粒子群算法的近红外光谱定性分析预处理及特征提取参数优化方法研究

李浩光[1,2]，于云华[1,2]，逄　燕[1]，沈学锋[1,2]

1. 山东石油化工学院机械与控制工程学院，山东 东营　257061
2. 中国石油大学(华东)新能源学院，山东 东营　257061

摘　要　在近红外光谱定性分析时，为取得良好识别效果，预处理及特征提取是不可或缺环节。预处理主要是为消除各种干扰因素对光谱数据影响，常用预处理方法有平滑、一阶导、归一化等；而特征提取方法能剔除数据中的无关信息，保留有效信息，常用特征提取方法有偏最小二乘、主成分分析、线性判别分析等。不同预处理及特征提取方法具有不同特点，构建定性分析模型时，单一使用某种预处理或特征提取方法往往难以取得理想效果，常需将多种预处理及多种特征提取方法组合使用以提升模型性能。在各预处理及特征提取环节中往往有可变参数如特征提取维数等需要设定，这些可变参数对模型性能有重要影响，因此采用多个预处理及多个特征提取方法就存在多参数需要确定的问题。研究中常采用试凑法求各待定参数最优值，欲求得多个待定参数中某一个参数最优值，首先需据经验固定其他参数值，然后将某一个待优化参数代入近红外定性分析模型进行试凑，以求得模型最优识别率所对应参数值，并将其作为最优值。利用试凑法逐个求得多个待优化参数后，再将参数组合设置到定性分析模型中，最后进行定性鉴别，但试凑法求得的参数组合难以保证为全局最优解。除试凑法外，还可通过多重循环嵌套方法来获取近定性分析模型预处理与特征提取环节最优参数组合，但是该方法需消耗大量计算机内存与计算时间，而且效率低。为此，提出一种基于粒子群算法的近红外光谱定性分析模型预处理与特征提取参数优化方法，可快速获得预处理与特征提取环节的最优参数组合，并保证代入最优参数组合的定性分析模型具有最优识别性能，采用粒子群算法对平滑系数、一阶导系数、偏最小二乘特征提取维数等参数进行寻优，并将该方法与多重循环嵌套方法进行对比实验，实验结果证明了方法的有效性。

关键词　粒子群算法；特征提取；参数寻优；定性分析

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)09-2742-06

引　言

在近红外光谱定性分析时，为取得最优识别效果，首先需对原始光谱数据预处理，并进行特征提取。预处理主要目的是为消除样品自身与外界环境干扰因素对近红外光谱数据影响，常用预处理方法包括平滑、一阶导、归一化等[1-2]；而特征提取方法能够剔除近红外光谱数据中的无关信息，保留有效信息，常用特征提取方法有偏最小二乘(partial least squares，PLS)、主成分分析(principal component analysis，PCA)、线性判别分析(linear discriminant analysis，LDA)、正交线性判别分析(orthogonal linear discriminant analysis，OLDA)等[3-5]。不同预处理方法及特征提取方法具有不同特点，构建近红外定性分析数学模型时，单独使用某种预处理方法或特征提取方法往往难以取得理想结果，常将多种预处理方法及多种特征提取方法组合使用以提升模型性能，在各种预处理及特征提取方法中往往需要设定一些可变参数，这些可变参数对定性分析的性能有重要影响，因此采用多个预处理方法及多个特征提取方法就存在多个参数需要确定的问题。

常用的确定各参数的方法包括试凑法及多重循环嵌套寻优方法。试凑法求取各待定参数最优值时难以求得全局最优

收稿日期：2021-01-20，**修订日期**：2021-04-06

基金项目：国家重大仪器设备开发专项(2014YQ470377)，山东省教育厅科技计划项目(J18KA329)，东营市科技发展基金项目(DJ2020032)资助

作者简介：李浩光，1981年生，山东石油化工学院副教授　　e-mail：lihaoguang@upc.edu.cn

解；多重循环嵌套寻优方法需要消耗大量计算机内存与时间，存在效率低的缺点。为高效确定预处理环节及特征提取环节的多个待定参数，提出了一种基于粒子群算法的近红外光谱预处理及特征提取参数组合寻优方法，并以玉米籽粒单倍体二倍体光谱为例，对两种方法进行了实验验证。可以快速获得预处理与特征提取环节的最优参数组合，并保证代入最优参数组合的近红外定性分析模型具有最优的识别性能。

