第41卷，第10期
2021年10月

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

Vol. 41, No. 10, pp3032-3038
October, 2021

改进团队进步算法的近红外光谱波长筛选

高美凤，陶焕明

江南大学轻工过程先进控制教育部重点实验室，物联网工程学院，江苏 无锡 214122

摘 要 针对近红外光谱波长选择问题，在团队进步算法(TPA)的基础上，提出一种改进团队进步算法(iTPA)的波长变量选择方法，将分子光谱的波段按照与其相应的理化值建模得到的评价值函数大小降序排列，顺序分为精英组、普通组和垃圾回收组。当新生波段选择学习行为时，若其产生于普通组，则需要向精英组样板的方向调节；若其产生于精英组，则需要改进其更新方向，向垃圾回收组样板的反方向调节。垃圾回收组成员的评价值不像精英组和普通组随着更新的过程一直上升，而是一直处于极低的状态，为产生于精英组的新生波段在学习时提供一个准确的更新方向，从而提升算法的全局寻优能力。通过不断的迭代更新，逐步提升整体评价值，最终选取评价值最高的波段作为筛选波段。该算法对玉米的淀粉和蛋白质含量数据集进行了实验测试，并与TPA、遗传算法(GA)、主成分分析(PCA)以及全谱方法进行了对比。实验结果表明，所提算法能够找出全谱范围内波长的最优组合，并且可以解释各含量的化学特性。玉米淀粉数据集运行的效果相比于全光谱，变量个数从700个减少到17.55个左右(50次试验求平均)，模型的RMSEC从0.335 7降到0.260 9，校正集预测精度提升了22.3%，模型的RMSEP从0.391 4下降到0.334 4左右，预测集预测精度提升了14.6%；在玉米蛋白质数据集运行的效果相比于全光谱，变量个数从700个减少到19.6个左右(50次试验求平均)，模型的RMSEC从0.147 4降到0.101 9，校正集预测精度提升了30.1%，模型的RMSEP从0.178 9下降到0.117 7，预测集预测精度提升了34.2%。

关键词 近红外光谱；波长选择；改进的团队进步算法；智能组合优化；特征波长

中图分类号：Q657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)10-3032-07

引 言

近红外光谱分析技术凭借其快速、无损以及低成本的特点，已经广泛运用在食品分析、生物医学、农业等领域。但由于近红外光谱不同的波长点对被测物质不同化学基团的吸收强度不同，被测物质浓度并不与光谱全谱波长信息相关，往往采集到的全光谱数据存在大量的冗余信息以及噪声信息，与之直接相关的有效信息仅仅存在于全谱信息中的一部分。如何快速有效地提取出有用光谱信息是近红外光谱技术研究的重点之一。

目前，国内外研究学者提出大量基于不同原理策略的波长筛选方法，主要有：以全谱PLS模型的某些参数作为变量选择的依据，如无信息变量消除法(UVE)对回归系数设定阈值限制来选择有效变量[1]；以光谱区间为筛选对象的方法，如区间偏最小二乘(iPLS)[2]、移动窗口偏最小二乘(MWPLS)[3]、向前和向后间隔偏最小二乘(FB-iPLS)[4]、区间随机蛙跳(iRF)[5]等；以连续投影策略进行变量排序筛选出最优变量子集，如连续投影算法(SPA)[6]；以模型集群分析策略的方法，如变量空间迭代收缩(VISSA)[7]、变量组合集群分析(VCPA)[8]；以智能算法为核心进行波长组合优化，如遗传算法(genetic algorithm，GA)[9-11]、免疫遗传算法(IGA)[12]、粒子群算法(PSO)[13]、蚁群算法(ACO)[14]、二进制蜻蜓算法(BDA)[15]等。团队进步算法(team progress algorithm，TPA)[16-17]是一种典型的智能组合优化算法，它是通过树立榜样引导团队进步的方向，利用合理的分工合作提高工作效率。

近红外光谱波长变量多，需要在多达几百甚至上千的波长点中，选择最有效的波长点，使得所建模型的预测精度最高，因此将波长筛选问题转化为波长点之间的组合优化问题，通过智能组合优化算法寻找光谱中的最佳波长。在团队进步算法的基础上，提出一种应用于近红外光谱波长筛选的智能组合优化算法：改进团队进步算法(improved team progress algorithm，iTPA)。首先将波长变量均分为若干波段，

