蜻蜓算法优选小麦粉蛋白质近红外建模校正集

胡云超，刘智健，汪　莹，黄浩冉，王红鸿，吴彩娥，熊智新*
（南京林业大学轻工与食品学院，江苏 南京　210037）

摘　要：为优选小麦粉蛋白质近红外建模校正集，在传统K/S（Kennard/Stone）方法划分的初始校正集基础上采用二进制蜻蜓算法（binary dragonfly algorithm，BDA）挑选代表性样品，建立小麦粉蛋白质含量偏最小二乘回归（partial least square regression，PLSR）模型，并用预测集检验评估模型的稳定性及预测性能。结果表明：BDA挑选出的最佳校正集样品数量为30 个，所建模型的预测决定系数（R_p^2）为0.956 4，预测标准偏差（root mean square errors of prediction，RMSEP）为0.278 1，与传统K/S划分的100 个初始校正集的建模效果（R_p^2：0.938 8，RMSEP：0.329 4）相比，R_p^2提高了1.87%，RMSEP降低了15.57%。10 次BDA实验优选出校正集的平均数量为30.2 个，且所建10 个模型蛋白质含量预测效果均优于初始校正集建模。综上，BDA算法可以优选出数量少、具有代表性的校正集样品，建立的小麦粉蛋白质PLSR模型稳定性好、预测精度高，可为小麦粉品质近红外检测分析提供一种高效的校正集优选方法。

关键词：蜻蜓算法；近红外光谱；校正集优选；小麦粉蛋白质含量

Calibration Set Optimization by Dragonfly Algorithm for Near-Infrared Modeling of Wheat Flour Protein Content

HU Yunchao, LIU Zhijian, WANG Ying, HUANG Haoran, WANG Honghong, WU Cai'e, XIONG Zhixin*
(College of Light Industry and Food Engineering, Nanjing Forestry University, Nanjing　210037, China)

Abstract: In order to optimize the calibration set for near-infrared modeling of the protein content in wheat flour, the binary dragonfly algorithm (BDA) was used to select representative samples from the primary calibration set divided by the traditional Kennard/Stone (K/S) method. Based on the representative samples, a partial least square regression (PLSR) model for estimating the protein content in wheat flour was established, and the prediction set was employed to evaluate the stability and prediction performance of the model. The results indicated that an optimal calibration set with 30 samples was selected finally by BDA, and the proposed model exhibited a coefficient of determination of prediction (R_p^2) of 0.956 4 and a root mean square errors of prediction (RMSEP) of 0.278 1, which increased by 1.87% and decreased by 15.57% compared with those (0.938 8 and 0.329 4) from K/S partition of 100 primary calibration sets, respectively. The average number of calibration sets selected from 10 BDA experiments was 30.2, and the protein content of wheat flour was predicted better by the 10 models developed than that obtained based on the primary calibration set. Therefore, BDA can select a small number of representative calibration set samples based on which a PLSR model with good robustness and high prediction accuracy for the protein content of wheat flour can be established. The proposed method can provide an efficient tool for calibration set selection in near-infrared spectroscopic analysis of the quality of wheat flour.

Keywords: dragonfly algorithm; near-infrared spectroscopy; optimization of calibration set; protein content of wheat flour
DOI:10.7506/spkx1002-6630-20230317-170
中图分类号：TS207.3　　文献标志码：A　　文章编号：1002-6630（2024）09-0009-07
引文格式：
胡云超, 刘智健, 汪莹, 等. 蜻蜓算法优选小麦粉蛋白质近红外建模校正集[J]. 食品科学, 2024, 45(9): 9-15. DOI:10.7506/spkx1002-6630-20230317-170.　http://www.spkx.net.cn

收稿日期：2023-03-17
基金项目：“十三五”国家重点研发计划重点专项（2019YFD1002300）
第一作者简介：胡云超（2000—）（ORCID: 0009-0006-0008-0807），女，硕士研究生，研究方向为农林产品无损检测。
E-mail: nlhyc@njfu.edu.cn
*通信作者简介：熊智新（1973—）（ORCID: 0000-0001-9720-922X），男，副教授，博士，研究方向为农林产品无损检测。
E-mail: Leo_xzx@njfu.edu.cn

HU Yunchao, LIU Zhijian, WANG Ying, et al. Calibration set optimization by dragonfly algorithm for near-infrared modeling of wheat flour protein content[J]. Food Science, 2024, 45(9): 9-15. (in Chinese with English abstract)
DOI:10.7506/spkx1002-6630-20230317-170. http://www.spkx.net.cn

