第 41 卷 第 15 期/2021 年 8 月/光学学报 研究论文

光学学报

基于随机蛙跳算法的特征波长优选

撒继铭[1,2,3]，江河[1*]，谢凯文[1]，顾瀚文[1]，罗怡杰[2]，张朱珊莹[3,4,5**]

[1]武汉理工大学信息工程学院，湖北 武汉 430070；

[2]宽带无线通信与传感器网络湖北省重点实验室，湖北 武汉 430070；

[3]医学信息分析及肿瘤诊疗湖北省重点实验室，湖北 武汉 430074；

[4]中南民族大学生物医学工程学院，湖北 武汉 430074；

[5]中南民族大学认知科学国家民委重点实验室，湖北 武汉 430074

摘要 在随机蛙跳算法的基础上，提出了一种改进的窗口随机蛙跳算法，即使用连续窗口代替单个波长点，所提算法提高了原蛙跳算法的寻优精度，降低了随机蛙跳算法的迭代次数，从而提高了算法的收敛速度。血液样本的实验结果表明，相较于全波段，建立的偏最小二乘法模型的预测均方根误差下降了 47.9%，预测集相关系数 R_p 提高了 4.07%，证明了所提算法的有效性。通过对 3 种主流算法及改进随机蛙跳算法筛选出的特征波长建立回归分析模型，证明了改进算法在特征波长选择上的优越性。

关键词 光谱学；红外光谱；随机蛙跳算法；定量分析；特征波长选择

中图分类号 TN929.11 **文献标志码** A **doi**：10.3788/AOS202141.1530001

Characteristic Wavelength Optimization Based on Random Frog Algorithm

Sa Jiming[1,2,3], Jiang He[1*], Xie Kaiwen[1], Gu Hanwen[1], Luo Yijie[2], Zhang Zhushanying[3,4,5**]

[1] *School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei* 430070, *China*;

[2] *Hubei Key Laboratory of Broadband Wireless Communication and Sensor Networks, Wuhan, Hubei* 430070, *China*;

[3] *Hubei Key Laboratory of Medical Information Analysis and Tumor Diagnosis & Treatment, Wuhan, Hubei* 430074, *China*;

[4] *College of Biomedical Engineering, South-Central University for Nationalities, Wuhan, Hubei* 430074, *China*;

[5] *Key Laboratory of Cognitive Science (South-Central University for Nationalities), State Ethnic Affairs Commission, Wuhan, Hubei* 430074, *China*

Abstract Based on the random frog algorithm, a window-based random frog algorithm is proposed. With a continuous window instead of a single wavelength point, the proposed algorithm improves the optimization accuracy of the original random frog algorithm and reduces the iteration times of the algorithm, thus improving the convergence speed. The results of the blood samples show that compared with the full-band results, the root mean square error of prediction (RMSEP) of the built partial least square model decreases by 47.9%, and the correlation coefficient of the prediction set, R_p, increases by 4.07%, proving the validity of the proposed algorithm. The regression analysis of the characteristic wavelengths selected by the three mainstream algorithms and the window-based random frog algorithm is carried out, which demonstrates the superiority of the improved algorithm in selecting the characteristic wavelength.

收稿日期：2020-12-30；**修回日期**：2021-02-02；**录用日期**：2021-03-08

基金项目：国家自然科学基金(61501526，61178087)

通信作者：*hejiang2019@whut.edu.cn；**syzhu@mail.scuec.edu.cn

Key words spectroscopy; infrared spectroscopy; random frog algorithm; quantitative analysis; characteristic wavelength optimization

OCIS codes 300.6170; 150.1135; 200.4560; 040.3060

1 引 言

红外光谱是一种无损、可用于快速分析生物医学中蛋白质、葡萄糖、脂类和核酸等生物分子的指纹技术，其与多变量分析化学计量学结合，可为生物医学中的生理参量检测提供强大快速的分析方法[1]。但是，采用原始光谱进行定量分析，波长变量过多，导致算法迭代次数过多，影响算法的复杂性，运算速度慢，且未经处理的红外光谱谱峰宽，相邻谱峰之间重叠现象严重[2]，因此有效提取被测物质的相关特征波长信息，进行特征波长选择，关系到后续定量和定性分析的准确性，在红外光谱的定量分析中有着较大的研究意义。目前常用的波长优选方法有协同区间偏最小二乘法（SiPLS）[3]、连续投影算法（SPA）[4]、竞争自适应重采样算法（CARS）[5]等。

SiPLS 将光谱区间划分为若干等宽子区间[6]，每个子区间建立回归分析，在利用交互验证法来确定每个主因子数量的基础上，将交叉验证均方根误差（RMSECV）作为衡量标准，在精度最高的几个区间内筛选并进行多种组合，从而达到特征波长优选的目的。De 等[7]在利用红外光谱预测茶叶中多酚含量的研究中，使用偏最小二乘（PLS）算法生成粒子群优化（PSO）遗传算法的适应度曲线，用特定的波长窗口进行测试和培训，确定了波长的最佳范围。刘振尧等[8]采用移动窗口法进行特征波长优选，并对 492～890 nm 波段建立了血红蛋白含量的 PLS 定量模型，其预测集相关系数和均方根误差分别为 0.988 和 1.97%，预测效果较好。

