第37卷，第1期
2017年1月

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

Vol. 37，No. 1，pp95-102
January，2017

智能优化算法应用于近红外光谱波长选择的比较研究

宾 俊[1]，范 伟[1*]，周冀衡[1*]，李 鑫[1]，梁逸曾[2]

1. 湖南农业大学生物科学技术学院，湖南 长沙 410128
2. 中南大学化学化工学院，湖南 长沙 410083

摘 要 近红外光谱(NIRS)是一种间接分析技术，其应用需建立相应的校正模型。为了提高模型的解释能力、预测准确度和建模效率，需要对NIRS进行波长选择，优选最小化冗余信息。智能优化算法是以生物的行为方式或物质的运动形态为背景，经过数学抽象建立算法模型，通过迭代计算来求解组合最优化问题，其核心策略是以某种目标函数为标准，基于多元校正建模并以逐步逼近的方法筛选出有效的波长点。选用蚁群优化(ACO)、遗传优化(GA)、粒子群优化(PSO)、随机青蛙(RF)和模拟退火(SA)5种智能优化算法对烟叶总氮和烟碱近红外光谱数据进行特征波长选择，结合偏最小二乘(PLS)算法，构建了多个烟叶总氮和烟碱的校正模型，结果显示：所选用两个数据集的总氮最优模型分别为PSO-PLS和GA-PLS模型，烟碱最优模型分别为GA-PLS和SA-PLS模型，五种智能优化算法所建模型预测性能并非全部优于全谱PLS模型，但是通过智能优化算法进行波长选择后建立的PLS模型大大简化，模型的预测精度、可解释性和稳定性均有所提高。同时也对优选波长进行了解释和分析，烟叶总氮特征波长优选组合为4 587～4 878和6 700～7 200 cm^{-1}；烟叶烟碱特征波长优选组合为4 500～4 700和5 800～6 000 cm^{-1}，优选出来的特征波长具有实际物理意义。

关键词 近红外光谱；智能优化算法；波长选择；总氮；烟碱

中图分类号：TH741.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2017)01-0095-08

引 言

近年来，近红外光谱技术(NIRS)以其独特的优势在食品、制药、烟草、石化、纺织、农产品等行业得到日益广泛的应用。由于NIRS是一种间接分析技术，其应用依赖近红外光谱仪、化学计量学算法和应用模型三者的有机结合。在建立应用模型的研究领域，多元校正技术获得了飞速发展，为了解决光谱数据高度共线性的问题，大量基于潜变量的线性回归算法[1-2]被提出，偏最小二乘(PLS)是其中最核心的算法之一，在这些潜变量方法应用过程中，最佳潜变量数目并不容易确定，因此会发生模型的欠拟合或过拟合等问题，影响预测效果。而目前有研究表明[3-5]，对光谱数据进行特征波长选择有可能改善上述问题、提高模型的预测能力。

波长选择的本质是Occam's razor思想的继承，从一组波长中挑选出一些最有效的波长，使构造出来的光谱模型更优。通过波长选择，可以从原始光谱的波长集合中选择出使某种评估标准最优的波长子集，该子集能在最大程度上反映原始光谱信息，对因变量的解释性最强，且对模型的贡献处于较高水平，使机器学习(回归和分类)能达到与波长选择前近似甚至更好的效果。光谱数据进行波长选择的好处在于：(1)避免过度拟合，改进模型预测性能；(2)使模型训练器运行更快、效能更高；(3)剔除无信息波长和干扰波长，建立一个稳健性高、易解释、高精度的校正模型。而智能优化算法是目前多元校正中应用非常广泛的波长组合优化算法之一。

智能优化算法[6-7]又叫现代启发式算法，是以生物的行为方式或物质的运动形态为背景，经过数学抽象建立算法模型，通过迭代计算来求解组合最优化问题。此类算法通过模拟自然生态系统机制以求解复杂优化问题，具有严密的理论依据，为传统优化技术难以处理的组合优化问题提供了切实可行的解决方案。与其他类型的优化方法相比，其实现较简单、效率较高、通用性强，具有全局优化性能、跳出局部极值能力强、并且适合于并行处理，理论上可以在一定时间内找到最优解或近似最优解。

收稿日期：2015-10-23，**修订日期**：2016-02-22

基金项目：国家自然科学基金项目(21275164)，湖南省研究生科研创新项目(CX2015B237)资助

作者简介：宾 俊，1987年生，湖南农业大学生物科学技术学院博士研究生 e-mail：binjun2009@gmail.com

*通讯联系人 e-mail：wei_fan@foxmail.com；jhzhou2005@163.com

蚁群优化(ant colony optimization，ACO)算法[8-9]、遗传算法(genetic algorithm，GA)[10]、粒子群优化(particle swarm optimization，PSO)算法[11-12]、随机青蛙(random frog，RF)算法[13-14]和模拟退火(simulated annealing，SA)算法[15]是当前极具代表性、较新颖的五种智能优化算法，单独报道较多，但未见将它们进行综合比较分析的研究。鉴于此，本文选用这五种方法结合 PLS 技术分别对烤后烟叶中的总氮和烟碱含量建立了近红外定量分析模型，并对这些模型的稳健性和准确度进行了系统的对比分析，试图找出一种较好的烟叶近红外光谱样本特征波长选择方案，同时对优选波长进行了解释，找出了总氮和烟碱对应的最佳光谱波长组合。所建立的烟叶总氮和烟碱含量快速定量分析模型不仅能为烟叶实时品质分析提供技术支持，也可为智能优化算法的研究提供理论基础，还能为烟草专用近红外光谱仪的开发提供参考。