小麦是世界上种植面积最广、总产量和营养价值最高的粮食作物，提供了人类20%的能量[1]。小麦行业的发展对国家的粮食安全和社会稳定具有重要意义，2022年，国内粮食市场“麦强面弱”格局明显，产品品质、品牌成为企业赢得小麦粉市场的关键[2]。小麦粉中有三大营养素，分别是蛋白质、淀粉和脂类，其中蛋白质（含量约为7%～15%）决定着小麦粉的加工品质和营养品质[3]。小麦粉可根据其蛋白质含量分为高筋粉（大于10.5%）、中筋粉（8.0%～10.5%）和低筋粉（小于8.0%）[4]。小麦粉中蛋白质含量的不同使得小麦粉具有不同的用途，例如高筋粉一般用于制作面包，而点心和菜肴一般使用低筋粉进行制作加工，所以在生产过程中对小麦粉蛋白质含量的快速精确检测就显得尤为重要。

近红外光谱分析技术是21世纪发展起来的一种快速、无损、绿色、可用于在线监测的分析技术，广泛应用于食品[5]、农业[6]、医药[7]、林业[8]等各个领域，随着科学技术的发展，结合化学计量学的近红外光谱分析技术在小麦粉蛋白质定量分析中的应用逐渐广泛[9-11]。近红外光谱所分析的对象大多是复杂的、未预处理的样品体系，通常会收集大量的实验样本，但这些样本可能80%以上是重复样本或无效样本，因此有必要从中挑选出具有一定代表性的校正样本代替原始数据集进行建模，提高建模的效率和模型精度，减少数据库的存储空间。常用的样本划分方法有随机采样法、K/S（Kennard/Stone）法、SPXY（sample set partitioning based on joint X-Y distances）法等。随机采样法是从样品集中随机选择一定数量的样品组成校正集[12]。K/S法是以光谱变量间的欧氏距离为基础，挑选分布范围广且代表性强的样品作为校正集[13-14]。SPXY法是在K/S法的基础上引入样品化学值信息，用光谱间距离以及化学值浓度之间的距离选择代表性样品[15-16]。由于K/S法和SPXY法以样本间的距离为标准对样品集进行划分，可能会将异常或者不合适的样本挑选入校正集，进而影响所建模型性能。群智能优化算法是化学计量学方法的重要组成部分，其主要思路是基于对自然生物群体（例如狼群、蚁群、蜻蜓等）生存现象的观察，将其生存现象量化并应用在数学模型优化中，特点为群个体之间相对独立，通过更新策略在搜索空间中寻找最优解。群智能优化算法在光谱分析领域中已有许多成功的研究及应用案例，主要应用在特征波长优选及建模方法参数优化等方面。Guo Zhiming等[17]利用近红外光谱分析技术结合模拟退火、蚁群优化、遗传算法等群智能优化算法，选择信息丰富的光谱变量，建立了准确、稳健的绿茶活性成分和抗氧化能力定量分析模型。王仲雨等[18]提出改进鲸鱼优化算法并用于近红外建模过程中的波长选择，该算法能有效筛选出波长变量并建立玉米脂肪、蛋白质、淀粉和水的预测模型。蜻蜓算法（dragonfly algorithm，DA）作为群智能优化算法的一种，将群体行为的所有可能因素都考虑在内，使其能够将目标函数快速地收敛在最优解附近，具有良好的全局寻优能力[19-20]。陈勇等[21]采用衰减消退蜻蜓算法优选小麦粉蛋白质近红外特征波长，筛选出的波长数量少，所建模型稳定性高。Chen Yuanyuan等[22]提出了一种新的基于二进制蜻蜓算法（binary dragonfly algorithm，BDA）的波长选择方法，针对汽油近红外光谱数据集选择有效波长，结果表明基于多BDA和集成学习BDA算法可以提高波长选择的稳定性。蜻蜓算法在近红外特征波长优选、建模方法参数优化等方面有着良好的应用性能，但在模型建立过程中优选校正集的应用鲜见报道。本研究采用BDA算法挑选具有代表性的校正集样品，以迭代过程中BDA选出的校正集建模的交互验证标准偏差（root mean square error of cross validation，RMSECV）与所建模型对验证集预测的预测标准偏差（root mean square errors of prediction，RMSEP）之和构建适应度函数，从而在适应度函数构建中引入校正集信息，实现对校正集样品的优选，提高模型预测的精度，并以NeoSpectra Micro型便携式近红外光谱仪所测的小麦粉近红外光谱和蛋白质数据为例，与传统的校正集优选算法（K/S法、SPXY法）的预测结果进行对比和分析，探讨BDA算法优选小麦粉蛋白质近红外建模校正集样品的可行性。

1 材料与方法

1.1 材料

实验所用样品为超市购买不同品牌、不同批次的小麦粉，共计160 个样品，包含低筋粉23 份、中筋粉82 份和高筋粉55 份，收集到的样本置于保鲜袋内常温储存备用，取出小麦粉后于室温（20～23 ℃）条件下采集光谱。