SPA 算法在进行波长选择时，从一个波长点出发连续不断地采用投影策略筛选出与已有波长线性相关度最小的波长点，将这些点组成一个波长子集，并重复该过程直到指定数量，最终的波长组合即为线性关系最小的波长组合。李冠稳等[9]结合 SPA 算法从全波段光谱中筛选特征变量，并利用全波段和特征波段建立模型，从而提高了模型运算效率，且模型预测能力较全波段有所提高。

CARS 采用蒙特卡罗采样技术，结合 PLS 算法筛选出最佳的特征波长区间，不同波长通过权值的形式互相竞争，淘汰竞争力较弱的波长组合，体现“适者生存”的竞争原则。孙涛等[10]针对多金属离子混合溶液紫外可见光谱存在严重重叠和分离困难的问题，使用 CARS 算法，降低了波长选择的复杂性，并且保证了波长选择过程的稳定性。

上述研究并未考虑到光谱连续性和光谱区间的衔接性对定量分析的影响。随机蛙跳（RF）算法是近年提出的新型特征波长选择算法，其根据不同变量被选择的可能性不同，通过多次迭代来确定每个变量被选择的概率，进而选择概率高的变量作为特征波长。但该算法在运行过程中迭代次数 N 过大，导致算法复杂度高，收敛速度慢。针对 RF 算法的缺点，本文在 RF 算法的基础上，通过引入连续窗口的思想，得到改进的窗口随机蛙跳算法，该算法使得光谱的连续性和不同连续区间的衔接性得到优化，从而有效地挑选出能充分反映样本特征的波长变量。与 3 种主流特征波长优选算法进行仿真分析及结果对比，证明了改进算法在特征波长选择上的优越性。

2 特征波长优选

2.1 红外光谱数据的采集

1）光谱数据的采集

本实验选用日本岛津公司生产的 UV-VIS-NIR 分光光度计，光度计型号为 SolidSpec-3700。光谱仪的光谱范围为 200～2500 nm，是一款覆盖紫外、可见、近红外区域的光谱仪，采样间隔为 1 nm，总计有 2300 个光谱波长点，光谱分辨率为 0.1 nm，环境温度为 25 ℃，湿度为 40%，海拔 40 m，采集血红蛋白仿体溶液的吸收光谱。光谱的波长范围为 600～1800 nm。采集到的血红蛋白仿体溶液光谱如图 1 所示。

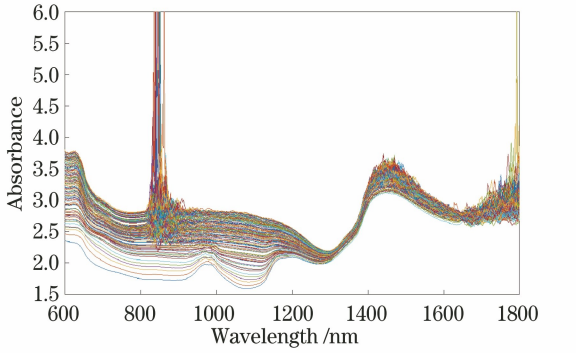

图 1 血红蛋白仿体溶液样本光谱

Fig. 1 Spectra of hemoglobin phantom solution sample

2）血液样本介绍

血液样本由 Karl Norris 提供，使用 NIRSystems 6500 光谱仪进行采集。样品池是一个直径为 2 cm 的不锈钢圆筒和石英窗口，近红外光谱的波长范围为 1100～2498 nm，采样间隔为 2 nm，总共包含 700 个波长点，选取 190 个样本的透射光谱数据作为本实验的血红蛋白含量分析样本，血液样本光谱如图 2 所示。

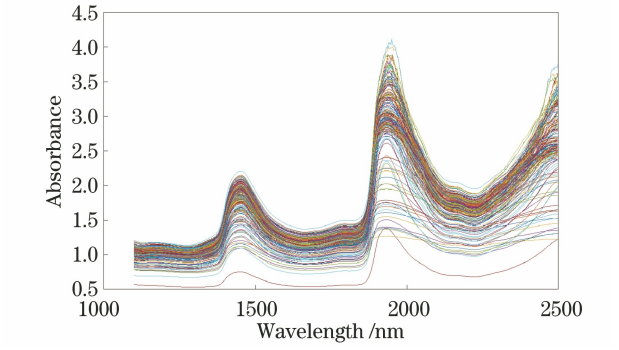

图 2 血液样本光谱

Fig. 2 Spectra of blood sample

2.2 样本集划分

红外光谱分析方法是否完善、可靠，依赖于样品集的划分，划分时应尽量保证样品集的随机性和代表性，一般校正集的参数包含预测集的参数范围，对于同一个样品来说，划分的校正集、预测集的平均值和标准偏差应当与全体集相近。本实验采用等间隔窗口法划分校正集和预测集，校正集和预测集的数量之比为 3∶1，该方法划分样本集分为如下两个步骤：