收稿日期：2020-08-28，**修订日期**：2020-12-12

基金项目：国家自然科学基金项目(61833007)资助

作者简介：高美凤，1963年生，江南大学物联网工程学院副教授 e-mail：mfgao@jiangnan.edu.cn

对波段按照与其相应的理化值进行 PLS 建模，将得到的评价值函数按大小降序排列，波段分组中增设一个垃圾回收组。按评价值从高到低依次分为精英组、普通组和垃圾回收组，建立精英组和垃圾回收组两个学习样板，结合学习和探索的过程，通过不断的迭代更新，选取评价值最高的波段作为筛选波段。该算法在计算迭代过程中，每个波段的波长点在不断进行更新，各小组在搜索过程中出现明显的分工，使算法具备良好的全局搜索能力。

1　算法原理及实现

1.1　TPA 算法实现原理

1.1.1　数学模型

TPA 是一种双群体搜索算法[16]，模仿团队两个小组(精英组和普通组)的学习和探索过程，并设计合理的成员更新规则，逐步提升其评价值以达到全局最优。实验表明，该算法能够在较少的计算量前提之下快速寻找全局最优。

算法模型多变量无约束最大化问题可表示为

$$\begin{cases} Y = \min\{f(x)\} \\ x = [x_1, x_2, \cdots, x_n]^{\mathrm{T}} \\ x_i \in [a_i, b_i],\ i = 1, 2, \cdots, n \end{cases} \tag{1}$$

式(1)中，向量 x 代表一个成员，即为包含多个波长的波段；x_i 表示该成员的第 i 个能力因素，即该波段中的第 i 个波长点；a_i 和 b_i 分别表示 x_i 的上下边界值；而函数 $f(x)$代表该成员 x 的评价值。通过更新成员以逐步提升或降低评价值来寻求最优成员。

1.1.2　分组规则

将整条光谱波段均分为 P 个波段同时确定其评价值，评价值为该波段与测得的含量理化值进行 PLS 建模得到的校正集均方差(RMSEC)和相关系数(r)为变量的函数值，本工作设定的评价值为

$$f(x) = \frac{r}{1 + \mathrm{RMSEC}} \tag{2}$$

每一个波段就相当于向量 x。将 P 个波段按评价值从大到小排列分成 $N+M$ 个波段(N 和 M 均为整数)，N 个评价值较高的波段组成精英组$\{x_{e1}, x_{e2}, \cdots, x_{en}\}$，$M$ 个评价值较低的波段组成普通组$\{x_{p1}, x_{p2}, \cdots, x_{pm}\}$。通过团队进步过程更新组内成员。

1.1.3　行为定义

需要产生一条新生波段 x_r，新生波段可从任意一组产生，其波长点从当前组随机波段同一波长点中继承。若新生波段出身自精英组，且该新生波段的第 i 个波长点如果是在精英组第 n 个波段中产生，那么该波长点需继承精英组第 n 个波段中第 i 个波长点；若新生波段出身自普通组，且该新生波段的第 i 个波长点如果是在普通组第 m 个波段中产生，那么该波长点需继承普通组第 m 个波段中第 i 个波长点。继承下来的新生波段通过设定概率选择一次学习或者探索行为以更新自身的波长点，才能成为候选波段 x_c。

新生成员 x_r 如果选择进行学习行为，则需要向参照目标方向调节。参照方向产生于精英组和普通组，分别称为精英组样板 e_e 和普通组样板 e_p，且样板值取所在组波段波长的平均值。若新生波段产生于普通组，则其需向精英组样板调节；若新生波段产生于精英组，则向普通组样板的反方向调节，如式(3)所示。

$$\begin{cases} x_c = (1-\gamma)\ x_r + \gamma e_e \\ x_c = (1+\gamma) x_r - \gamma e_p \end{cases} \tag{3}$$

式(3)中，γ 为区间[0，1]内随机数。若 x_c 某个波长点范围越界，则改用其边界值。

新生成员如果进行探索行为，则其各波长点 x_{ri}($i=1$，2，…，n)将做随机改变。并且探索强度逐步减小。两组新生波段 x_r 经过探索行为生成 x_c 的表达式如式(4)和式(5)所示。

$$x_c = [x_{c1}, x_{c2}, \cdots, x_{cn}]^{\mathrm{T}} \tag{4}$$

$$x_{ci} = \begin{cases} x_{ri} + \gamma_i t_{e,p}(b_i - x_{ri}), & m_i = 0 \\ x_{ri} - \gamma_i t_{e,p}(x_r i - a_i), & m_i = 1 \end{cases} \tag{5}$$