1.2 仪器与设备

NeoSpectra Micro型便携式近红外光谱仪 埃及Si-ware公司；D200杜马斯定氮仪 济南海能仪器股份有限公司。

1.3 方法

1.3.1 光谱采集

NeoSpectra Micro型便携式近红外光谱仪的波长范围

为1 350～2 550 nm，波数范围为7 407～3 922 cm^{-1}，采样间隔为13.62 cm^{-1}，分辨率为16 cm^{-1}。采集小麦粉样品的近红外光谱时，NeoSpectra Micro型便携式近红外光谱仪机身采用金属试管架夹持固定，探头向下垂直对准深1 cm圆盘样品池，样品池顶部与探头底部相距1 cm，面粉样品铺平深1 cm圆盘样品池，按120°间隔采集得到3 条不同检测点的光谱，取它们的平均作为该样品的最终采集光谱，共得到160 个小麦粉的光谱数据。

1.3.2 蛋白质含量测定

小麦粉样品的蛋白质含量参照GB 5009.5—2016《食品中蛋白质的测定》[23]中的燃烧法测定。

1.3.3 建模与模型评估

采用偏最小二乘回归（partial least square regression，PLSR）法建立小麦粉蛋白质定量校正模型[24]，采用留一法交互验证，限定最大主成分数为12，选取最佳主成分数，即交叉验证的预测残差平方和（prediction residual error sum of square，PRESS）最小时对应的主成分数。

模型建立过程中采用RMSECV对模型的性能进行评价，建立最优的校正模型。模型建立完成后，通常采用RMSEP、决定系数（R^2）[25]等指标对模型的预测性能进行综合评价，R^2越接近1，表示模型的预测效果越好；如果R^2为负值，表明模型拟合效果极差。RMSECV和RMSEP值越小，所建模型的稳定性与预测精确度越高。

1.3.4 蜻蜓算法优选校正集

蜻蜓算法是Mirjalili[26]在2016年通过对自然界蜻蜓行为进行观察、总结和抽象后，提出的一种新的智能群体优化算法，并通过对几类典型函数优化验证了连续DA算法、BDA算法的有效性。生物学家观察到，蜻蜓主要通过5 种主要策略来改变其位置：分离（Separation）、对齐（Alignment）、聚集（Cohesion）、觅食（Attraction to food）、避敌（Distraction from enemy），这5 种策略的数学模型表达式分别如式（1）～（5）所示：

$$S_i=-\sum_{j=1}^{N}X-X_j \tag{1}$$

$$A_i=\frac{\sum_{j=1}^{N}V_j}{N} \tag{2}$$

$$C_i=\frac{\sum_{j=1}^{N}X_j}{N}-X \tag{3}$$

$$F_i=X^+-X \tag{4}$$

$$E_i=X^-+X \tag{5}$$

式中：i表示第i个蜻蜓；X表示当前蜻蜓的位置，X_j表示第j个邻近蜻蜓的位置；N表示邻近蜻蜓的数量；V_j表示第j个邻近蜻蜓的速率；X^+表示食物的位置；X^-表示危险或敌人的位置。

通过上述5 种策略位置，在搜索范围空间更新蜻蜓的位置并模拟它们运动，考虑了步长向量（ΔX）和位置向量（X），并在粒子群算法的框架基础上开发了一种基于步长向量（ΔX）和位置向量（X）的人工蜻蜓搜索算法。步长向量表明了蜻蜓的运动方向，如式（6）所示：

$$\Delta X_{t+1}=（sS_i+aA_i+cC_i+fF_i+eE_i）+w\Delta X_t \tag{6}$$

式中：s为分离权重；a为对齐权重；c为聚集权重；f为觅食权重；e为避敌权重；w为惯性权重；t为当前迭代次数。得出步长向量后，蜻蜓的位置更新如式（7）所示：

$$X_{t+1}=X_t+\Delta X_{t+1} \tag{7}$$

群智能优化算法在连续空间和离散空间中的优化方式不同。在连续搜索空间中，DA的搜索代理通过在位置向量上添加步进向量更新种群的位置，而在利用蜻蜓算法优选近红外建模校正集时，需将连续域转换为离散域，在离散域空间中寻找最优解。Mirjalili等[27]利用传递函数将蜻蜓算法进行改进，提出BDA，传递函数接收步长值作为输入并输出一个0或1的数字，表示位置变化的概率。V型传递函数如式（8）所示：

$$T（\Delta x）=\left|\frac{\Delta x}{\sqrt{\Delta x^2+1}}\right| \tag{8}$$

式中：Δx为传递函数的输入，即步长值。

用传递函数得出位置变化率后更新蜻蜓在搜索空间中的搜索位置（式（9））：

$$X_t+1=\begin{cases}-X_t & r<T（\Delta x_{t+1}）\\ X_t & r>T（\Delta x_{t+1}）\end{cases} \tag{9}$$