步骤 1 定义窗口大小为 4（可根据样本数量进行调整），将数据集划分为若干个连续的窗口大小为 4 的区间。

步骤 2 取出步骤 1 中的一个窗口，将前 3 个数据集划分为校正集，最后一个数据集划分为预测集，其他窗口类似。

146 个仿体溶液样本和 185 个血液样本（已剔除 5 个异常样本）的划分结果如表 1 所示，从划分结果可以看出，样本集的划分正常，覆盖范围分布均匀。

表 1 血红蛋白含量值统计

Table 1 Statistics of hemoglobin content

Sample	Sample set	Total number of samples	Mass concentration of hemoglobin /(g·L^{-1})			
			Maximum	Minimum	Average	Standard deviation
Phantom solution sample	Entire set	146	150	5	77.50	42.29
	Calibration set	110	150	5	77.34	42.53
	Validation set	36	148	8	78.00	42.14
Blood sample	Entire set	185	173	103	137.08	16.49
	Calibration set	139	173	103	137.95	16.40
	Validation set	46	173	106	134.46	16.66

2.3 所提算法

RF 算法是一种仿生类算法，它运用可逆跳跃马尔可夫链蒙特卡罗（RJMCMC）方法，在模型空间中构造一个 MCMC 链，当过渡概率满足建模所需的平衡条件时，得到的 MCMC 链将会收敛到模型指标和模型特定参数的联合分布中，使后续对变量数和模型参数的计算更简便。它集合了模因算法和粒子群优化算法的思想，通过对两者的优化和扩展，借鉴了两者的优点，因此它具有适者生存和随机搜寻两大突出特点[3]。对于 RF 算法，每只青蛙代表一个候选解，青蛙种群的初始化过程对应着初始位置的具体分布，种群的划分相当于组团搜索，之后进行局部搜索，此过程是 RF 算法的最关键步骤，这个过程对青蛙的具体位置进行了更新。全局信息的交换依赖于子群之间的混合，局部深度搜索的辅助为 RF 算法提供最后的局部极值。

RF 算法选择部分最小二乘线性判别构造分类器，兼具 SiPLS 算法对高度相关数据处理能力较强的特点；RF 算法集合了模因算法和粒子群优化算法的思想，因此具有适者生存和随机搜寻两大特点，兼具 CARS 算法降低波长选择复杂性的优势。

文献[3]中 RF 算法通过单个波长点实现特征波长的优选，考虑到光谱的连续性和不同的连续区间，本实验对 RF 算法进行了改进，引入连续窗口的思想，即窗口随机蛙跳算法（WRF），连续的窗口区间能够降低 RF 算法的迭代次数，从而加快算法的收敛速度。

WRF 算法需要设置迭代次数 N、变量子集中

变量数量 Q、窗口宽度 w。首先，在整个频谱上，通过固定宽度为 w 的移动窗口将频谱划分为子区间，并获得所有可能的区间，这些重叠的区间被视为变量。WRF 算法的流程如图 3 所示，其中 V_0 表示变量子集，V^* 表示候选变量子集，Q^* 表示候选变量子集的变量数量，i 表示迭代次数。