1　算法原理

粒子群算法(particle swarm optimization，PSO)首先由 Eberhart 博士和 Kennedy 博士在 1995 年提出，该算法是一种进化算法，从随机解出发，通过迭代搜寻最优解，具有较强的自适应能力及解决问题能力，在很多领域获得了成功应用[4]。

PSO 算法基于对鸟群觅食行为的模仿：鸟群在自然界随机搜寻食物时，若所在区域里只有一块食物，所有的鸟在搜索前均不知食物具体位置，但是鸟群可以判断感知当前位置与食物的距离，最有效食物搜索策略就是搜索当前离食物目标距离最近的鸟的附近空间。

用于参数组合寻优的 PSO 算法流程如图 1 所示。

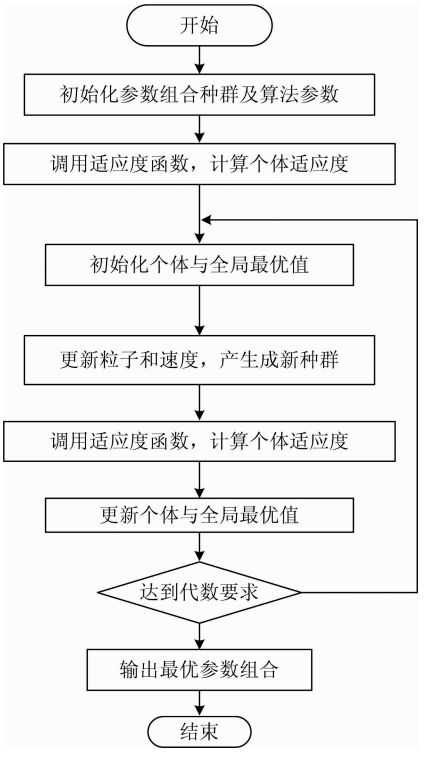

图 1　预处理及提取最优参数组合粒子群算法流程图

Fig. 1　Flow chart of Pretreatment and extraction parameters combination search based on PSO method

在参数寻优过程中，待求最优解等价于鸟类欲搜寻的食物，而鸟类觅食的搜索区域则对应于待求解问题的变量范围，在搜寻食物目标过程中，各个个体的鸟不仅需要自己的知识积累，还需根据整个鸟群的搜索经验来修正自己的速度，从而使自己不断靠近食物。

提出的基于粒子群算法的特征提取参数优化方法实现步骤如下：

(1)参数初始化：首先设定各待定参数变化范围，学习因子设定为 C_1 和 C_2，最大进化代数 G，kg 表示当前进化代数。在一个 D 维的搜索空间中，粒子种群规模为 size，每个粒子代表解空间的一个候选解。其中，第 $i(1<i<\text{size})$ 个粒子在解空间的位置表示为 X_i，速度表示为 V_i。第 i 个粒子从搜寻开始到当前迭代次数搜索得到最优解、个体极值 P_i、整个种群目前的最优解 BestS。初始化步骤随机生成 size 个粒子，并随机生成初始种群的位置及速度矩阵。

(2)个体评价(适应度评价)：将各粒子初始位置作为个体极值，计算群体中各粒子的初始适用值 $f(x_i)$，并求出种群最优位置，在本节中使用预处理、特征提取、分类器等环节构成近红外定性分析模型对实验数据所得的鉴别准确率作为适应度函数值。

(3)更新粒子速度和位置，生成新种群，并对粒子速度与位置进行越界检查。

$$V_i^{kg+1} = w(t)V_i^{kg} + c_1 r_1 (P_i^{kg} - X_i^{kg}) + c_2 r_2 (\text{BestS}_i^{kg} - X_i^{kg}) \tag{1}$$

$$X_i^{kg+1} = X_i^{kg} + V_i^{kg} \tag{2}$$

其中，$kg=1,2,\cdots,G$，$i=1,2,\cdots,\text{size}$，r_1 和 r_2 为 0 到 1 的随机数，c_1 为局部学习因子，c_2 为全局学习因子，一般取 c_2 为较大值。