式中：r为[0, 1]之间的随机数；负号表示逻辑取反运算。

采用BDA算法优选校正集，首先使用K/S法将样本初步划分为初始校正集和预测集，初始校正集用于建立定量校正模型以及待优化，预测集在建模结束后用于评估优选的校正集建模的预测效果，接下来采用BDA算法，在初始校正集中进一步挑选出数量更少、更具有代表性的样品组成新的校正集，实现对校正集样品的优选。采用K/S法将初始校正集划分为子校正集和验证集，BDA的作用是在子校正集中挑选一定数量的样品作为新的校正集，根据其全局搜索能力强的特性在子校正集样本空间中大范围搜索合适的校正集，适应度函数值为优选出的校正集建立PLSR模型的RMSECV与该模型预测验证集的RMSEP之和（sum），如式（10）所示。每次实验迭代计算时，如果本次迭代最优解优于上次，则记录该最优

解对应的sum、RMSECV和RMSEP。经过不断的迭代更新，最终选取sum最小的样品集作为最优校正集。BDA算法优选校正集的流程如图1所示。

$$sum=0.6\times RMSECV+0.4\times RMSEP \quad (10)$$

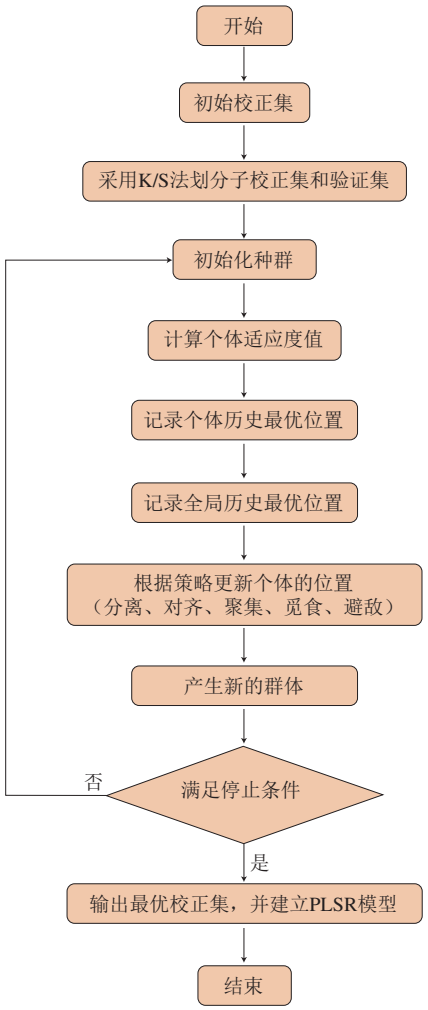


图 1　BDA算法优选校正集的流程图

Fig. 1　Flow chart of calibration set optimization by BDA

1.4　数据处理与分析

采用实验室自主研发的NIRSA 5.9.4系统[28]（计算机软件著作权登记号为2007SR06801）、Matlab 2016a等软件平台进行数据处理与分析。

2　结果与分析

2.1　样品划分

本研究所选样品的小麦粉蛋白质含量测定结果如表1所示，其含量基本覆盖小麦粉蛋白质量分数（7%～15%），并且分布较为均匀，表明该样品具有代表性。

表 1　小麦粉蛋白质含量统计

Table 1　Statistics of the protein content in wheat flour

样品数量	蛋白质量分数/%				
	最小值	最大值	极差	平均值	标准差
160	6.34	14.83	8.49	10.20	1.71

在采集的所有样品数据中，受样品、采集环境和仪器的影响，一定程度上会存在异常样品数据，影响所建模型的稳定性与预测能力，因此在建模之前必须将异常样品从集合中剔除。采用主成分分析（principal component analysis，PCA）与马氏距离相结合的方法检测异常样本，剔除马氏距离大于3*f*/*m*的样本，其中*f*为PCA所用主因子数，*m*为样本数，共剔除20 个异常样本。采用K/S方法将140 个正常样品划分为初始校正集（100 个）和预测集（40 个），其小麦粉蛋白质含量分布如表2所示，初始校正集与预测集的样本化学值分布较宽，具有良好的代表性。

表 2　初始校正集与预测集小麦粉蛋白质含量统计

Table 2　Statistics of the protein content in wheat flour in initial calibration and prediction sets

项目	样品数量	蛋白质质量分数/%				
		最小值	最大值	极差	平均值	标准差
初始校正集	100	6.34	14.10	7.76	10.19	1.73
预测集	40	6.99	13.78	6.79	10.23	1.33

2.2　初始校正集建模

以100 个初始校正集样品的近红外光谱及其蛋白质含量数据为研究对象，建立PLSR模型。为了消除光谱数据中无关信息和噪声的干扰，使用移动平均平滑（moving average filter，MAF）、Savitaky-Golay卷积平滑（Savitaky-Golay filter，SGF）、标准正态量变换（standard normal variate transformation，SNV）、一阶导数（1st derivative，1st D）、标准化及组合的预处理方法对样品进行预处理[29]，建立PLSR校正模型以评价预处理方法的优劣，选定最佳的预处理方法。不同预处理方法的校正模型评价结果如表3所示。