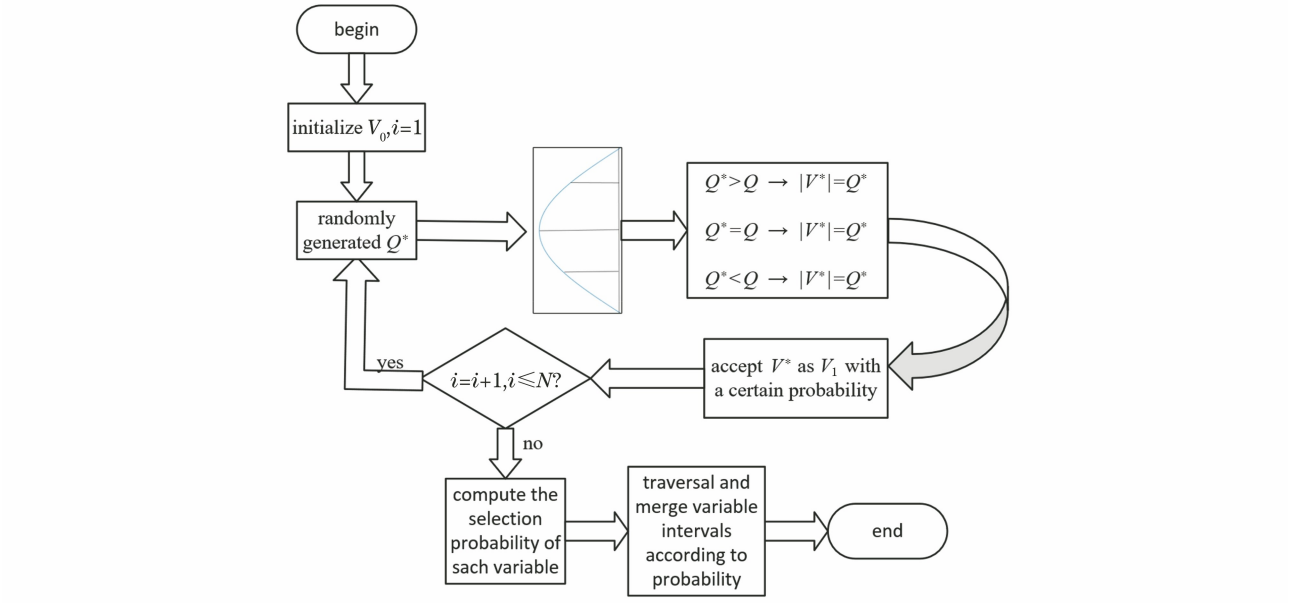

图 3 WRF 算法的流程

Fig. 3 Process of WRF algorithm

WRF 算法的主要运算步骤包括：1）初始化包含 Q 个变量的变量子集 V_0，根据均值为 Q 的正态分布范数随机生成 Q^*；2）比较 Q^* 和 Q 的大小，根据全部变量集和变量子集 V_0 生成包含 Q^* 个变量的候选变量子集 V^*；3）根据 PLS 算法比较 V^* 与 V_0 集合的预测误差，给 V^* 赋值相应的接收概率，并通过比较接收概率与 0～1 内的随机概率，决定是否将 V_0 更新为 V^*，若迭代次数超过 N 次，则退出循环；4）计算对变量的选择概率，根据概率对变量进行排序，以最高选择概率为起始，遍历宽度为 w 的窗口，合并子窗口，通过 RMSECV 值筛选出最优的合并区间。最终完成对高维波长变量的降维，得到的变量即为 WRF 算法筛选的特征波长变量。

3 结果与讨论

3.1 红外光谱分析算法评价指标

为了衡量 WRF 算法在各项任务上的性能，选择如下参数作为算法性能的评估指标。各指标的具体计算方法如下：

1）RMSECV，其计算公式为

$$f_{\mathrm{RMSECV}}=\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-1)}, \quad (1)$$

式中：n 表示测量的样本数量；y_i 表示第 i 个血红蛋白浓度的真实值；$\hat{y}_i$ 表示第 i 个血红蛋白浓度的测量值。RMSECV 值越小，性能越好。

2）相关系数 R，其计算公式为

$$R=\sqrt{1-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/\sum_{i=1}^{n}(y_i-\bar{y})^2}, \quad (2)$$

式中：$\bar{y}$ 表示所有血红蛋白浓度的真实值对应的平均值。相关系数 R 表征血红蛋白浓度的真实值和测量值之间的差异程度，其取值范围为[0，1]。所计算的相关系数 R 值越大，表明该算法测量得到的值越接近真实值，拟合能力越强。校正集和预测集的相关系数分别为 R_{c} 和 R_{p}。

3）预测均方根误差(RMSEP)，其计算公式为

$$f_{\mathrm{RMSEP}}=\sqrt{\sum_{i=1}^{n_{\mathrm{p}}}(y_{\mathrm{p}i}-\hat{y}_{\mathrm{p}i})^2/(n_{\mathrm{p}}-1)}, \quad (3)$$

式中：n_{p} 表示验证集的样本数量；$y_{\mathrm{p}i}$ 表示第 i 个血红蛋白浓度的真实值；$\hat{y}_{\mathrm{p}i}$ 表示第 i 个血红蛋白浓度的预测值。RMSEP 用来衡量验证集中血红蛋白浓度真实值和预测值之间的离散程度，RMSEP 差值越小，表明所建模型的预测能力越强。

3.2 仿体溶液样本结果分析

仿体溶液光谱吸收值在 880 nm 和 1800 nm 波长附近出现了明显的吸收饱和情况，因此，本实验选择 900～1750 nm 波长范围用于仿体溶液样本血红

蛋白定量分析。

1）基于 SiPLS 的特征波长优选结果

将光谱波长范围 901～1700 nm 划分为 100 nm 宽的 8 个区间，对每个区间初步进行 PLS 回归分析，8 个区间对应的 RMSECV 值和主因子数如图 4 所示。仿体溶液样本各区间的 RMSECV 的极差值较大，最优区间为第 4 区间，RMSECV 值为 2.8266，波长范围为 1201～1300 nm，占全波段的 11.8%。

2）SPA 特征波长优选结果

连续投影算法能够很好地消除数据的共线性问题，使用连续投影算法对特征波长优选时，需要对 Maximum 进行设置。为了找到最优的 Maximum，以 10 为间隔，对 Maximum 进行遍历，遍历范围为 10～70，并进行 PLS 回归分析，根据 RMSECV 值筛选出最优的 Maximum 值，遍历结果如表 2 所示。

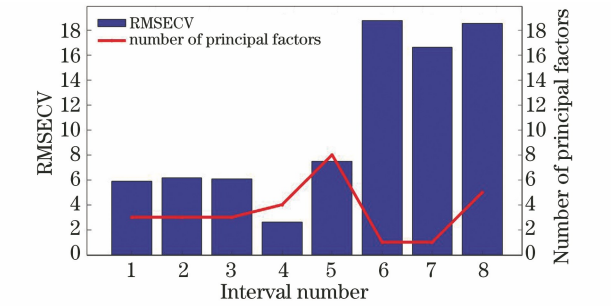

图 4　仿体溶液样本不同区间对应的 RMSECV 和主因子数

Fig. 4　RMSECV and number of principal factors corresponding to different intervals for phantom solution sample

表 2　连续投影算法参数优化结果

Table 2　Parameter optimization results of successive projections algorithm

Maximum	10	20	30	40	50	60	70
Number of wavelengths	2	2	4	5	5	5	5
RMSECV	4.3313	4.3313	2.2511	2.1494	2.1494	2.1494	2.1494