1 实验部分

1.1 材料

数据集 A 所用烟叶样本为云南烟草公司红云红河公司提供，数据集 B 所用烟叶样本为云南烟草公司昆明市公司提供，样本包含上、中、下部叶，品种为 K326，样品总氮含量根据《YC/T161—2002 烟草及烟草制品总氮的测定》行业标准用连续流动法测定，烟碱含量根据《YC/T160—2002 烟草及烟草制品总植物碱的测定》行业标准用连续流动法测定。具体样本信息见表 1。

表 1 烟叶样本统计信息

Table 1 Statistics for information of tobacco leaf samples

数据集	样本个数	成分	含量范围/%	平均值/%	标准差
数据集 A	258	总氮	1.690 0～2.810 0	2.151 6	0.261 54
		烟碱	0.980 0～3.710 0	2.645 4	0.570 14
数据集 B	165	总氮	0.880 0～2.580 0	1.648 2	0.378 95
		烟碱	0.920 4～4.665 5	2.844 5	0.810 12

1.2 光谱采集

数据集 A 是用 Nicolet Antaris FT-NIR 光谱仪以漫反射模式测量 258 个烟叶粉末样品得到的光谱数据。扫描次数为 32 次，分辨率为 8 cm^{-1}，光谱采集范围为 10 000～4 000 cm^{-1}(1 000～2 500 nm)，每个样本测量三次所得的平均值为测量光谱。

数据集 B 是用 BWTEK i-Spec 近红外光谱仪以漫反射模式测量 165 个烟叶粉末样品得到的光谱数据，扫描次数为 32 次，分辨率为 3.5 nm，光谱采集范围为 11 111～5 882 cm^{-1}(900～1 700 nm)，每个样本测量 3 次所得的平均值为测量光谱。

利用 Kennard-Stone 样本划分方法将数据集 A 的 181 个样品划作训练集，77 个样品划作预测集，将数据集 B 的 116 个样品划作训练集，49 个样品划作预测集。

1.3 方法

1.3.1 蚁群优化算法(ACO)

ACO 是根据自然界中真实蚁群觅食行为提出的一种简单、正反馈、分布式的群集智能演化计算算法。ACO 模拟蚂蚁的合作和适应机制等自然行为，每个蚂蚁在其所经过的路径上会遗留一种叫做信息素的挥发性物质，蚂蚁通过信息素和信息素强度的反馈机制选择路径，最终，可以找到更好的路径有更多的信息素沉积，所有的蚂蚁找到的特定路径将是解决目标问题的最优方案。自被开发以来，ACO 已在组合优化、函数优化及数据挖掘等问题上获得了广泛应用，基本的 ACO 模型有三个核心的算法步骤：

(1)选择概率：每个蚂蚁通过下述规则搜索下一个位置，假设蚂蚁当前所在节点为 i，可以到达的下一个节点为 j，概率为 P_{ij}，由两个参数决定：跟踪水平 τ^{α} 和期望信息 η^{β}。

$$P_{ij}=\begin{cases}\underset{k\notin \text{tabu}}{\arg\max}[\tau_{ik}^{\alpha}\eta_{ik}^{\beta}] & q\leqslant q_0\\ \dfrac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum\limits_{k\notin \text{tabu}}\tau_{ij}^{\alpha}\eta_{ij}^{\beta}} & j\notin \text{tabu},\ q>q_0\\ 0 & j\in \text{tabu},\ q>q_0\end{cases} \tag{1}$$

$$\eta_{ij}=\frac{E}{d(P_j,G)}=\frac{E}{\sqrt{(x_{P_j}-x_G)^2+(y_{P_j}-y_G)^2+(z_{P_j}-z_G)^2}} \tag{2}$$

(2)局部信息素更新：当任何一只蚂蚁经过一个边缘就按照下列式(3)和式(4)更新每个边缘的信息素。

$$\tau_{ij}\leftarrow(1-\rho_1)\tau_{ij}+\rho_1\tau_{ij}\frac{z^*}{z} \tag{3}$$

$$\text{fitness}=(y_i^*-\hat{y}_i)^2/2 \tag{4}$$

(3)全局信息素更新：全局更新的平均值可以减少信息素更新的计算。

$$\tau_{ij}\leftarrow\tau_{ij}+\rho_2\Delta\tau_{ij}^{BS} \tag{5}$$

$$\Delta\tau_{ij}^{BS}=\left(\frac{\bar{z}}{z^*}-1\right)\tau_{ij} \tag{6}$$

1.3.2 遗传算法(GA)

GA 是一种借鉴生物进化和自然选择的机制，利用选择、交换和突变等进化算子的操作使目标函数值变量“优胜劣汰”，并最终达到最优结果的自适应启发式全局搜索算法。经过蓬勃发展和不断完善，GA 算法已经被成功应用到了生物信息学、计算科学、经济学和化学等领域。GA 的一般步骤如下：

(1)寻找一种对问题潜在解进行“数字化”编码的方案；

(2)用适应性函数对每一个基因个体进行适应性评估；

(3)用选择函数按照规则从种群中择优选择两个个体作为父方和母方；

(4)基因交叉变异、产生子代；

(5)对子代基因进行变异；

(6)重复步骤(3)—步骤(5)，直到新种群的产生，找到满意的解。

1.3.3 粒子群优化算法(PSO)