(4)比较粒子当前适应值 $f(x_i)$ 与自身历史最优值 P_i，如果 $f(x_i)$ 优于 P_i，则设置为当前值 $f(x_i)$，并更新粒子位置。

(5)比较粒子当前适应 $f(x_i)$ 与种群最优值 BestS，如果 $f(x_i)$ 优于 BestS，则 BestS 为当前值 $f(x_i)$，更新种群全局最优值。

(6)判断粒子群算法收敛条件，若满足，则结束寻优，输出最优参数组合及其对应适应度函数值，否则 $kg=kg+1$，转至步骤(3)。结束条件一般是最大迭代次数或评价值小于设定精度。

2　算法设计

基于上述粒子群算法原理，未对该算法进行验证，选择如下近红外光谱数据集作为实验数据集：

以中国农业大学国家玉米改良中心提供的某品种玉米单倍体和二倍体籽粒作为研究对象，分 5 日连续采集其近红外光谱，使用自制近红外光谱采集装置[5-7]，并以漫透射采集方式交替采集单倍体、二倍体单籽粒近红外光谱各 100 条，共 5 组数据，5 个实验数据集按时间顺序依次编号为 T1—T5。

针对近红外光谱定性分析模型中预处理及特征提取参数优化问题对算法设计如下：

(1)适应度函数设计

适应度函数可用于评价粒子群算法所搜寻的各个参数组合的质量，根据适应度函数值的变化，进行迭代进化搜索粒

子最优值，并对粒子群中其他粒子状态进行更新，利用粒子适应度函数值能够反映粒子质量，即粒子是否能够使适应度函数取得最优值。拟进行参数优化的适应度函数模型如图 2 所示，在特征提取参数优化问题中将整个定性分析模型作为适应度函数，由图 2 可知，适应度函数由平滑、一阶导、归一化、PLS 特征提取、OLDA 特征提取、SVM 分类器等环节构成，其中待寻优参数有：平滑系数、一阶导系数、PLS 特征提取维数、OLDA 特征提取维数。

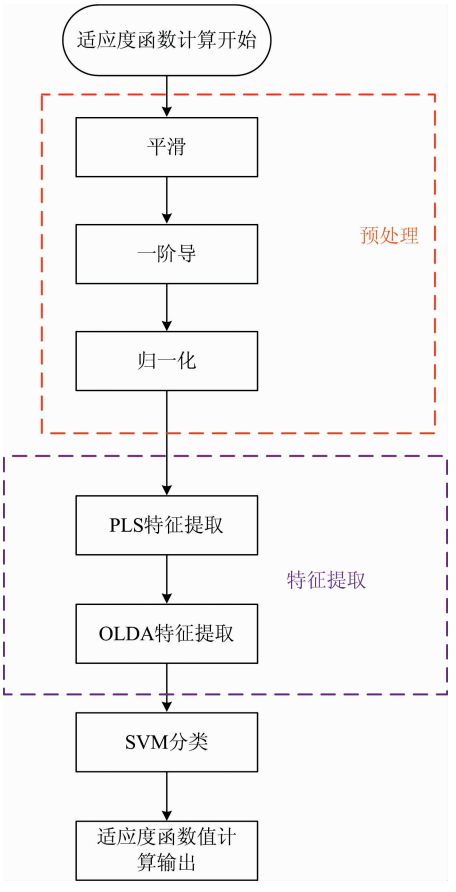

图 2 粒子群算法适应度函数流程图

Fig. 2 Flow chart of fitness function of PSO algorithm

(2)算法参数设计

粒子群算法相对于其他优化算法，其特点是需要设置参数相对较少，参数变化与算法寻优能力、收敛速度密切相关。

算法中需要确定的参数有：种群规模、粒子长度、粒子范围、粒子速度范围、加速度常数等。

①种群规模

图 3 是模型识别率随粒子种群规模变化曲线，由图 3 可以看出，种群规模，即同一批粒子的种群数量增加到 40 后便对识别率的影响很小，种群数目取 40 时，适应度函数值即分类器对待分类数据的识别率就能够达到 95%左右，其后识别率增长速度较慢。