式(5)中，γ_i 是区间(0，1)的随机数，m_i 为 0 和 1 二值随机整数。收缩系数 $t_{e,p}$ 为

$$t_{e,p} = \left(1 - \frac{k}{K}\right)^{\alpha_{e,p}} \tag{6}$$

式(6)中，K 为算法最大迭代次数，k 表示当前累计的迭代次数。收缩指数 $\alpha_{e,p}$ 表示当新生波段继承自精英组时选取 α_e，当新生波段继承自普通组时选取 α_p。

1.1.4　更新规则

若候选波段 x_c 的评价值高于精英组末位 x_{ewst} 的评价值，则 x_c 进入精英组，同时精英组再次进行评价值排序，精英组末位 x_{ewst} 不进入普通组直接淘汰。这是因为 TPA 设置了学习行为，加强了算法的定向搜索和局部搜索能力，遭到淘汰的精英组末位 x_{ewst} 如果进入普通组的话容易导致算法陷入局部最优。若候选波段 x_c 的评价值劣于 x_{ewst} 但优于普通组末位 x_{pwst}，还需检查 x_c 是否由探索行为得到，若是，则 x_c 进入普通组，淘汰 x_{pwst}。若不是，直接丢弃 x_c。这是因为学习行为产生高评价值候选波段的可能性比较大，且趋同性强，容易导致普通组波段同化，降低全局寻优能力。

1.2　改进 TPA 算法

在 TPA 算法中，继承于精英组的新生波段 x_r 在选择学习行为时，需要向普通组样板 e_p 的反方向调节。但在迭代过程中，精英组和普通组的成员在不断进行更新优化，从而在迭代后期，普通组的样板值 e_p 也随着精英组样板值增大，在选择学习行为时，并不能为继承于精英组的新生波段 x_r 提供一个良好的调节方向，容易陷入局部最优。因此对 TPA 算法进行如下改进：

1.2.1　改进分组规则

在对 P 个成员进行分组时，由原来的 $N+M$ 模式改为 $N+M+L$ 模式，评价值依次降序排列。N 为评价值高的波段组成的精英组$\{x_{e1}, x_{e2}, \cdots, x_{en}\}$，$M$ 为评价值适中的波段组成的普通组$\{x_{p1}, x_{p2}, \cdots, x_{pm}\}$，新增添的 L 组为评价值最低的波段组成的垃圾回收组$\{x_{g1}, x_{g2}, \cdots, x_{gl}\}$。再通过团队进步过程更新成员。

1.2.2　学习行为重定义

新增添的 L 组不参与继承、学习以及探索行为。新生波

段进行学习行为时，仍需要向参照目标方向调节，参照方向更改为分别产生于精英组和垃圾回收组，称为精英组样板 e_e 和垃圾回收组样板 e_g，样板值依旧取所在组波段波长的平均值。即若新生波段产生于普通组，则其需向精英组样板调节；若新生波段产生于精英组，就向垃圾回收组样板的反方向调节，如式(7)所示。

$$\begin{cases} x_c = (1-\gamma)x_r + \gamma e_e \\ x_c = (1+\gamma)x_r - \gamma e_g \end{cases} \tag{7}$$

式(7)中，γ 为区间[0，1]内随机数。若 x_c 某个波长点范围越界，则改用其边界值。

1.2.3 *更新规则的修改*

在进行波段更新的时候需要将低评价值的波段回收进 L 组，使 L 组的波段评价值极低。新的波段更新规则在精英组和普通组不做修改，当候选波段 x_c 的评价值低于垃圾回收组首位(垃圾回收组中评价值最高)x_{gbst} 的评价值，则 x_c 进入垃圾回收组，同时垃圾回收组中的 x_{gbst} 波段遭到淘汰，促使垃圾回收组评价值一直保持低的状态，为继承于精英组的新生波段进行学习行为时提供一个正确的更新方向，有效地避免算法陷入局部最优。