从表 2 可以看出，在 Maximum 达到 40 以后 RMSECV 和选择出来的波长数量保持不变，而随着 Maximum 的增加，算法的运算时间也会增加，因此选择 40 作为 Maximum 的最优参数。当 Maximum 取值为 40 时对应的波长数量为 5，RMSECV 有最小值 2.1494；入选波长在光谱中的位置如图 5 所示。

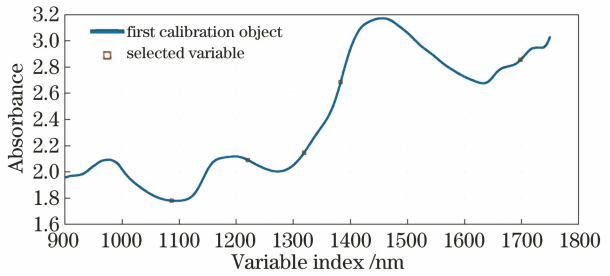

图 5　仿体溶液样本基于 SPA 特征波长分布

Fig. 5　Characteristic wavelength distribution based on SPA for phantom solution sample

3）CARS 特征波长优选结果

CARS 是基于“适者生存”原则的变量选择方法，采用十折交叉验证法，将采样总次数设置为 200 次，经特征波长优选得知，特征波长变量数随着采样次数的增加逐渐减少，最终接近 0，下降趋势符合衰减函数原理。基于 CARS 算法的仿体样本红外光谱特征波长优选过程如图 6 所示。根据波长变量回归系数的变化趋势，在 127 次采样时有最好的效果，此时对应的最小 RMSECV 值为 2.1420，筛选出仿体溶液在血红蛋白定量分析中的特征波长，一共选出 18 个特征波长（1096，1149，1182，1220，1222，1223，1224，1225，1226，1289，1290，1291，1292，1296，1297，1315，1345，1346 nm），主要分布在 1190～1350 nm 范围内。

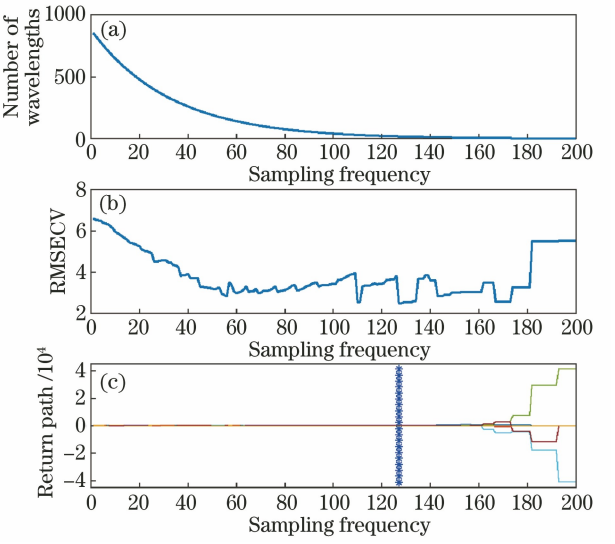

图 6　仿体溶液样本基于 CARS 特征波长优选过程。(a)波长数量；(b) RMSECV；(c)回归路径

Fig. 6　Selection process of characteristic wavelength based on CARS for phantom solution sample. (a) Number of wavelengths; (b) RMSECV; (c) return path

4）WRF 特征波长优选结果

仿体溶液光谱数据包括 850 个波长点，对仿体

溶液光谱数据调试后发现窗口宽度 w 设置在 2～10 范围内时有较好的预测效果。当 w 设置为 2 时，一共包含 849 个连续窗口，每个窗口包含 2 个波长点，可获得最好的处理效果。迭代次数 N 设置较大，有助于寻找最佳的光谱区间，综合考虑预测精度和计算成本，迭代次数 N 设置为 1000 次时可满足实验的要求，经过交叉验证后将初始变量集的变量数量 Q 设置为 10。选择概率较高的区间集中在中间窗口部分。

根据连续窗口的被选概率，对窗口波段进行合并后得到仿体溶液基于 WRF 特征波长优选结果，如图 7 所示。图 7(a)为第 3～849 个合并区间的 RMSECV 值，由于第 1、2 合并区间的波长数量过少，RMSECV 值过大，故不在图中展示。可以看到，在第 22 个合并区间有最小的 RMSECV 值(2.2962)，合并区间的波长数量为 40 个，随着合并区间波长数量的增加，RMSECV 值存在一定的波动，但是总体上带来了更多的无用信息。文献[3]中 RF 算法得到的 RMSECV 为 2.3984，而 WRF 算法的 RMSECV 降低了 0.1022，说明了 WRF 算法的有效性。图 7(b)为特征波长的分布图，大部分波长分布在 1200～1300 nm 范围内。