图 4 是算法收敛速度随粒子种群规模变化曲线，由图 4 可以看出。随着粒子种群数目增大，粒子间相互配合能力随之增强，每个粒子负责搜寻空间相对变小，较大的粒子种群数更易搜索到全局最优解，但易带来负面问题，即算法运行时间直线上升。

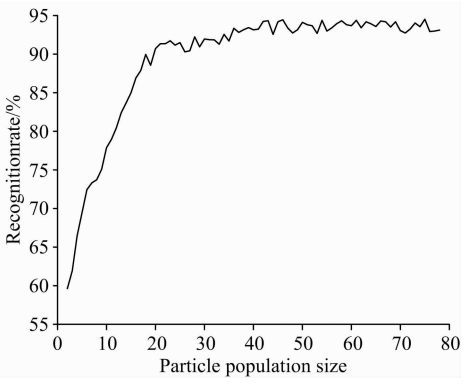

图 3 识别率随粒子种群规模变化曲线图

Fig. 3 Recognition rate curve with particle population size changing

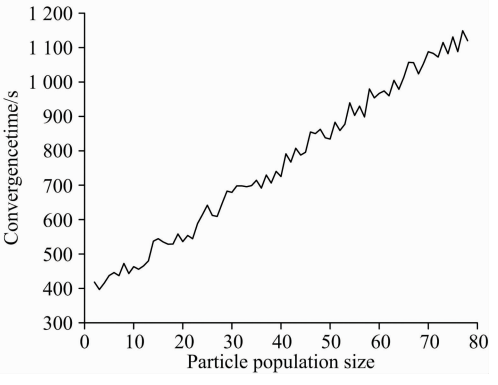

图 4 算法收敛速度随粒子种群规模变化曲线图

Fig. 4 Curve of convergence time with particle populationsize changing

综合考虑寻优能力与收敛时间，本节中粒子群种群规模适宜设置为 40。

②粒子长度

粒子长度即每个粒子所包含的待定参数的维数，对应本节中的适应度函数，本节中粒子的长度为 4 维。

③粒子范围

粒子范围指粒子在多维解空间中寻优区域，由具体优化问题与目标确定，一般将待优化参数取值范围固定为粒子范围，本研究中粒子的每一维搜索范围并不相同，具体设置如下：

平滑参数(smooth parameter)用 SP 表示，其最大值用 maxSP 表示，步长为 2，变化范围为[3，5，7，…，maxSP]；一阶导参数(FD parameter)用 FDP 表示，其最大值用 maxFDP 表示，步长为 2，变化范围为[3，5，7，…，maxFDP]；PLS 特征提取的维数(PLS parameter)用 PLS 表示，其最大值用 maxPLS 表示，步长为 1，变化范围为[3，4，5，…，maxPLS]；OLDA 特征提取后的维数(OLDA parameter)用 FDP 表示，因为其最大值不可能超过 PLS 特征提取后的维

数，因此其最大值用 PLS 表示。

④粒子速度范围

粒子飞行速度范围表示粒子搜索过程中单次运动距离，若粒子飞行的速度过高，粒子飞行跨度过大，易错过最优解。若飞行的速度太低，粒子只能在一个小的局部范围内进行搜索，可能搜寻到局部最优解。本节规定粒子飞行速度为 $k \cdot x\max$，其中 k 在 0.1 至 1 之间变化，$x\max$ 为各维粒子飞行速度的最大值，粒子每一维都采用相同取值方法。

⑤加速度常数 c_1，c_2

加速度常数代表粒子群算法中的学习因子，两值分别代表各粒子向个体极值与全局极值飞行时的加速度权重比值。较小的加速度值允许粒子在被拉回目标值前能够扩展搜索目标值范围之外的区域。加速度值设置过大则会导致粒子越过目标搜索范围。根据实际调试过程，本节中 PSO 算法加速度常数 c_1 为 1.48，c_2 设为 1.85。

