang G Q, Liu X, et al. Energy & Fuels, 2014, 28(12): 7474.
[12] Yang Y, Wang L, Wu Y J, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017, 182: 73.
[13] Sheykhizadeh S, Naseri A. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 194: 202.
[14] Kutsanedzie F Y H, Chen Q, Hassan M M, et al. Food Chemistry, 2018, 240: 231.
[15] Kim J S, Lee Y Y, Kim T H. Bioresource Technology, 2016, 199: 42.


# Optimization of Characteristic Wavelength Variables of Near Infrared Spectroscopy for Detecting Contents of Cellulose and Hemicellulose in Corn Stover


LIU Jin-ming[1, 2], CHU Xiao-dong[1], WANG Zhi[1], XU Yong-hua[3], LI Wen-zhe[1], SUN Yong[1*]

1. College of Engineering, Northeast Agricultural University, Harbin 150030, China
2. College of Electrical and Information, Heilongjiang Bayi Agricultural University, Daqing 163319, China
3. School of Electrical and Information, Northeast Agricultural University, Harbin 150030, China



**Abstract** Pretreatment is an effective way to improve the utilization efficiency of the corn stover biotransformation. The conversion rate is directly related to contents of the cellulose and hemicellulose in corn stover during the bio-refinery conversion to biofuels. To achieve an effective control for the corn stover bio-refining process after the pretreatment, the near infrared spectroscopy (NIRS) was used to quickly detect contents of the cellulose and hemicellulose, solving the problems of being time consuming and high-cost in the traditional chemical analysis method. To improve the efficiency and precision of the NIRS detection, the genetic simulated annealing algorithm (GSA) based on genetic algorithm (GA) combined with simulated annealing algorithm (SA)

was presented for optimizing the characteristic wavelength variables of NIRS. In the GSA, firstly, the number of the NIRS wavelengths was used as the code length for binary coding; secondly, the root mean square error of cross-validation (RMSECV) of the partial least squares (PLS) regression model was used as the objective function; thirdly, the fitness function was designed combining with the temperature parameter; and last, the selective replication of the perturbation solution was realized based on the Metropolis criterion. Therefore, GSA can effectively improve the search efficiency at the later stage of evolution while avoiding premature convergence. 120 samples of corn stover were prepared by using the pretreatments of alkaline, biology, and the combination of alkaline and biology. The contents of cellulose and hemicellulose were measured using the wet chemistry methods. The NIRS were collected using the Nicolet Antaris Ⅱ Fourier near infrared spectrometer. The spectrum was pretreated by 7 points Savitzky-Golay smoothing combining with multivariate scattering correction and standard normal variate transformation. The samples were divided into correction set and validation set by using Kennard-Stone algorithm at a ratio of 3∶1. The GSA is used for the characteristic wavelength variables optimizations of the NIRS whole wavelengths (Full-GSA), the synergy interval partial least squares selected spectral region (SiPLS-GSA), and the backward interval partial least squares selected spectral region (BiPLS-GSA), respectively. And then, the optimized results of the characteristic wavelength variables were evaluated by the PLS regressive model with the validation set. In Full-GSA, 1 557 wavelength points were used as chromosome genes in whole wavelengths, 118 cellulose characteristic wavelength points and 164 hemicellulose characteristic wavelength points were selected after 16 executions. In SiPLS-GSA, the cellulose and hemicellulose wavelength points of spectral region optimized by SiPLS were 388 and 160, respectively, and 157 cellulose characteristic wavelength points and 148 hemicellulose characteristic wavelength points were gotten after the further optimization by GSA. In BiPLS-GSA, the cellulose and hemicellulose wavelength points of spectral region optimized by BiPLS were 358 and 180, respectively, and 130 cellulose characteristic wavelength points and 153 hemicellulose characteristic wavelength points were selected after the further optimization by GSA. It was shown that not only the number of wavelengths was significantly decreased after the optimization, but also the performance of regressive model was obviously better than that of the whole wavelengths. The best performance of regressive model for cellulose characteristic wavelengths was obtained by Full-GSA, and the best performance for hemicellulose characteristic wavelengths was obtained by SiPLS-GSA. The mean relative error (MRE) values of validation set for cellulose and hemicellulose in the best model were 1.752 4% and 2.020 8%, which were decreased by 13.636 6% and 25.368 4% compared with the whole wavelengths, respectively. The GSA combining with temperature parameters to design the fitness function is suitable for the NIRS characteristic wavelength selection of the cellulose and hemicellulose contents in corn stover, and has a good global search capability. The encoding scheme of GSA using each wavelength point in whole wavelengths as chromosome gene is suitable for the characteristic wavelength selection of NIRS whole spectrum. GSA is also suitable for the characteristic wavelength selection of the spectral region optimized by SiPLS and BiPLS, and the selection of wavelength points in the optimized spectral region can also be achieved effectively.

**Keywords** Corn stover; Near infrared spectroscopy (NIRS); Genetic simulated annealing algorithm (GSA); Synergy interval partial least squares (SiPLS); Backward interval partial least squares (BiPLS); Characteristic wavelength

(Received Mar. 10, 2018; accepted Jul. 28, 2018)

* Corresponding author