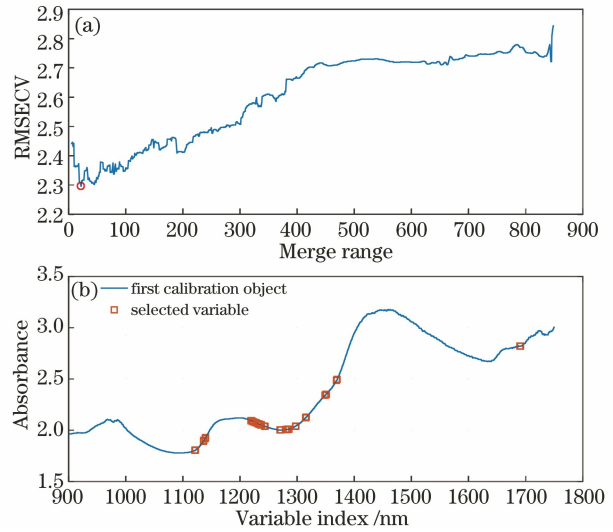

图 7 仿体溶液样本基于 WRF 特征波长优选结果。(a)合并区间的 RMSECV 值；(b)波长分布

Fig. 7 Optimization results of characteristic wavelengths based on WRF for phantom solution sample. (a) The RMSECV value of the merge intervals; (b) the wavelength distribution

综上所述，SiPLS、SPA、CARS、WRF 特征波长变量筛选数目分别为 100、5、18、40，其中 SiPLS 筛选出的特征变量最多，其他 3 类特征变量较少。对 4 种特征波长优选结果和全波段建立 PLS 定量分析方法。

4 种特征波长优选的 PLS 算法结果如表 3 所示，可以看出算法效果比较稳定，预测集的相关度都在 99%以上。相较于全波段谱，经过特征波长优选后，大部分算法的性能都有了提升，其中 CARS 算法在提取特征波长后，算法性能相对最优，RMSECV 减少了 1.3%，R_p 提高了 0.37%；WRF 算法的性能稍次于 CARS 算法，但是其主元个数为 4，相较于其他算法，复杂度最低。

表 3 仿体溶液特征波长优选结果

Table 3 Optimization results of characteristic wavelength of phantom solution sample

Optimization algorithm	Number of principal components	RMSECV	RMSEP	R_p
Full	9	3.3454	4.4792	0.9948
SiPLS	13	2.8266	3.8059	0.9960
SPA	6	2.1494	3.4661	0.9965
CARS	5	2.1420	2.1282	0.9985
WRF	4	2.2962	2.7742	0.9984

3.3 血液样本结果分析

血液样本定量分析时选择 1100～2498 nm 全波长，跨越 1400 nm 的波长范围，采样间隔为 2 nm，包括 700 个波长点。

1) SiPLS 特征波长优选结果

将光谱波长范围 1100～2498 nm 划分成宽为 200 nm、包含 100 个变量的 7 个波长区间，对每个区间初步建立 PLS 回归分析，7 个区间对应的 RMSECV 值和主因子数如图 8 所示。血液样本各个区间的 RMSECV 值区别较小，较优的 4 个区间为 1、3、4、6 区间，这 4 个区间对应的波长范围为 1100～1298 nm、1500～1698 nm、1700～1898 nm、2100～2298 nm。最优区间 6 的 RMSECV 值为

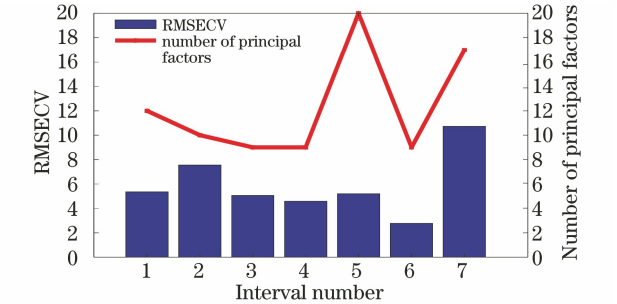

图 8 血液样本不同区间对应的 RMSECV 和主因子数

Fig. 8 RMSECV and number of principal factors corresponding to different intervals for blood sample

2.3614。

2）SPA 特征波长优选结果

当 Maximum 设置为 40 时，对应的波长数量为 17，RMSECV 有最小值 3.7454；入选波长在光谱中的位置如图 9 所示。可以看出，SPA 筛选的波长分别为 1100，1374，1388，1448，1508，1542，1578，1602，1824，1908，2012，2028，2092，2172，2220，2350，2450 nm，在 1100～2500 nm 范围内分布相对较均匀。

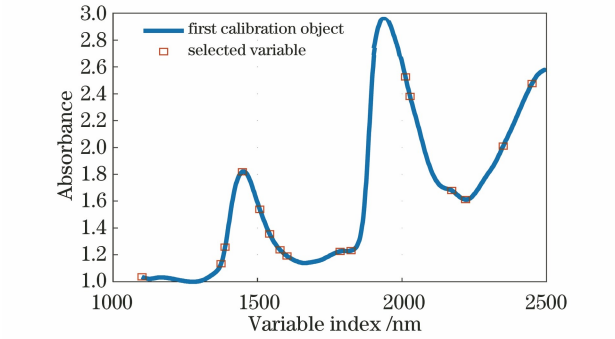

图 9 血液样本基于 SPA 特征波长分布

Fig. 9 Characteristic wavelength distribution based on SPA for blood sample

3）CARS 特征波长优选结果

采用十折交叉验证法，采样总次数设置为 200 次，基于 CARS 算法的血液样本红外光谱特征波长优选过程如图 10 所示。从图 10(a)所示的特征波长数量可以看出，特征波长数量随着采样次数的增加逐渐减少，最终接近 0。从图 10(b)所示的 RMSECV 值看出：在 20～60 次采样过程中，RMSECV 值出现波动；在 60～109 次采样过程中，RMSECV 值缓慢下降，说明此阶段剔除无用信息的效果较好，结合图 10(c)可以看出，109 次采样时得到最小的 RMSECV 值 5.1743，109 次采样后，RMSECV 值总体呈上升趋势，算法性能总体呈下降趋势。根据 109 次采样的结果，筛选出 29 个特征波段，主要分布在 1480，1870，2170 nm 3 个波长点附近。