基于上述分析，针对参数优化问题，粒子群算法参数设置如下：

粒子种群规模为 40，最大进化代数为 200，c_1 为 1.48，c_2 设为 1.85。

适应度函数：使用平滑、一阶导、归一化、PLS 特征提取、OLDA 特征提取、SVM 分类器等环节构成近红外定性分析模型，随机抽取 T1 数据集的一半作为训练集，另一半作为测试集，重复 20 次，将识别率取平均后作为适应度函数值。

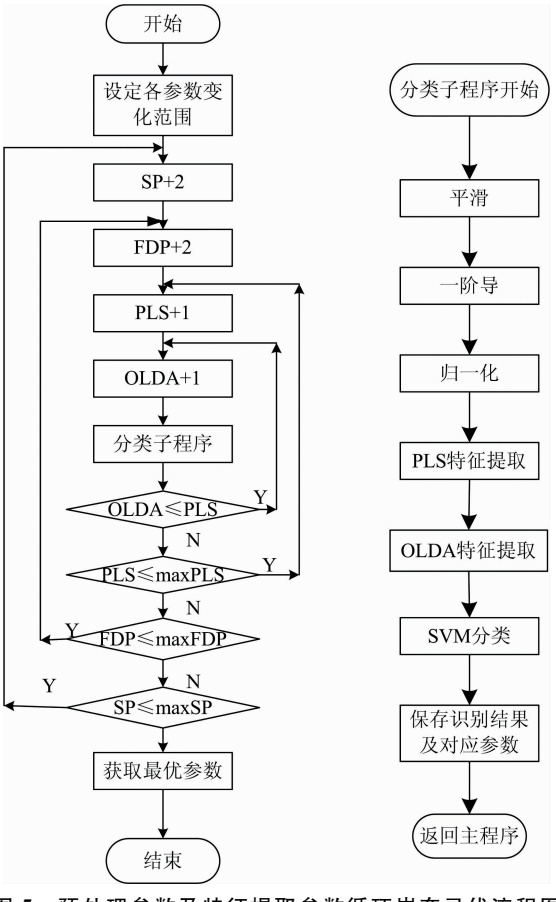

图 5　预处理参数及特征提取参数循环嵌套寻优流程图

Fig. 5　Flow chart of loop nesting optimization method

对比实验算法设计：循环嵌套方法与所提出的粒子群算法进行对比实验，循环嵌套方法的流程图如图 5 所示，程序利用 4 个循环嵌套实现 4 个待确定参数寻优。

图 5 中近红外定性鉴别子程序包括平滑、一阶导、归一化、PLS 特征提取、OLDA 特征提取、SVM 分类器等环节，多个环节组合实现对单倍体与二倍体两类籽粒的分类。在分类时，随机抽取 T1 数据集的一半作为训练集，另一半作为测试集，重复 20 次，将识别率取平均得到平均识别率，分类完成后保存所得平均识别率与对应的参数组合，并对识别率进行排序。

3　结果与讨论

利用图 2 所示的 PSO 算法流程以及图 5 所示的多重循环嵌套方法分别对上述近红外定性分析模型最优参数组合进行搜寻，近红外光谱定性分析模型中待寻优的参数包括平滑参数、一阶导系数、PLS 维数、OLDA 维数共 4 个。

实验数据：使用玉米单倍体二倍体光谱数据作为实验数据集。

针对上述数据建立近红外光谱定性分析模型，定性分析模型中数据预处理采用平滑(Smoothing)、一阶导(first Derivative，FD)、矢量归一化(vector normalization，VN)三种方法相结合[8-9]，特征提取环节采用 PLS＋OLDA 组合的方式，最后利用 SVM 方法进行分类鉴别。

将 T1 作为预处理与特征提取参数组合优化算法的实验数据集，T2—T5 数据集作为测试集验证所获得优化参数组合的推广性能，使用上述两种方法所得实验结果及其分析见表 1 和表 2。

表 1　识别率及其对应参数列表(PSO 方法)