PSO 是一种源于对鸟群捕食行为研究的进化计算技术。

同遗传算法类似，是一种基于叠代的优化工具，系统初始化为一组随机解，在解空间通过叠代搜寻最优值。在 PSO 中，每个优化问题的潜在解都可以想象成 D 维搜索空间上的一个点，可称之为“粒子”，所有的粒子都有一个被目标函数决定的适应值，每个粒子还有一个速度决定他们飞翔的方向和距离，然后粒子们就追随当前的最优粒子在解空间中搜索。作为近年来发展较快的优化计算方法，由于其容易实现且具有深刻的智能背景，已被广泛应用于函数优化、模式识别及模糊系统控制等领域。PSO 算法主要步骤如下：

(1)在 D 维问题中，定义粒子 i 根据邻近粒子和自身的经验不断调整其位置 x 和速度 v，用适应性函数来评估所有粒子并确定迭代次数。

$$x^i = (x^{i1}, x^{i2}, \cdots, x^{iD}) \tag{7}$$

$$v^i = (v^{i1}, v^{i2}, \cdots, v^{iD}) \tag{8}$$

(2)粒子 i 的速度通过跟踪当前最优粒子的位置 p_{best} 和所有粒子中的最优位置 g_{best} 来更新迭代

$$v_{k+1}^i = w_k v_k^i + c_1 r_1 (p_{\text{best}} - x_k^i) + c_2 r_2 (g_{\text{best}} - x_k^i) \tag{9}$$

$$x_{k+1}^i = x_k^i + v_{k+1}^i \tag{10}$$

$$w_k = w_{\max} - (w_{\max} - w_{\min}) \times (k-1)/\text{iter}_{\max} \tag{11}$$

式(9)中惯性权重 w_k 控制前期速度对当前速度的影响，正常数 c_1 和 c_2 表示加速度系数，随机系数 r_1 和 $r_2 \in [0, 1]$，k 表示当前迭代次数，$\text{iter}_{\max}$ 表示最大迭代次数。当满足迭代终止条件后，迭代终止。

1.3.4　随机青蛙算法(RF)

RF 是新近提出的一种基于逆跳马尔科夫蒙特卡洛(RJMCMC)的模型维数转换技术与模型集群分析(MPA)思想的特征波长选择算法。它利用大量序贯采样得到的子模型，计算出每个变量的选择频率并对变量的重要性进行评价，从而挑选出特征波长。其主要工作流程包括如下：

(1)初始化可调参数，随机产生一个含有 Q 个变量的初始变量子集 V_0；

(2)根据标准正态分布 $N(Q, \theta Q)$ 产生一个随机数，将离此随机数最近的整数 Q^* 作为候选变量子集 V^* 包含的变量个数；

(3)利用 Q^* 与 Q 的关系，建立一个含有 Q^* 个变量的候选变量子集 V^*；

(4)按 RMSECV 计算的概率确定 V^*，并令 $V_1 = V^*$，如此迭代 N 步；

(5)综合分析每步产生的候选变量子集，计算每个变量的选择频率。

1.3.5　模拟退火算法(SA)

SA 是由 Kerkpatrick 等提出的一种基于固体物理退火原理而研发的随机全局优化技术，与确定性优化技术相反，SA 作为一种概率优化技术，在解决组合优化问题时先从某一模拟较高初温开始，随着温度参数的不断下降，结合 Metropolis 标准在解空间中随机寻找目标函数的全局最优解，SA 最大的优势是其能以一定概率接受非局部最优解，经过大量的迭代变化后能跳出局部最优并趋于全局最优。由于 SA 能够遍历局部最优来寻找全局最优解，因而广泛应用于各类优化问题。其主要思想包括如下 6 个步骤：

(1)初始化：初始温度 T，初始解状态 v_0(算法开始迭代的起点)，每个 T 值的迭代次数 L。

(2)产生新解 v'。

(3)计算增量 $\Delta F = F(v') - F(v_i)$，其中 $F(v_i)$ 为评价函数。

(4)若 $\Delta F < 0$ 则接受 v' 作为新的解，否则以概率 $p(\Delta F) = \exp\left(\frac{-\Delta F}{T}\right)$ 接受 v' 作为新的当前解。

(5)对 $k = 1, 2, \cdots, L$ 执行第(2)—(4)步。

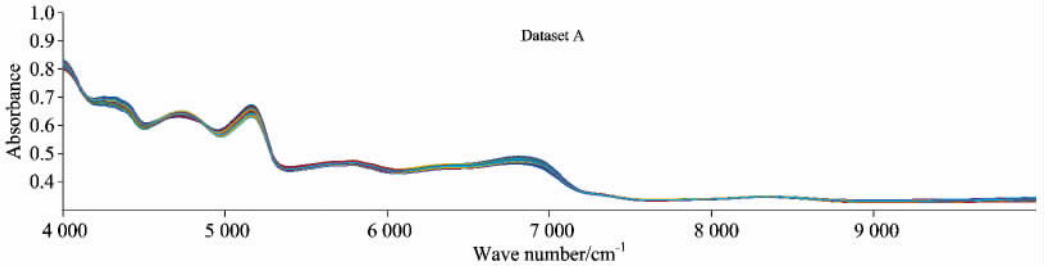
(6)如果连续若干个新解都没有被接受时算法终止，输出当前解作为最优解。

采用上述算法结合 PLS 建立总氮和烟碱的预测模型时，利用交互验证法来优化模型的相关参数，以样品的测量值和光谱预测值相关系数(R^2)、预测集相关系数(Q^2)、交互验证均方根误差(RMSECV)及预测均方根误差(RMSEP)作为评价模型的有效指标。所有数据分析都是使用 Matlab R2015a 软件完成的。