1.3 iTPA 算法实现步骤

Step1：初始化基本参数。包含总波段成员个数，精英组、普通组以及垃圾回收组的波段个数，新生波段选择学习或探索行为的概率，精英组和普通组的收缩系数，迭代次数。

Step2：分组。将各波段按照与其相应的理化值 PLS 建模得到的评价值函数大小降序排列，评价值函数如式(2)所示，依次放入精英组、普通组和垃圾回收组，使得精英组中评价值最高，普通组次之，垃圾回收组最低。

Step3：产生新生波段。新生波段随机选择从精英组或普通组中产生，其波长点从当前组随机波段同一波长点中继承。

Step4：新生波段选择学习或探索行为生成候选波段。生成一个随机数，若随机数满足选择学习行为的概率，则需进行学习行为。在进行学习行为之前，需判断该新生波段是从精英组还是普通组产生，以便选择不同的学习行为，如式(7)所示。若随机数满足选择探索行为的概率，则需进行探索行为。在进行探索之前，仍需判断该新生波段是从精英组还是普通组产生，从而为探索行为选择不同的收缩系数，如式(5)和式(6)所示。若行为过程中某个波长点范围越界，则改用其边界值。

Step5：波段更新。候选波段的评价值需要跟三个小组中的波段评价值进行对比更新。更新规则参阅本文 1.1.4 和 1.2.3。

Step6：循环 Step3—Step5 进行迭代更新。迭代结束，选出精英组中评价值最高的波段为最终筛选波段。

2 实验部分

2.1 数据准备

2.1.1 *数据来源*

为考察新提出的变量筛选算法对建模预测的效果，将其应用在一组标准玉米近红外光谱数据集。该光谱数据集引用自 eigenvector 网站上开源的玉米样本光谱数据集，网址 https://eigenvector.com/resources/data-sets/。该数据集为 80 个玉米样品用 mp5spec 仪器扫描得到的光谱数据，并用化学方法测定了其淀粉和蛋白质含量。以间隔为 2 nm 在波长范围为 1 100～2 498 nm 上收集(700 个波长点)。

2.1.2 *剔除异常数据及样本划分*

考虑到仪器测量光谱数据时因误差而得到异常光谱，会影响模型性能，故先用马氏距离剔除光谱中异常数据。图 1 为各个数据样本点到数据中心点的马氏距离分布图。马氏距离最远的两个样本(75 和 77 号样本)作为异常点剔除，图 2 为剔除异常样本之后的原始光谱图。

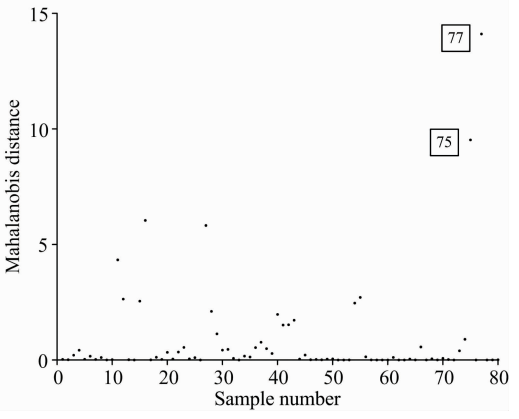

图 1 样本马氏距离分布图

Fig. 1 Mahalanobis distance distribution of samples

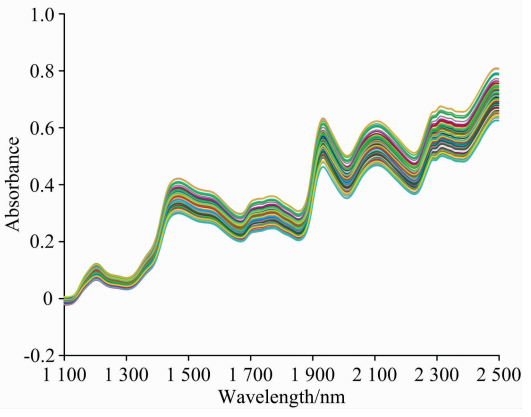

图 2 剔除异常样本的光谱图

Fig. 2 Spectra of abnormal samples removed

采用 Kennard-Stone(KS)方法[18]将剩余的 78 个样本点分成校正集和预测集。划分结果为校正集样本 50 个，预测集样本为 28 个。校正集和预测集淀粉和蛋白质含量值统计如表 1。由表 1 可知，校正集样本与预测集样本的平均值和标准差相差不大，通过 KS 方法划分数据集保证了校正集样本均匀分布。