Table 1　Recognition rate and its corresponding parameter list (PSO method)

序号	SP	FDP	PLS	OLDA	平均识别率/%
1	11	7	11	5	96.2
2	13	9	9	4	96.5
3	9	9	11	4	96.9
4	11	13	7	4	97.0
5	11	11	9	5	96.3

表 2　识别率及其对应参数列表(多重循环嵌套方法)

Table 2　Recognition rate and its corresponding parameter list (multiple loop nesting method)

序号	SP	FDP	PLS	OLDA	平均识别率/%
1	13	7	11	4	96.9
2	13	9	9	4	96.6
3	9	9	11	5	97.2
4	11	11	9	4	96.8
5	11	11	9	5	96.4

表 1 及表 2 是分别使用粒子群算法与多重循环嵌套方法寻优得到模型识别率最高时的几组参数组合，由两表可以看

出，在两种方法中识别率较高时对应参数有多种组合。

选择表 2 中任意一组数据可以发现：PSO 算法与多重循环嵌套方法获得的参数值比较接近，PSO 算法与多重循环嵌套方法所获的第二组参数组合则完全一致。

此外，在同一种方法所获的几种参数组合中，参数之间相差并不大，基本在小范围内波动。以 PSO 算法为例，平滑参数在 9～11 范围内波动，而 OLDA 维数在 4～5 之间波动。因此，在模型实际使用时，为获得较优或者最优的识别性能，可以将参数设置在一定范围内。

对于本实验数据与对应分类任务，最优识别效果对应平滑参数一般可设置为 9，11 和 13，而一阶导参数可设置为 7，9，11 和 13，PLS 参数可设置为 7，9 和 11，OLDA 可设置为 4 和 5。

在构建近红外光谱定性分析模型时，选择表格中任一组参数组合代入模型，均可获得最优或者较优的预测性能。

表 3 最优参数组合在其他数据集的识别结果表

Table 3 Recognition results of optimal parameter combination tested with other data sets

测试集	T1	T2	T3	T4	T5
正确识别率/%	96.2	95.6	95.2	96.2	94.6
正确拒识率/%	97.4	95.2	94.8	93.4	94.0
平均识别率/%	96.9	95.4	95.0	94.9	94.3

表 3 是利用 PSO 方法在数据集 T1 上获得的最优预处理与特征提取参数组合代入近红外定性鉴别模型后，在多个测试数据集上所得识别率。对每一个数据集进行实验时，从各数据集中随机抽取一半作为训练集建立定性分析模型，剩余一半作为测试集对所建模型进行测试，抽取样本时，两类样本各占 50%，20 次实验所得识别率取均值。

建立模型阶段代入 PSO 方法搜寻得到的预处理与特征提取参数组合 SP＝9，FDP＝9，PLS＝11，OLDA＝4。由表 3 可以看出，所获参数组合在几个数据集上均能获得高于 95%的识别率，说明粒子群方法所获取优化参数组合在不同数据集均具有较好推广性能。

表 4 程序消耗时间对比表

Table 4 Comparison of program consumption time

方法	多重循环嵌套方法	PSO 方法
时间	2 825 s	586 s

表 4 是在同一台计算机分别使用多重循环嵌套方法以及 PSO 方法对最优参数组合进行寻优所消耗的时间。由表 4 可知，PSO 方法只需 586 s，而循环嵌套方法需要 2 825 s，PSO 方法寻优效率较高，而多重循环嵌套方法寻优效率需要消耗大量的计算机内存与计算时间，效率较低。

4 结 论

针对近红外光谱定性分析模型中预处理及特征提取环节多参数需要寻优，当前常用试凑法及多重循环嵌套方法存在无法获得全局最优解、效率低的问题。研究中提出了一种基于粒子群算法的定性分析模型预处理与特征提取参数组合优化方法，首先采用粒子群算法与 SVM 算法对预处理与特征提取的多环节参数组合进行寻优，再利用多个测试集对代入最优参数组合的定性分析模型进行测试，实验结果证明了方法的有效性。

References

[1] Chu X L, Shi Y Y, Chen B, et al. Journal of Instrumental Analysis, 2019, 38(5): 603.