2　结果与讨论

2.1　光谱预处理

采集到的烟草原始近红外光谱数据夹杂仪器操作、样品背景和杂散光等引起的噪声和无关信息，因此在对光谱数据进行操作之前，需要对光谱进行预处理，预处理有助于消除干扰因素、促进有用信息的提取。两个光谱数据皆采用如下预处理方法：(1)使用蒙特卡洛采样法剔除光谱数据中的奇异样本；(2)采用 Savitzky-Golay 平滑法对光谱进行平滑，平滑点数为 5，多项式次数为 3；(3)应用多元散射校正法消除样品颗粒分布不均匀及颗粒大小产生的散射影响。预处理后的光谱见图 1。



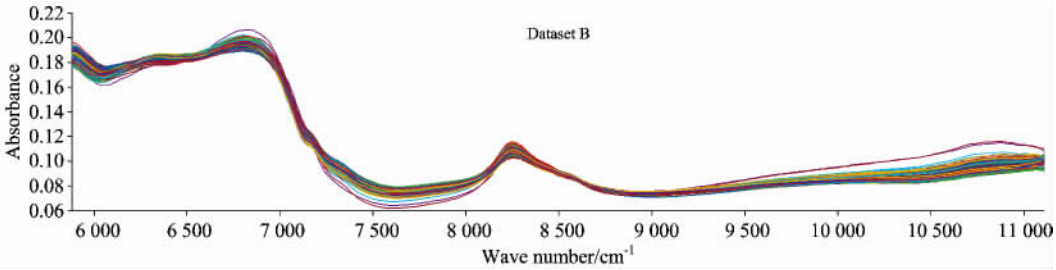


图 1 数据集 A 和数据集 B 的预处理后的近红外光谱

Fig. 1 Processed NIR spectra of dataset A and B

2.2 特征波长选择

2.2.1 ACO 算法优选波长

使用 ACO 算法进行特征波长选择时，首先要对影响 ACO 算法运算性能的参数进行初始化设置。蚁群大小是蚁群算法的重要参数之一，一般根据待处理问题规模大小来设置，既要保证算法的全局搜索能力和稳定性，同时兼顾收敛速度，本文设置为 40；初始化时由于所有节点信息素强度相同，则信息素强度 $\tau=1$，蚂蚁对每个节点的选择概率相同，则表示跟踪水平的信息启发因子 $\alpha=1$；能见度 $\eta=1$，为了降低算法的随机性，令表示能见度的期望启发因子 $\beta=2$，信息素耗散常数 $\rho=0.95$，种群进化代数设为 150。运行 ACO 算法，设置光谱窗口为 1，蚁群的适应度采用函数 $F=\frac{1+\mathrm{RMSECV}}{R^2}$ 来评估，一般在进化至 100 代算法收敛，组合优化后建立的 PLS 模型性能指标见表 2。

表 2 ACO-PLS 模型的性能指标

Table 2 Performance index of ACO-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs	nVAR
数据集 A	总氮	0.972 5	0.044 8	0.919 5	0.068 7	15	70
	烟碱	0.995 3	0.041 3	0.987 0	0.054 7	15	80
数据集 B	总氮	0.893 7	0.130 7	0.864 9	0.154 0	11	75
	烟碱	0.961 2	0.16	0.909 4	0.237 6	12	80

2.2.2 GA 优选波长

GA 用于光谱波长组合优化时，为了保证算法能在进化代数内收敛，令进化代数为 200，同时为了保证这种随机搜索算法优化结果的可靠性和稳定性，算法运行次数设置为 50。按照 GA 算法原理，其他参数设置如下：种群大小为 30，初始时平均 5 个波长(基因)构成 1 个染色体，染色体个数为 30，染色体交叉概率为 50%，变异概率为 1%，每次进化淘汰 5 个组合。选择适应度函数 $F=\frac{R^2}{1+\mathrm{RMSEP}}$，$R^2$ 为交互验证预测值和量测值的相关系数，R^2 越大，RMSEP 越小，F 就越大，表明模型的预测性能越好，建立的 GA-PLS 模型性能指标见表 3。

2.2.3 PSO 算法优选波长

PSO 具有深刻的智能背景，同遗传算法比较，PSO 简单易实现且没有过多参数需要调整。运行时首先设置粒子群搜索参数：粒子种群大小为 50，迭代次数为 200，算法运行 10 次，以 $F=Q^2$ 作为适应度函数，建立的 PSO-PLS 模型性能指标见表 4。

表 3 GA-PLS 模型的性能指标

Table 3 Performance index of GA-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs	nVAR
数据集 A	总氮	0.970 3	0.046 4	0.925 6	0.066 0	13	165
	烟碱	0.996 5	0.035 5	0.994 5	0.039 9	15	155
数据集 B	总氮	0.897 1	0.128 6	0.885 3	0.147 2	9	50
	烟碱	0.965 0	0.152 0	0.913 7	0.231 8	10	59

表 4 PSO-PLS 模型的性能指标

Table 4 Performance index of PSO-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs	nVAR
数据集 A	总氮	0.973 5	0.043 8	0.926 0	0.065 8	13	773
	烟碱	0.995 9	0.038 5	0.990 0	0.049 1	15	605
数据集 B	总氮	0.901 4	0.125 9	0.862 9	0.154 7	10	225
	烟碱	0.982 9	0.106 3	0.919 2	0.224 4	13	173