2.2 方法

实验使用的是一台戴尔计算机，处理器是 Intel(R) Core i5-9400，CPU 主频为 2.90 GHz，操作系统为 Windows10，

表 1　校正集和预测集中淀粉和蛋白质含量值 g/100 g 统计

Table 1　Statistics of starch and protein contents in correction set and prediction set

	最大值	最小值	平均值	标准差
淀粉校正集	66.472 0	62.826 0	64.809 9	0.799 7
淀粉预测集	65.841 0	63.099 0	64.432 3	0.813 3
蛋白质校正集	9.595 0	7.654 0	8.569 2	0.485 2
蛋白质预测集	9.711 0	8.112 0	8.848 8	0.498 0

所有计算均在 MATLAB 2016a 中进行。为了验证 iTPA 算法的有效性和优越性，分别用全谱波长、GA、TPA 以及主成分分析(principal component analysis，PCA)算法对玉米的淀粉和蛋白质含量的建模效果进行了对比。进行 GA 算法时，由于原始光谱波长点数众多，如对波长点进行优选组合，运算效率将会非常低，因此将原光谱均分若干个子区间，用 GA 算法进化迭代获取最大适应度值所对应的优选子区间组合。根据基因选出的波段建立 PLS 模型，计算出模型的相关系数(r)和校正集预测均方根误差(RMSEC)，适应度函数 F 跟 iTPA 算法评价值函数 $f(x)$保持一致为

$$F = \frac{r}{1 + \mathrm{RMSEC}} \tag{8}$$

将原光谱在 1 100～2 498 nm 范围之间共 700 个波长点数划分 35 个等距区间，即遗传编码长度为 35，每一个基因包含 20 个波长点数。设定群体个数为 50 个，交叉概率为 0.85，变异概率为 0.1，最大迭代数为 200 代。在种群进化过程中寻找最大迭代次数内进化过程中最优适应度个体。

同样的，采用 TPA 算法时将原光谱等分 35 个波段区间，每个波段内含 20 个波长点数，即每个成员对应 20 个能力因素。按照算法经验，精英组由 15 个波段组成，普通组由 20 个波段组成，该分组模式算法性能最优。设定选择学习概率为 0.35，循环尝试不同的收缩指数，当精英组收缩指数为 20 时，算法预测效果最优，普通组收缩指数设定为精英组的 0.5 倍。迭代次数为 1 000。迭代完毕之后，取精英组评价值最高的成员为优选成员，优选成员中包含的能力因素就是所要筛选的波长点。如出现重复波长点则去除即可。因此，采用 TPA 算法筛选出来的波长点数最多不超过 20 个。

对于 iTPA 算法，同样分为 35 个成员，按照算法经验，其中精英组 10 个成员，普通组 10 个成员，垃圾回收组 15 个成员，该分组模式算法效果最优，其余设定条件与原 TPA 算法保持一致。由于 GA 算法、TPA 算法、iTPA 算法都具有的随机性，因此将以上三种算法分别运行 50 次求平均。

3　结果与讨论

3.1　算法性能分析

改进算法 iTPA 优化了 TPA 算法的更新方向，促使算法很大程度上避免陷入局部最优。图 3 为用 TPA 和 iTPA 筛选之后的变量分别与淀粉和蛋白质含量测试 50 次进行建模分析的预测均方根误差值。