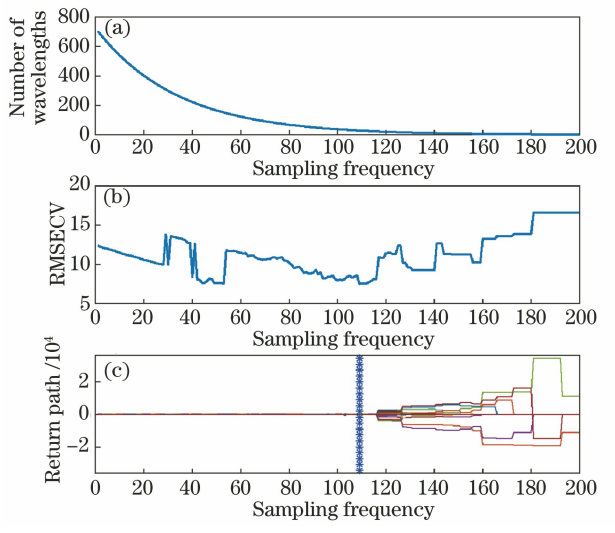

图 10 血液样本基于 CARS 特征波长优选过程。(a)波长数量；(b) RMSECV；(c)回归路径

Fig. 10 Selection process of characteristic wavelength based on CARS for blood sample. (a) Number of wavelengths; (b) RMSECV; (c) return path

4）WRF 特征波长优选结果

血液样本与仿体溶液样本的光谱特点类似，N、Q、w 参数分别设置为 1000、10、2，其中血液样本包含 700 个波长点，在 w 设置为 2 的情况下，连续窗口变量为 699 个。根据窗口波段的被选概率，对波段进行合并后得到血液样本基于 WRF 特征波长优选结果，如图 11 所示。图 11(a)为第 5～699 个合并区间的 RMSECV 值，前 4 个 RMSECV 值过大，在第 18 个合并区间有最小的 RMSECV 值(2.7972)，合并区间的波长数量为 31 个，随着合并区间变量数的增加，RMSECV 呈明显的上升趋势。图 11(b)为第 31 个波长的分布图，特征波长主要分布在 2100～2300 nm 范围内，特征波长点分布集中。

血液样本成分复杂，其包含的多种物质对血红

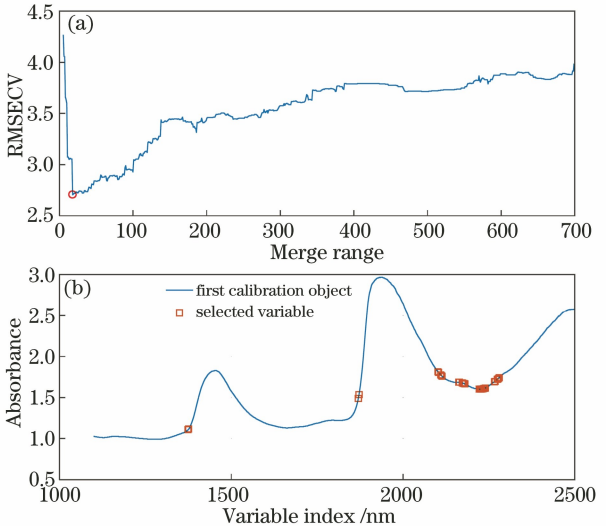

图 11 血液样本基于 WRF 特征波长优选结果。(a)合并区间的 RMSECV 值；(b)波长分布

Fig. 11 Optimization results of characteristic wavelengths based on WRF for blood sample. (a) The RMSECV value of the merge intervals; (b) the wavelength distribution

蛋白的测试会有较大的干扰，测试结果存在较多的未知因素。SiPLS、SPA、CARS、WRF 算法的特征波长变量筛选数量分别为 200、17、29、31。对 4 种特征波长优选结果和全波段建立 PLS 定量分析方法。

4 种特征波长优选算法对血液样本的优选结果如表 4 所示：全波长主元数量为 14，相对于其他四类特征波长优选方法偏高；经过 SiPLS、SPA、CARS、WRF 算法特征波长优选后定量分析效果都优于全波段分析效果，说明这 4 种特征波长优选方法在剔除无用信息的同时，较好地保留了有用信息，SiPLS 预测集的相关系数提高了 3.42%，但是特征变量数最多，选取了 200 个特征变量，占全波段变量总数的 28.5%，算法复杂性高。WRF 算法的精度优于其他三类特征波长优选方法，相较于全波段，其 RMSEP 下降了 47.9%，R_p 提高了 4.07%，主元数量为 9，算法复杂度最低。