表 3　不同预处理方法的样品蛋白质PLSR校正模型评价

Table 3　Evaluation of PLSR calibration models developed using different pretreatment methods

预处理方法	窗口宽度	初始校正集		预测集	
		R^2_{cv}	RMSECV	R^2_p	RMSEP
无预处理	—	0.960 6	0.343 1	0.938 8	0.329 3
MAF+标准化	5	0.962 3	0.335 7	0.938 8	0.329 4
SGF+标准化	5	0.960 6	0.343 4	0.938 8	0.329 3
SNV+标准化	—	0.957 3	0.357 3	0.936 7	0.334 9
MAF+SNV+标准化	5	0.959 9	0.346 1	0.931 8	0.347 7
SGF+SNV+标准化	5	0.957 3	0.357 2	0.936 7	0.334 9
1st D+标准化	5	0.952 5	0.376 7	0.930 5	0.351 1

注：—.方法无需设置该参数。

由表3可知，对比不同预处理方法的建模效果，其中MAF＋标准化（MAF窗口宽度为5）的预处理方法除RMSEP略高于无预处理和SGF＋标准化外，各项指标均为最优，此时PLSR模型的R^2_{cv}为0.962 3，RMSECV为0.335 7，R^2_p为0.938 8，RMSEP为0.329 4，模型具有较高的预测精度，后续实验均采用MAF＋标准化（MAF窗口宽度为5）的预处理方法。

2.3　蜻蜓算法优选校正集

采用K/S方法将初始校正集划分为子校正集和验证集，比例为4∶1，子校正集80 个，验证集20 个，结合BDA算法优选校正集，设置迭代次数40 次，初始种群数500，优选校正集样品数量20～40 个。进行10 次BDA优选校正集实验，实验序号记为BK1～BK10，sum变化如图2所示，随着迭代的进行，sum越来越小，表明所挑选的校正集建模以及所建模型对验证集的预测评价参数越来越优。优选校正集的建模及预测结果如表4所示，10 次实验优选的校正集样品个数平均为30.2 个，平均R^2_p为0.949 5，RMSEP为0.299 0，平均预测性能R^2_p提高了1.14%，RMSEP降低了9.23%，10 次优选的校正集建模预测性能均优于初始校正集，实验BK1在10 次实验中优选出的30 个校正集样本建模预测效果最优（R^2_p：0.956 4，RMSEP：0.278 1），与初始校正集相比，R^2_p提高1.87%，RMSEP降低15.57%，实验BK3和BK10所优选出的校正集样品数仅24 个，且具有较好的模型稳定性和预测能力。

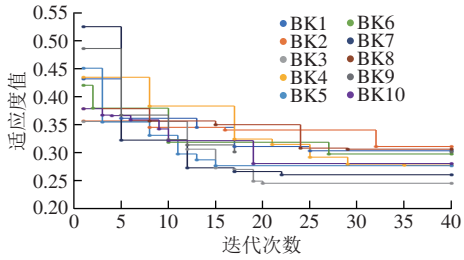


图 2　10 次BDA优选校正集实验适应度值变化

Fig. 2　Changes in fitness values for BDA experiments 1–10 for calibration set optimization with the number of iterations

表 4　10 次BDA优选校正集实验的建模及预测结果

Table 4　Modeling and prediction results from BDA experiments 1–10 for calibration set optimization

实验	优选校正集样品个数	适应度函数值（sum）	优选校正集		预测集	
			R^2_{cv}	RMSECV	R^2_p	RMSEP
BK1	30	0.304 3	0.972 6	0.287 3	0.956 4	0.278 1
BK2	31	0.312 2	0.966 2	0.3033	0.952 2	0.291 2
BK3	24	0.246 9	0.980 1	0.226 3	0.945 2	0.311 6
BK4	34	0.278 8	0.986 7	0.214 9	0.946 6	0.307 5
BK5	34	0.278 0	0.982 1	0.244 1	0.951 9	0.292 1
BK6	26	0.307 0	0.970 5	0.272 8	0.944 2	0.314 6
BK7	36	0.261 8	0.984 9	0.214 1	0.945 6	0.310 5
BK8	31	0.299 0	0.949 5	0.304 4	0.947 8	0.304 1
BK9	32	0.302 5	0.959 2	0.314 1	0.952 6	0.289 8
BK10	24	0.282 2	0.976 1	0.241 8	0.952 5	0.290 1
平均值	30.20	0.287 3	0.972 8	0.263 1	0.949 5	0.299 0

图3为初始校正集、BK1优选校正集和预测集的蛋白质含量分布图，BK1所挑选出的校正集样本含量分布较为均匀，基本涵盖了预测集样品的含量分布范围。将BK1优选的校正集和预测集取前两个主成分作主成分分布图，如图4所示，30 个校正集在42 个预测集样本中均匀分布，尽可能地用较少的样本包含整个数据集的特征，从而使所建立的预测模型可以对预测集进行良好预测。