由图 3 可知，iTPA 算法由于避免了陷入局部最优，使得整体预测效果得到了明显的提升。表 2 和表 3 分别为各算法测试淀粉含量和蛋白质含量的各项性能数据。

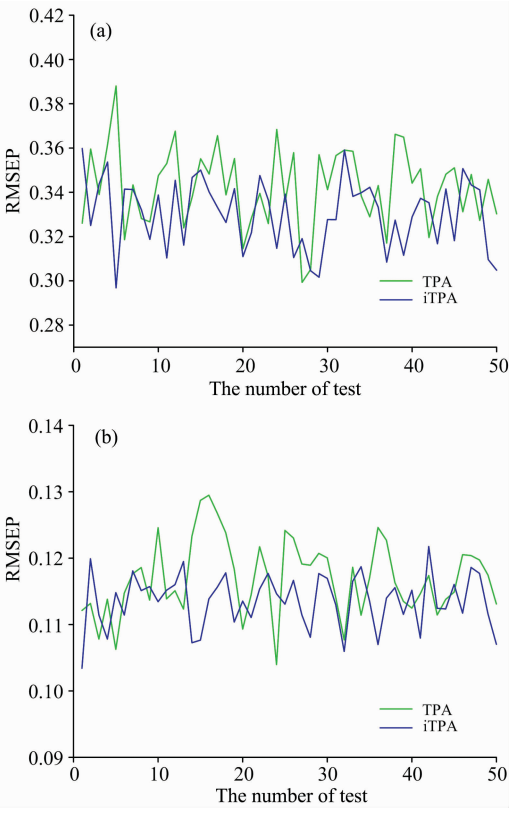

图 3　TPA 和 iTPA 算法运行 50 次(a)淀粉预测 RMSEP 值和(b)蛋白质预测 RMSEP 值

Fig. 3　Prediction RMSEP values of starch (a) and protein (b) after running TPA and iTPA for 50 times

表 2　不同变量筛选方法对玉米淀粉含量的预测结果

Table 2　Prediction results of corn starch content by different variable screening methods

	波长数	R_c	RMSEC	R_p	RMSEP	T/s
F-PLS	700	0.905 7	0.335 7	0.871 7	0.391 4	0.028
GA-PLS	350	0.958 2	0.226 4	0.926 2	0.300 6	21.980
TPA-PLS	17.64	0.945 3	0.258 1	0.901 3	0.345 3	0.680
iTPA-PLS	17.55	0.944 1	0.260 9	0.907 9	0.334 4	0.676
PCA	20	0.899 6	0.377 0	0.792 7	0.438 3	0.467

从表 2 可得，在玉米淀粉含量的预测上，iTPA 相比于全谱 PLS(F-PLS)，变量个数从 700 个减少到均值为 17.55 个(50 次运算求平均)，模型的校正均方根误差 RMSEC 从 0.335 7 降到 0.260 9 左右，校正集预测精度提升 22.3%。模型的预测均方根误差 RMSEP 从 0.391 4 下降到 0.334 4 左右，预测集预测精度提升 14.6%。相比原 TPA 算法有了提升，TPA 算法的 RMSEP 为 0.345 3。GA 算法总体预测效果是最佳，但其筛选出来的平均波长点数高达 350 个和算法运算时间 21.98 s，远远大于 iTPA 算法的 17.55 个和 0.676 s。通过对原光谱进行 PCA 降维，筛选出 20 个波长点(贡献率

已达 0.999 98)，以此与筛选出近乎相同波长数的 iTPA 算法作比较。经 PCA 算法筛选波长之后得到的 RMSEP 为 0.792 7，预测效果远不如 iTPA 算法。因此，iTPA 算法应用在淀粉含量光谱数据集中能在保持预测能力的前提下大幅度削减波长点数，有效地减小建模的计算量，同时算法速度更快。

表 3 不同变量筛选方法对玉米蛋白质含量的预测结果

Table 3 Prediction results of corn protein content by different variable screening methods

	波长数	R_c	RMSEC	R_p	RMSEP	T/s
F-PLS	700	0.951 7	0.147 4	0.930 7	0.178 9	0.028
GA-PLS	346	0.979 1	0.097 8	0.973 5	0.111 4	22.120
TPA-PLS	19.64	0.975 5	0.105 6	0.969 6	0.119 3	0.678
iTPA-PLS	19.60	0.975 9	0.101 9	0.970 4	0.117 7	0.666
PCA	20	0.956 9	0.141 7	0.891 6	0.236 2	0.469

从表 3 可得，对于玉米蛋白质的预测，iTPA 算法预测均方根误差 RMSEP 为 0.117 7，预测相关系数 R_p 为 0.970 4，校正均方根误差 RMSEC 为 0.101 9，校正相关系数 R_c 为 0.975 9，预测效果比全谱 PLS 和原 TPA 算法都有提升，全谱 PLS 的 RMSEP 为 0.178 9，TPA 算法的 RMSEP 为 0.119 3。GA 算法总体预测效果最佳，但其筛选出来的平均波长点数高达 346 个，以及算法运算时间 22.12 s，远远大于 iTPA 的 19.60 个和 0.666 s。通过对原光谱进行 PCA 降维，筛选出 20 个波长点(贡献率已达 0.999 997)。经 PCA 算法筛选波长之后得到的 RMSEP 为 0.236 2，预测效果远不及 iTPA 算法。因此，iTPA 算法应用在蛋白质含量光谱数据集中，同样能在保持预测能力的前提下大幅度削减波长点数，有效地减小建模的计算量，同时算法速度更快。