表 4 血液样本特征波长优选结果

Table 4 Optimization results of characteristic wavelength of blood sample

Optimization algorithm	Number of principal components	RMSECV	RMSEP	R_p
Full	14	3.8142	5.3758	0.9421
SiPLS	10	2.3614	3.3581	0.9763
SPA	15	3.7454	2.4686	0.9707
CARS	10	5.1743	5.2252	0.9470
WRF	9	2.7972	2.3205	0.9828

文献[11]中利用 SG 平滑与 PLS 相结合的方法对血液中的血红蛋白进行定量分析；为了证明 WRF 算法的有效性，将 WRF 算法与文献[11]的方法、RF 算法使用相同的血液样本，对比结果如表 5 所示。可以看出，WRF 算法筛选出 31 个特征波长，在简化算法的基础上，定量分析效果优于 SG＋PLS 算法和 RF 算法，证明了 WRF 算法的有效性。

表 5 不同方法结果对比

Table 5 Result comparison of different algorithms

Algorithm	Characteristic wavelength number	RMSEP	R_p
SG＋PLS	700	5.45	0.940
RF＋PLS	34	3.35	0.974
WRF＋PLS	31	2.79	0.983

4 结 论

提出一种引入连续窗口思想的 RF 算法，通过减少特征波长选择过程中的主元数量，降低 RF 算法的迭代次数，提高算法的收敛速度。提取 WRF 算法选择的特征波长并建立 PLS 模型，实验结果表明，定量分析效果远优于全波段分析效果，证明了 WRF 算法的有效性。所提改进算法在降低 RF 算法复杂度的基础上，获得优于原算法的定量分析结果，算法精度明显提高。目前在进行定量分析时，仅能对一种成分进行分析，在未来研究中，将会提高分析方法的性能，同时对血红蛋白、葡萄糖等多种成分进行定量分析。

参 考 文 献

[1] Rohman A, Windarsih A, Lukitaningsih E, et al. The use of FTIR and Raman spectroscopy in combination with chemometrics for analysis of biomolecules in biomedical fluids: a review [J]. Biomedical Spectroscopy and Imaging, 2020, 8(3/4): 55-71.

[2] Zou T, Lan S M, Yan W, et al. Soybean protein wavelength optimization based on portable NIR spectrometer[J]. Analytical Instrumentation, 2019(3): 94-99.
邹涛, 兰树明, 阎巍, 等. 基于便携式近红外光谱仪的大豆蛋白波长优选[J]. 分析仪器, 2019(3): 94-99.

[3] Li H D, Xu Q S, Liang Y Z. Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification [J]. Analytica Chimica Acta, 2012, 740: 20-26.

[4] Li S Z, Tong H X, Yuan L M, et al. Optimization of near infrared spectroscopy model for sugar content in apple by intervals successive projection algorithm[J]. Journal of Food Safety & Quality, 2019, 10(14): 4608-4612.
李速专, 童何馨, 袁雷明, 等. 间隔连续投影算法应用于近红外光谱苹果糖度模型的优化[J]. 食品安全质量检测学报, 2019, 10(14): 4608-4612.

[5] Li H D, Liang Y Z, Xu Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. Analytica Chimica Acta, 2009, 648(1): 77-84.

[6] Wang L Q, Kong Q M, Li G B, et al. Characteristic band selection of near-infrared spectrum for determining peroxide value of oil based on iPLS[J]. Food Science, 2011, 32(9): 97-100.
王立琦, 孔庆明, 李贵滨, 等. 基于 iPLS 的油脂过氧化值近红外光谱特征波段选择[J]. 食品科学, 2011, 32(9): 97-100.

[7] De A, Chanda S, Tudu B P, et al. Wavelength selection for prediction of polyphenol content in inward tea leaves using NIR[C]//2017 IEEE 7th International Advance Computing Conference (IACC), January 5-7, 2017, Hyderabad. New York: IEEE Press, 2017: 184-187.

[8] Liu Z Y, Pan T. Equivalent waveband selection of VIS-NIR spectroscopic measurement for hemoglobin [J]. Optics and Precision Engineering, 2012, 20(10): 2170-2175.
刘振尧, 潘涛. 可见-近红外光谱测定血红蛋白的等效波段选择[J]. 光学 精密工程, 2012, 20(10): 2170-2175.

[9] Li G W, Gao X H, Xiao N W, et al. Estimation of soil organic matter content based on characteristic variable selection and regression methods[J]. Acta Optica Sinica, 2019, 39(9): 0930002.
李冠稳, 高小红, 肖能文, 等. 特征变量选择和回归方法相结合的土壤有机质含量估算[J]. 光学学报, 2019, 39(9): 0930002.

[10] Sun T, Yang C H, Zhu H Q, et al. A wavelength selection method of UV-Vis based on variable stability and credibility[J]. Spectroscopy and Spectral Analysis, 2019, 39(11): 3438-3445.
孙涛, 阳春华, 朱红求, 等. 一种基于变量稳定性和可信度的紫外-可见特征波长选择方法[J]. 光谱学与光谱分析, 2019, 39(11): 3438-3445.

[11] Pan T, Yan B R, Tang Y, et al. Combination optimization of PLS regression and SG smoothing in NIR analysis of hemoglobin [C]// International Conference on Photonics and Imaging in Biology and Medicine 2017, September 26-28, 2017, Suzhou China. Washington D.C.: OSA, 2017: W3A.75.