2.2.4 RF 算法优选波长

RF 算法在运行时有 5 个参数需要初始化，包括迭代次数 N、初始化子集变量个数 Q、正态分布方差控制参数 θ、候选变量比例因子 ω 以及概率上限参数 η。根据经验，在数据运行中取 $N=10\ 000$，由于其与变量个数相关，必须足够大来保证算法收敛；Q 只对第一次迭代有影响，但对整体性能没有明显影响，一般将其设置为 50；参数 θ，ω 和 η 对运算结果影响不大，分别保持默认设置 0.3，3 和 0.1。建立的 RF-PLS 模型的各性能指标见表 5。

表 5 RF-PLS 模型的性能指标

Table 5 Performance index of RF-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs	nVAR
数据集 A	总氮	0.967 0	0.048 9	0.921 6	0.067 8	11	1 075
	烟碱	0.997 6	0.029 7	0.990 5	0.049 8	15	490
数据集 B	总氮	0.896 7	0.128 9	0.873 4	0.151 2	10	332
	烟碱	0.962 0	0.158 4	0.903 4	0.245 2	10	155

2.2.5 SA 算法优选波长

根据 SA 基本原理，要成功运用 SA 算法解决组合优化

问题，必须合理的选择算法控制进程参数。首先基于 SA 算法操作只能求最大值的原则，以 $F=\frac{1}{1+\mathrm{RMSECV}}$作为算法的适应度函数，当 RMSECV 取得最小值时，函数 F 取得最大值。SA 算法是根据模拟固体退火过程产生的，故其最大的特点就是冷却进度表的设计，直接影响算法性能。基于算法的准平衡概念，冷却进度表参数设置如下：初始温度 $T_0=100$ ℃，第 k 个温度控制参数值 $T_k=0.95\mathrm{T}$，终止温度 $T_f=0$ ℃，第 k 个马尔科夫链的长度 $L_k=50$。建立的 SA-PLS 模型性能指标见表 6。

表 6　SA-PLS 模型的性能指标

Table 6　Performance index of SA-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs	nVAR
数据集 A	总氮	0.972 9	0.044 3	0.925 2	0.066 2	13	404
	烟碱	0.994 4	0.045 1	0.989 2	0.049 9	15	400
数据集 B	总氮	0.886 6	0.135 0	0.876 8	0.150 1	10	99
	烟碱	0.977 2	0.122 6	0.938 0	0.196 5	15	122

2.3　特征波长解析

数据集 A 的总氮样本通过 ACO，GA，PSO，RF 和 SA 五种智能优化算法分别选出 70，165，773，1 075 和 404 个波长。ACO 优选出来的波长比较分散，在 4 700 和 5 500 cm^{-1} 附近优选波长较多，GA 优选波长主要分布在 4 500～5 000 cm^{-1}，PSO 优选波长较多，基本分布在 4 000～7 000 cm^{-1} 区间，RF 优选了 11 个波段，主要优选波长位于 4 000～5 000 和 8 000～10 000 cm^{-1}，SA 优选波长在整个区间都有分布，主要分布在波段 4 000～6 000 cm^{-1}。五种方法皆选出来的波长有 11 个，分别是 4 289，4 524，4 601，4 617，4 640，4 760，4 868，4 871，6 761，6 777 和 7 031 cm^{-1}，这与 4 587～4 878 cm^{-1} 附近是蛋白质、酶等物质的 N—H 伸缩振动的一级倍频和组合频吸收的特征谱带相符合。具体被选出的特征波长见图 2，图中阴影部分表示被选择的特征波长，图 3—图 5 同。

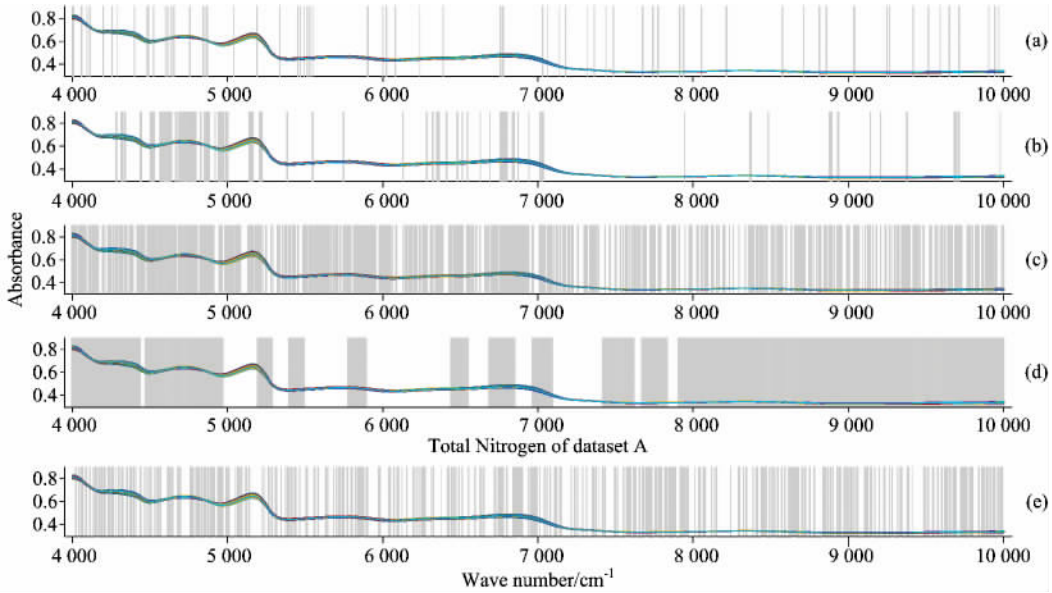


图 2　数据集 A 的总氮模型选择的波长，(a) ACO，(b) GA，(c) PSO，(d) RF 和(e) SA

Fig. 2　The selected wavelengths of total nitrogen in dataset A by (a) ACO, (b) GA, (c) PSO, (d) RF and (e) SA