[2] YAN Yan-lu, CHEN Bin, ZHU Da-zhou(严衍禄，陈 斌，朱大洲). Near Infrared Spectroscopy Analytical—Principles, Technology and Application(近红外光谱分析的原理、技术与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社), 2007.

[3] Mendes T O, Junqueira G M A, Porto B L S, et al. Journal of Raman Spectroscopy, 2016, 47(6): 692.

[4] Konstantinos E Parsopoulos, Michael N Vrahatis. Particle Swarm Optimization and Intelligence. 1st ed. Hershey, New York: Information Science Reference, 2010.

[5] Qin H, Ma J Y, Chen S J, et al. Infrared Technology, 2015, 1(37): 78.

[6] LI Hao-guang, LI Wei-jun, QIN Hong, et al(李浩光，李卫军，覃 鸿，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2016, 47(6): 259.

[7] LI Hao-guang, LI Wei-jun, QIN Hong, et al(李浩光，李卫军，覃 鸿，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2017, (S1): 422.

[8] Kurtulmus F, Alibas I, Kavdir I. Int. J. Agric. Biol. Eng., 2016, 9(1): 51.

[9] Le Cun Y, Bengio Y, Hinton G. Nature: Deep Learning, 2015, 521(7553): 436.

Research of Parameter Optimization of Preprocessing and Feature Extraction for NIRS Qualitative Analysis Based on PSO Method

LI Hao-guang[1,2], YU Yun-hua[1,2], PANG Yan[1], SHEN Xue-feng[1,2]

1. College of Mechanical and Control Engineering, Shandong Institute of Petrochemical and Chemical Technology, Dongying 257061, China
2. New Energy College, China University of Petroleum(East China), Dongying 257061, China

Abstract In the qualitative analysis of near-infrared spectroscopy, preprocessing and feature extraction are indispensable to achieve good recognition results. Preprocessing is mainly to eliminate the influence of various interference factors on the spectral data. The common preprocessing methods include smoothing, first-order derivatives, normalization, etc., while the feature extraction methods can eliminate the irrelevant information in the data and retain the effective information. The common feature extraction methods include partial least squares, principal component analysis, linear discriminant analysis, etc. Different preprocessing and feature extraction methods have different characteristics. When building a qualitative analysis model, it is often difficult to achieve ideal results by using a single preprocessing or feature extraction method. It is often necessary to use a combination of multiple preprocessing and feature extraction methods to improve the model's performance. Variable parameters such as feature extraction dimension need to be set in each preprocessing and feature extraction process. These variable parameters have an important impact on the performance of the model. Therefore, multiple parameters need to be determined in multiple preprocessing and multiple feature extraction methods. In practice, the trial and error method is often used to find the optimal value of each parameter. In order to get the optimal value of one of the parameters, it is necessary to fix the other parameter values according to experience. Then a parameter to be optimized is substituted into the NIR qualitative analysis model for trial and error to get the corresponding parameter value of the optimal recognition rate of the model, and take it as the optimal value. After several parameters to be optimized are obtained one by one by trial and error method, the combination of parameters is set into the qualitative analysis model, and finally, qualitative identification is carried out. However, the combination of parameters obtained by the trial and error method is difficult to guarantee the optimal optimal solution. In addition to the trial and error method, the multiple loops nesting method can also be used to obtain the optimal combination of parameters in the preprocessing and feature extraction of the near qualitative analysis model. However, this method consumes a lot of computer memory and computing time and has the disadvantage of low efficiency. In this paper, a method based on particle swarm optimization (PSO) is proposed to optimize the parameters of pre-processing and feature extraction of the NIR qualitative analysis model, which can quickly obtain the optimal parameter combination of pre-processing and feature extraction and ensure that the qualitative analysis model with the optimal parameter combination has the best recognition performance. The experimental results show that the method is effective.

Keywords Particle swarm optimization; Feature extraction; Parameter optimization; Qualitative analysis

(Received Jan. 20, 2021; accepted Apr. 6, 2021)