3.2 化学特性分析

图 4 表示 iTPA 算法分别在玉米淀粉和蛋白质数据集上运行 50 次后变量被选取的频率。

由图 4(a)可知，淀粉近红外光谱被 iTPA 算法筛选后的信息变量区域主要分布在 1 540～1 546，1 576～1 588，1 724～1 730 和 1 766～1 772 nm 等区域，而这些区域与淀粉中 O—H 的伸缩振动一级倍频以及 C—H 的伸缩振动一级倍频的频率一致，这与本次研究中淀粉的化学性质相一致。

由图 4(b)可知，蛋白质近红外光谱被 iTPA 算法筛选后的信息变量区域主要分布在 1 920，1 958～1 962，2 050，2 106～2 110，2 180～2 182 和 2 242～2 244 nm 等区域，而这些区域与蛋白质中 C═O 的伸缩振动一级倍频、N—H 键对称、不对称、伸缩振动以及 Amide Ⅰ(Ⅱ，Ⅲ)(分别为酰胺分子中羰基与胺基不同的耦合方式)的频率一致。同时筛选的变量也有分布在 1 700 nm 附近，这些区域主要对应着 C—H 伸缩振动一级倍频，这是因为蛋白质中含有 C 元素，这与本次研究中蛋白质的化学性质相一致。

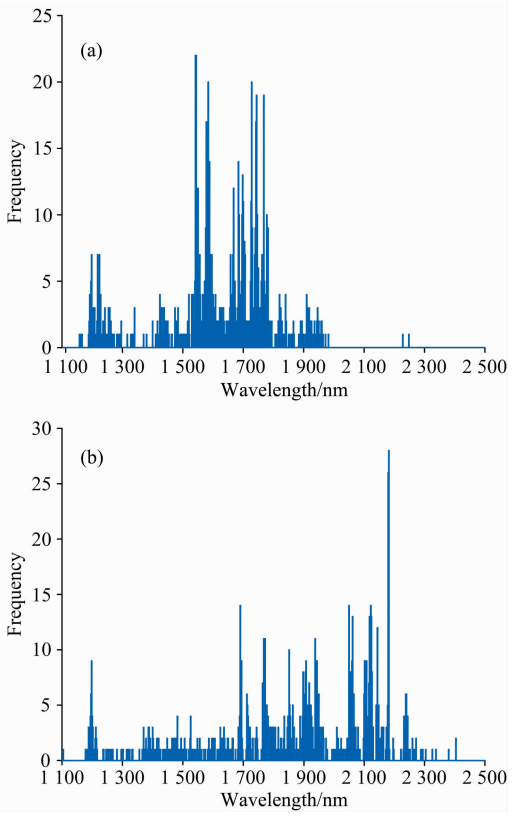

图 4 iTPA 算法运行 50 次后(a)淀粉光谱变量被选取的频率和(b)蛋白质光谱变量被选取的频率

Fig. 4 Frequencies of spectral variable selections for starch (a) and protein (b) after running iTPA for 50 times

因此 iTPA 算法能够有效地消除光谱信息中的一些干扰波长点，达到筛选出有效波长点目的。

4 结 论

提出了一种近红外光谱波长变量选择算法，即改进的团队进步算法(iTPA)，对玉米淀粉和蛋白质的近红外谱波长进行选择，建立了更加稳健的 PLS 模型，模型的预测均方根误差 RMSEP 分别为 0.334 4 和 0.117 7，获得了满意的预测精度。结果表明：iTPA 与其他波长选择算法相比性能均有一定优化，特别是筛选出的有效波长数目最少，达到 20 个以下，从而降低了模型的复杂度；快速的寻优能力也有利于近红外光谱检测在工业现场的实时应用。