数据集 A 的烟碱样本通过五种智能优化算法分别选出 80，155，605，490 和 400 个波长，ACO 算法优选波长度集中在 5 000～7 000 cm^{-1} 区间，GA 算法优选波长基本位于 4 000～6 000 cm^{-1} 范围，PSO 优选波长多位于 4 000～7 000 cm^{-1} 波段，而 RF 除位于 9 000～10 000 cm^{-1} 区间的 2 个波段外，在 4 000～7 000 cm^{-1} 有 6 个波段，SA 算法优选的波长分布较广，但在 4 000～7 000 cm^{-1} 范围也有大量分布。具体被选出的特征波长见图 3。五种方法皆选出的波长共 4 个，分别是 4 042，5 612，5 747 和 8 404 cm^{-1}。烟碱分子式中含有一个吡啶环和一个吡咯环，4 500～4 700 cm^{-1} 为吡啶的第一组合频吸收区，也是烟碱的一个特征谱带之一。

数据集 B 的总氮样本通过五种智能优化算法分别选出 75，50，225，332 和 99 个波长，从图 4 可知，ACO 算法选择的波长大多位于 6 400～6 600 以及 8 500 cm^{-1} 附近，而 GA 算法只选择了 50 个波长，主要分布在 6 000，7 200 和 8 700 cm^{-1} 附近，PSO 算法的优选波长多集中于 6 000～6 500，7 200～7 400 及 8 700～8 800 cm^{-1} 区间，RF 优选波长在区间 5 900～6 200 和 6 700～8 600 cm^{-1}，而 SA 算法选择的特征波长在 6 000～8 300 cm^{-1} 都有分布。五种方法同时有 7 个波长被选出，分别是 5 996，7 051，7 291，7 343，8 574，8 660 和 8 735 cm^{-1}。6 700～7 200 cm^{-1} 为 N—H 的合频吸收谱带，五种算法所选波长在此波段皆有分布。

数据集 B 的烟碱样本通过五种智能优化算法分别选出 80，59，173，155 和 122 个波长，从图 5 可以看出，五种算法优选波长在区间 5 900～6 200 cm^{-1} 都有较大分布，这与文献研究的烟碱在 5 900～6 400 cm^{-1} 有 N—H 特征吸收相吻合；

除 ACO 外，其他四种算法优选波长在 7 100～7 400 cm^{-1} 波段也有分布，而 ACO，PSO 和 RF 算法在 8 500 cm^{-1} 附近集中选择了一定的特征波长，这可能是烟碱的其他特征吸收。五种方法选出来 5 个共有波长，分别是 5 892，5 974，5 991，8 042 和 8 053 cm^{-1}。

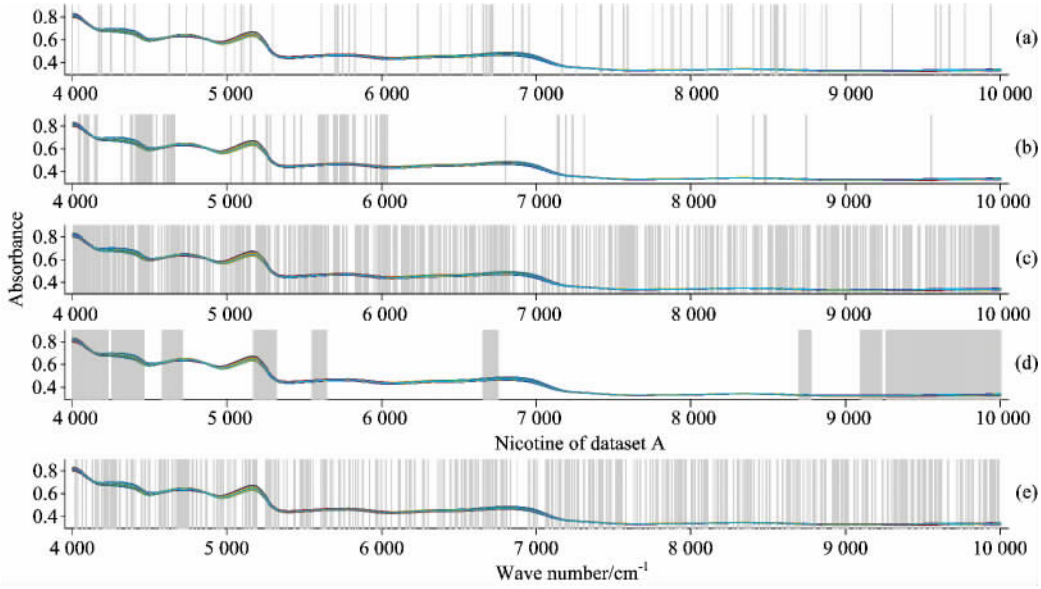


图 3 数据集 A 的烟碱模型选择的波长，(a) ACO，(b) GA，(c) PSO，(d) RF 和(e) SA

Fig. 3 The selected wavelengths of nicotine in dataset A by (a) ACO, (b) GA, (c) PSO, (d) RF and (e) SA