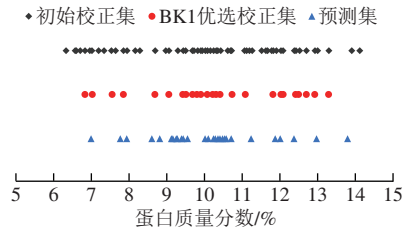


图 3　校正集和预测集样本的蛋白质含量分布

Fig. 3　Protein content distribution of calibration set and prediction set samples

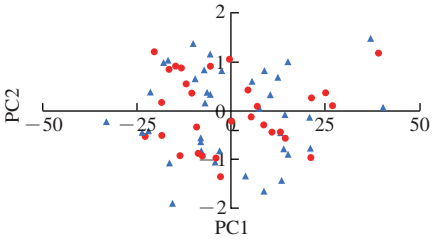


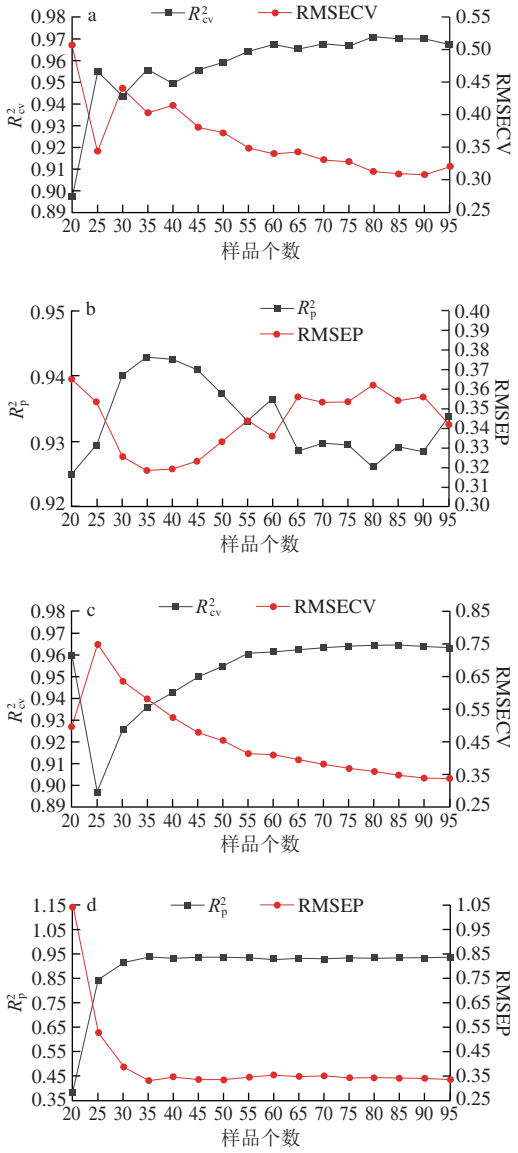
图 4　BK1优选校正集和预测集主成分分布

Fig. 4　Principal component analysis showing the distribution of calibration set and prediction set samples in BK1 for calibration set optimization

3　讨　论

在校正模型建立的过程中，选取参与校正的样本对建立稳健的模型是十分必要的，目前最常用的方法是K/S法和SPXY法。潘国锋[30]使用K/S算法对41 个水体中总氮光谱数据进行优选，用30 个样本建立了较为理想的硝酸盐定量校正模型。王世芳等[31]以小型西瓜为研究对象，校正集与预测集通过SPXY法进行划分，建立了西瓜瓜梗、瓜脐和赤道3 个部位的可溶性固形物含量模型，预测精度较好。朱荣光等[32]采用浓度梯度法、随机法、K/S以及SPXY法共4 种校正集划分方法对牛肉嫩度高光谱数据进行划分和比较，结果发现在偏最小二乘回归和主成分回归建模时，SPXY法所挑选出的校正集建模效果均较优。本研究将与传统的K/S法和SPXY法优选校正集进行对比，利用传统方法从初始校正集中分别采用K/S和SPXY法进一步挑选出k（k＝20, 25,…, 90, 95）个样品作为新的校正集建立PLSR模型，并对预测集进行预测，结果如图5所示。由图5a、b可知，K/S法所挑出的校