References

[1] Li Cheng, Zhao Tianlun, Li Cong, et al. Food Chemistry, 2017, 221: 990.
[2] Zhang Hua, Li Changcheng, Huang Jian, et al. Polish Journal of Environmental Studies, 2018, 27(4): 1859.
[3] Liu Haojie, Li Minzan, Zhang Junyi, et al. International Journal of Agricultural and Biological Engineering, 2019, 12(5): 149.
[4] QU Fang-fang, REN Dong, HOU Jin-jian, et al(瞿芳芳, 任 东, 侯金健, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(2): 593.
[5] Yun Yonghuan, Li Hongdong, Wood Leslie R E, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013, 111: 31.
[6] Tang Rongnian, Chen Xupeng, Li Chuang. Applied Spectroscopy, 2018, 72(5): 740.
[7] Deng Baichuan, Yun Yonghuan, Liang Yizeng, et al. Analyst, 2014, 139(19): 4836.
[8] ZHAO Huan, HUAN Ke-wei, ZHENG Feng, et al(赵 环, 宦克为, 郑 峰, 等). Journal of Changchun University of Science and Technology · Natural Science Edition(长春理工大学学报·自然科学版), 2016, 39(5): 51.
[9] Liu Haojie, Li Minzan, Zhang Junyi, et al. International Journal of Agricultural and Biological Engineering, 2019, 12(5): 149.
[10] HUANG Chang-yi, FAN Hai-bin, LIU Fei, et al(黄常毅, 范海滨, 刘 飞, 等). Journal of Instrumental Analysis(分析测试学报), 2014, 33(5): 520.
[11] JIANG Shui-quan, SUN Tong(江水泉, 孙 通). Food & Machinery(食品与机械), 2020, 36(2): 89.
[12] Zhao Xin, Zhu Qibing, Huang Min, et al. Analytical Methods, 2013, 5(18): 4811.
[13] Hu Leqian, Yin Chunling, Ma Shuai, et al. Food Analytical Methods, 2019, 12(3): 633.
[14] Mojtaba Shamsipur, Vali Zare-Shahabadi, Bahram Hemmateenejad, et al. Journal of Chemometrics, 2006, 20(3-4): 146.
[15] Chen Yuanyuan, Wang Zhibin. Molecules, 2019, 24(421): 1.
[16] BAO Yi, GAO Mei-feng(包 怡, 高美凤). Computer Engineering and Application(计算机工程与应用), 2010, 46(25): 171.
[17] Mokhtari Hadi. Journal of Intelligent & Fuzzy Systems, 2016, 31(1): 487.
[18] Kennard R W, Stone L A. Technometrics, 1969, 11(1): 137.

Near Infrared Spectral Wavelength Selection Based on Improved Team Progress Algorithm

GAO Mei-feng, TAO Huan-ming

Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

Abstract Aiming at the problem of near-infrared spectroscopy wavelength selection, an improved team progress algorithm (iTPA) is proposed based on the team progress algorithm (TPA). The bands of molecular spectrum are arranged in descending order according to the evaluation value function obtained by modeling corresponding physical and chemical values and are divided into elite group, plain group and garbage collection group. When the new wave band selects learning behavior, if it is generated in the plain group, it needs to adjust to the direction of the elite group template; if it is generated in the elite group, its updating direction needs to be improved to adjust to the reverse direction of garbage collection group template. Unlike the elite group and the plain group, members' evaluation value of the garbage collection group is always in a deficient state, which provides an accurate update direction for the new band generated from the elite group during the learning procedure to improve the global optimization ability of the algorithm. Through continuous iterative updating, the overall evaluation value is gradually improved, and finally, the band with the highest evaluation value is selected as the screening band. The algorithm is tested on the data set of corn starch and protein content and compared with TPA, genetic algorithm (GA), principal component analysis (PCA) and complete spectrum method. The experimental results show that the proposed algorithm can find the optimal combination of wavelengths in the whole spectrum range and explain each component's chemical characteristics. Compared with the full spectrum, for the corn starch data set, the number of variables of iTAP was decreased from 700 to 17.55 (averaged by 50 tests), RMSEC of the model was reduced from 0.335 7 to 0.260 9, and the prediction accuracy of the correction set was improved by 22.3%. The RMSEP of the model decreased from 0.391 4 to 0.334 4, and the prediction accuracy of the prediction set increased by 14.6%; For the corn protein dataset, the number of variables decreased from 700 to 19.6 (averaged by 50

tests), RMSEC of the model was reduced from 0.147 4 to 0.101 9, and the prediction accuracy of correction set was improved by 30.1%. The RMSEP of the model decreased from 0.178 9 to 0.117 7, and the prediction accuracy of the prediction set increased by 34.2%.

Keywords Near infrared spectrum; Wavelength selection; Improved team progress algorithm; Intelligent combination optimization; Characteristic wavelength

(Received Aug. 28, 2020; accepted Dec. 12, 2020)