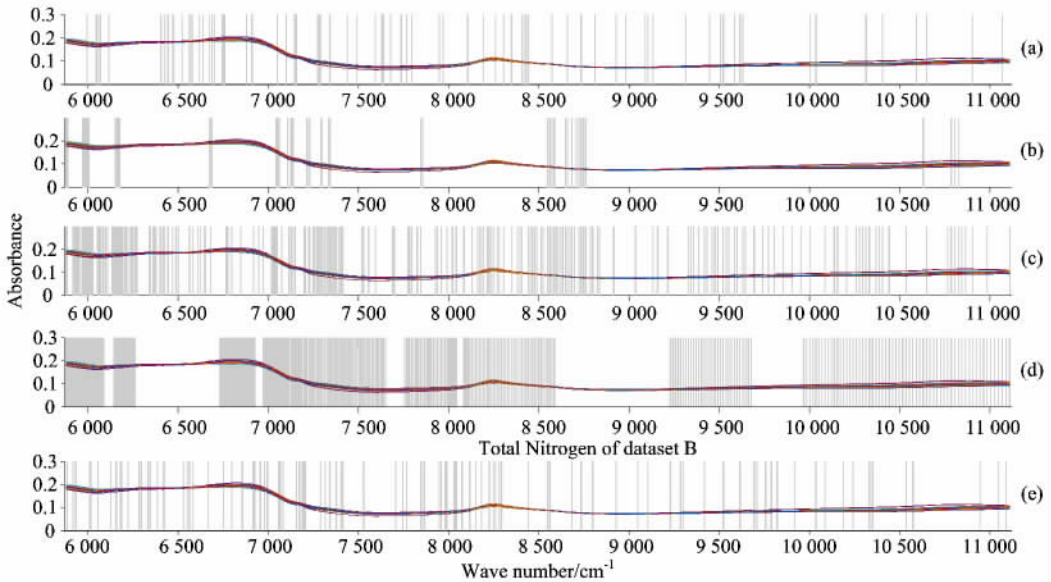


图 4 数据集 B 的总氮模型选择的波长，(a) ACO，(b) GA，(c) PSO，(d) RF 和(e) SA

Fig. 4 The selected variables of total nitrogen in dataset B by (a) ACO, (b) GA, (c) PSO, (d) RF and (e) SA

2.4 小结

全谱 PLS 模型性能指标见表 7。对于数据集 A 的总氮样本，ACO-PLS 与 RF-PLS 建立的模型与全谱 PLS 模型性能相当，但是 ACO-PLS 建模所用波长 70，较全谱(1 557)有极大的减少，所建模型更加简单，稳定性更强，其他三种方法建立的模型较全谱模型性能更优，所用波长也较少，分别为 163，773 和 404，总的说来，PSO-PLS 模型效果最佳。

在数据集 A 的烟碱样本中，由于全谱 PLS 模型相关系数已达 0.995 3，故而经过波长优化选择后，模型性能指标改善不明显，但模型更加简单、易解释，GA-PLS 仅用 155 个波长建模，结果优于全谱 PLS 模型，波长选择优势明显。

数据集 B 的总氮样本所建各模型中，RF-PLS 和 PSO-PLS 建模所用波长较多，且性能仅有少量提升，而 GA-PLS 仅用 50 个波长建模性能超过了全谱 PLS 和其他优化算法所建模型，简化了模型。

尽管数据集B的烟碱样本建立的模型中，ACO-PLS，

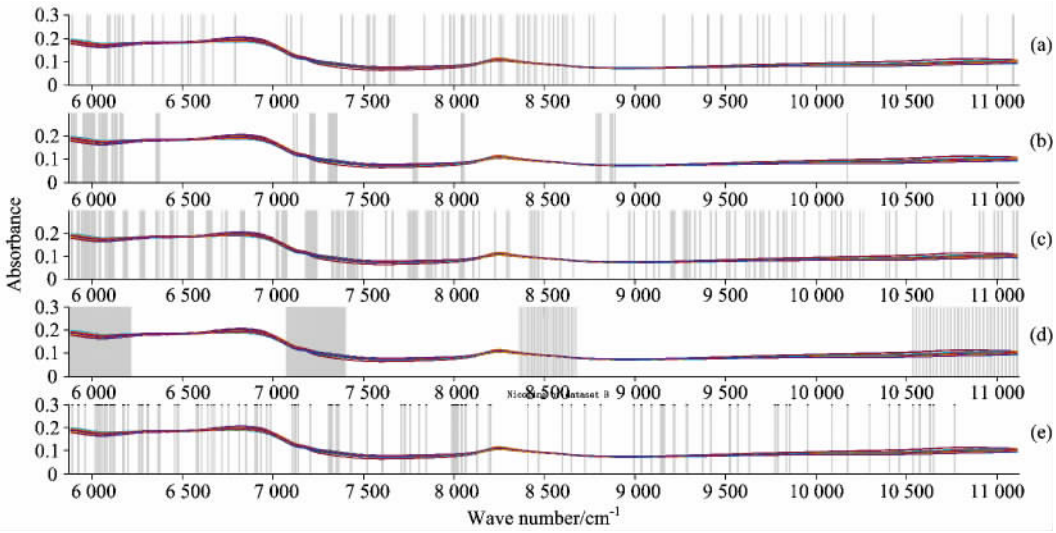


图 5　数据集 B 的烟碱模型选择的波长，(a) ACO，(b) GA，(c) PSO，(d) RF 和(e) SA

Fig. 5　The selected wavelength of nicotine in dataset B by (a) ACO, (b) GA, (c) PSO, (d) RF and (e) SA

表 7　全谱 PLS 模型的性能指标

Table 7　Performance index of full-PLS models

数据集	成分	R^2	RMSEC	Q^2	RMSEP	optPCs
数据集 A	总氮	0.967 2	0.048 8	0.924 7	0.066 4	12
	烟碱	0.995 3	0.041 6	0.990 7	0.040 6	15
数据集 B	总氮	0.894 0	0.130 6	0.870 6	0.152 1	10
	烟碱	0.985 5	0.097 8	0.918 0	0.226 0	15