正集随着样品数量的增加模型稳定性整体上越来越好，当所选样品个数为80、85以及90时所建模型稳定性最优，当所选样品个数为35时，模型预测效果最好（R_p^2：0.942 8，RMSEP：0.318 4）。由图5c、d可知，SPXY法所挑选出的校正集随着样品数量的增加模型稳定性整体变好；当样品个数为20时，所建模型稳定性较优，但预测性能差（R_p^2：0.385 6，RMSEP：1.043 6）；当样品个数为85时，所建模型稳定性最优，预测性能较好（R_p^2：0.933 4，RMSEP：0.343 5）；当样品个数为35时，所建模型稳定性较优，且预测性能最好（R_p^2：0.938 1，RMSEP：0.331 3）。



a. K/S法优选的校正集建模参数；b. K/S法优选的校正集预测参数；c. SPXY法优选的校正集建模参数；d. SPXY法优选的校正集预测参数。

图 5　K/S、SPXY法优选校正集建模及预测参数

Fig. 5　Modeling and prediction parameters of K/S and SPXY optimal calibration sets

通过K/S和SPXY法挑选出的校正集建模和预测结果可以看出，K/S法从初始校正集100 个样品中挑选出35 个样品作为新校正集，所建模型的预测精度相较于初始校正集而言也略有提升，R_p^2从0.938 8上升到0.942 8，初步达到了优选校正集的效果；SPXY法在挑选出35 个样品建模时预测性能最好，但预测精度略低于初始校正集建模，R_p^2为0.938 1，不符合挑选出数量更少的校正集建立预测精度更高的模型的目标。而采用BDA算法从初始校正集中优选校正集，10 次实验所选出的新校正集建模预测精度均高于初始校正集，挑选出30 个样品进行建模时，预测R_p^2高达0.956 4，样品个数为24时，预测R_p^2也可以达到0.952 5，说明采用BDA算法可以优选出数量更少的校正集建立预测精度更高的小麦粉蛋白质定量模型。

4　结　论

本研究在传统挑选校正集样品的基础上引入BDA算法进行优化，以所选校正集建立的模型RMSECV与其对验证集的RMSEP之和构建适应度函数，并与传统校正集挑选方法K/S和SPXY法进行比较。结果表明，BDA算法优选出的校正集有最优的预测性能，在10 次BDA优选实验中，平均挑选出的校正集个数约占原校正集个数的30%（从100 个降低到30.2 个），平均预测性能R_p^2提高了1.14%（从0.938 8提升至0.949 5），RMSEP降低了9.23%（从0.329 4降低至0.299 0）。采用BDA算法可以优选出数量少、具有代表性的校正集样品，建立的小麦粉蛋白质PLSR模型稳定性好、预测精度高，可为小麦粉品质近红外检测分析提供一种高效、准确的校正集优选方法。

参考文献：

[1]　陶金亚. 新型改良小麦可能有助于解决全球粮食短缺[J]. 中国食品学报, 2020, 20(12): 330.

[2]　张春良. 2022年国内面粉市场回顾及2023年展望[J]. 现代面粉工业, 2023, 37(1): 47-52; 57.

[3]　付苗苗. 面粉中三大营养组分对馒头品质影响的研究进展综述[J]. 粮食加工, 2014, 39(5): 20-23.

[4]　田炎炎. 基于小麦粉中菌落增长特征的预警指标研究[D]. 天津: 天津科技大学, 2020: 1. DOI:10.27359/d.cnki.gtqgu.2020.000271.

[5]　DOUGLAS R K, NAWAR S, ALAMAR M C, et al. Rapid detection of alkanes and polycyclic aromatic hydrocarbons in oil-contaminated soil with visible near-infrared spectroscopy[J]. European Journal of Soil Science, 2019, 70(1): 140-150. DOI:10.1111/ejss.12567.

[6]　ASSI S, KHAN I, EDWARDS A, et al. On-spot quantification of modafinil in generic medicines purchased from the Internet using handheld Fourier transform-infrared, near-infrared and Raman spectroscopy[J]. Journal of Analytical Science and Technology, 2020, 11(1): 2231-2236. DOI:10.1186/s40543-020-00229-3.

[7]　BERNHARD T, TRUBERG B, FRIEDT W, et al. Development of near-infrared reflection spectroscopy calibrations for crude protein and

dry matter content in fresh and dried potato tuber samples[J]. Potato Research, 2016, 59(2): 149-165. DOI:10.1007/s11540-016-9318-8.

[8] DOS SANTOS L M, AMARAL E A, NIERI E M, et al. Estimating wood moisture by near infrared spectroscopy: testing acquisition methods and wood surfaces qualities[J]. Wood Material Science and Engineering, 2020, 16(5): 1-8. DOI:10.1080/17480272.2020.1768143.

[9] JIN H, WANG J, YAN L, et al. Establishment of nondestructive testing model of the protein content in wheat flour by near infrared spectroscopy[C]//International Conference on New Technology of Agricultural Engineering, 2011: 1125-1129.

[10] CHEN J, ZHU S P, ZHAO G H. Rapid determination of total protein and wet gluten in commercial wheat flour using siSVR-NIR[J]. Food Chemistry, 2017, 221: 1939-1946. DOI:10.1016/j.foodchem.2016.11.155.