RF-PLS 和 GA-PLS 模型性能与全谱 PLS 相比略有不及，但是 PSO-PLS 和 SA-PLS 模型预测性能都优于全谱 PLS，而 SA-PLS 模型更优。

综上所述，五种方法表现各有优劣，但 GA 算法在两种不同近红外仪器所测的四个光谱数据中均有较好表现，可认为 GA 算法是近红外光谱波长优化的较好选择。

3　结　论

利用智能优化算法在搜索过程中不容易陷入局部最优以及较高的搜索效率等优点，从目前近红外光谱分析中最常用的定量分析模型入手，建立了五种智能优化算法用于烟叶近红外光谱的特征波长选择结合 PLS 的烟叶总氮和烟碱的定量预测模型，并比较了不同智能优化方法选择波长以及建立模型之异同，也比较了两种不同分辨率、不同波长扫描范围的主流近红外仪器测量同一化学指标之异同，结果表明：GA 算法在组合优化问题上具有较大的搜索优势，适应性较广，且优选出来的特征波长有实际物理意义，是近红外光谱波长优化的较好选择，优选波长后所建模型较全谱 PLS 模型大大简化，模型预测精密度有所提高，建模变量有所减少。利用智能优化算法建立的烟叶总氮和烟碱含量快速定量分析模型不仅能为烟叶实时品质分析提供技术支持，也可为智能优化算法的研究提供理论基础，还能为烟草专用近红外仪的开发提供参考。

References

[1] Fang Yi, Park Jong I, Jeong Young-Seon, et al. Annals of Operations Research, 2009, 190(1): 3.
[2] Bin Jun, Ai Fangfang, Liu Nian, et al. Journal of Chemometrics, 2013, 27(12): 457.
[3] Swierenga H, Wulfert F, de Noord O E, et al. Analytica Chimica Acta, 2000, 411: 121.
[4] Balabin R M, Smirnov S V. Analytica Chimica Acta, 2011, 692(1-2): 63.
[5] Roy P P, Roy K. QSAR & Combinatorial Science, 2008, 27(3): 302.
[6] Wang Ling. Intelligent Optimization Algorithms with Applications. Beijing: Tsinghua University & Springer Press, 2001.
[7] Jin Huimin, Ma Liang, Wang Zhoumian. Computer Engineering, 2006, 32(10): 201.
[8] Allegrini F, Olivieri A C. Analytica Chimica Acta, 2011, 699(1): 18.
[9] Zhu Yaodi, Zou Xiaobo, Huang Xiaowei, et al. Journal of Applied Solution Chemistry and Modeling, 2013, 2: 25.
[10] Jarvis R M, Goodacre R. Bioinformatics, 2005, 21(7): 860.
[11] Khajeh A, Modarress H, Zeinoddini-Meymand H. Journal of Chemometrics, 2012, 26(11-12): 598.
[12] Jiang Huaqin, Yan Zhengbing, Liu Xinggao. Neurocomputing, 2013, 119: 469.

[13] Li Hongdong, Xu Qingsong, Liang Yizeng. Analytica Chimica Acta, 2012, 740: 20.
[14] Yun Yonghuan, Li Hongdong, Wood L R, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013, 111: 31.
[15] Brusco Michael J. Computational Statistics & Data Analysis, 2014, 77: 38.

Application of Intelligent Optimization Algorithms to Wavelength Selection of Near-Infrared Spectroscopy

BIN Jun[1], FAN Wei[1]*, ZHOU Ji-heng[1]*, LI Xin[1], LIANG Yi-zeng[2]
1. College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha 410128, China
2. College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China

Abstract Near infrared spectroscopy (NIRS) is a kind of indirect analysis technology, whose application depends on the setting up of relevant calibration model. In order to improve interpretability, accuracy and modeling efficiency of the prediction model, wavelength selection becomes very important and it can minimize redundant information of near infrared spectrum. Intelligent optimization algorithm is a sort of commonly wavelength selection method which establishes algorithm model by mathematical abstraction from the background of biological behavior or movement form of material, then iterative calculation to solve combinatorial optimization problems. Its core strategy is screening effective wavelength points in multivariate calibration modeling by using some objective functions as a standard with successive approximation method. In this work, five intelligent optimization algorithms, including ant colony optimization (ACO), genetic algorithm (GA), particle swarm optimization (PSO), random frog (RF) and simulated annealing (SA) algorithm, were used to select characteristic wavelength from NIR data of tobacco leaf for determination of total nitrogen and nicotine content and together with partial least squares (PLS) to construct multiple correction models. The comparative analysis results of these models showed that, the total nitrogen optimums models of dataset A and B were PSO-PLS and GA-PLS models. GA-PLS and SA-PLS models were the optimums for nicotine, respectively. Although not all predicting performance of these optimization models was superior to that of full spectrum PLS models, they were simplified greatly and their forecasting accuracy, precision, interpretability and stability were improved. Therefore, this research will have great significance and plays an important role for the practical application. Meanwhile, it could be concluded that the informative wavelength combination for total nitrogen were 4 587～4 878 and 6 700～7 200 cm^{-1}, and that for tobacco nicotine were 4 500～4 700 and 5 800～6 000 cm^{-1}. These selected wavelengths have actually physical significance.

Keywords Near-infrared spectroscopy; Intelligent optimization algorithm; Wavelength selection; Total nitrogen; Nicotine

(Received Oct. 23, 2015; accepted Feb. 22, 2016)

* Corresponding authors