[11] 陈嘉, 叶发银, 赵国华. 基于信息融合的小麦粉品质快速检测[J]. 食品与发酵工业, 2019, 45(15): 243-250. DOI:10.13995/j.cnki.11-1802/ts.020329.

[12] CHEN Y, ZHONG Y C, QI Y, et al. Near-infrared spectroscopy for rapid evaluation of different processing products of *Sophora japonica* L.[J]. Spectroscopy Letters, 2018, 51(1): 37-44. DOI:10.1080/00387010.2017.1416478.

[13] KENNARD R W, STONE L A. Computer aided design of experiments[J]. Technometrics, 2012, 11(1): 137-148. DOI:10.1080/00401706.1969.10490666.

[14] 武晴滢, 祝震予, 吴剑鸣, 等. 泛Kennard-Stone算法的数据集代表性度量与分块采样策略[J]. 高等学校化学学报, 2022, 43(10): 150-157. DOI:10.7503/cjcu20220397.

[15] GALVÃO R K, ARAUJO M C, JOSÉ G E, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736-740. DOI:10.1016/j.talanta.2005.03.025.

[16] 李艳芬, 马瑞峻, 陈瑜, 等. 利用光谱技术结合化学计量学分析方法快速检测生物农药阿维菌素的试验研究[J]. 农业环境科学学报, 2023, 42(9): 2140-2146. DOI:10.11654/jaes.2022-1269.

[17] GUO Z M, BARIMAH A O, SHUJAT A, et al. Simultaneous quantification of active constituents and antioxidant capability of green tea using NIR spectroscopy coupled with swarm intelligence algorithm[J]. LWT-Food Science and Technology, 2020, 129: 109510. DOI:10.1016/j.lwt.2020.109510.

[18] 王仲雨, 高美凤. 基于改进鲸鱼优化算法的近红外光谱波长变量选择方法及其应用[J]. 分析测试学报, 2023, 42(1): 37-44. DOI:10.19969/j.fxcsxb.22081803.

[19] MERAIHI Y, RAMDANE-CHERIF A, ACHELI D, et al. Dragonfly algorithm: a comprehensive review and applications[J]. Neural Computing and Applications, 2020, 32(21): 1-22. DOI:10.1007/s00521-020-04866-y.

[20] SANKA S N, YARRAM T R, YENUMALA K C, et al. Dragonfly algorithm based spectrum assignment for cognitive radio networks[J/OL]. Materials Today: Proceedings, 2021. DOI:10.1016/J.MATPR.2020.11.301.

[21] 陈勇, 吴彩娥, 熊智新. 基于衰减消去蜻蜓算法的小麦粉蛋白质近红外特征波长优选[J]. 食品科学, 2022, 43(14): 219-225. DOI:10.7506/spkx1002-6630-20210608-102.

[22] CHEN Y Y, WANG Z B, HUCK C. Wavelength selection for NIR spectroscopy based on the binary dragonfly algorithm[J]. Molecules, 2019, 24(3): 421. DOI:10.3390/molecules24030421.

[23] 何杏宗, 冯雪雅, 赵悦, 等. 杜马斯燃烧法测定饼干中蛋白质的方法研究[J]. 食品工业, 2018, 39(12): 169-171. DOI:CNKI:SUN:SPGY.0.2018-12-043.

[24] 田静, 陈斌, 陆道礼, 等. 不同分光原理近红外光谱仪光谱标准化方法在小麦粉品质检测中的应用[J]. 中国食品学报, 2022, 22(10): 286-294. DOI:10.16429/j.1009-7848.2022.10.031.

[25] LIU Q, ZHANG W, ZHANG B, et al. Determination of total protein and wet gluten in wheat flour by Fourier transform infrared photoacoustic spectroscopy with multivariate analysis[J]. Journal of Food Composition and Analysis, 2022, 106: 104349. DOI:10.1016/j.jfca.2021.104349.

[26] MIRJALILI S. Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems[J]. Neural Computing and Applications, 2016, 27(4): 1503-1573. DOI:10.1007/s00521-015-1920-1.

[27] MIRJALILI S, LEWIS A. S-shaped versus V-shaped transfer functions for binary particle swarm optimization[J]. Swarm and Evolutionary Computation, 2013, 9: 1-14. DOI:10.1016/j.swevo.2012.09.002.

[28] XIONG Z X, PFEIFER F, SIESLER H W. Evaluating the molecular interaction of organic liquid mixtures using near-infrared spectroscopy[J]. Applied Spectroscopy, 2016, 70(4): 635-644. DOI:10.1177/0003702816631301.

[29] 褚小立. 现代光谱分析中的化学计量学方法[M]. 北京: 化学工业出版社, 2022: 79-93.

[30] 潘国锋. 基于K-S算法的水质硝酸盐含量光谱检测方法研究[J]. 光谱实验室, 2011, 28(5): 2700-2704. DOI:10.3969/j.issn.1004-8138.2011.05.132.

[31] 王世芳, 韩平, 崔广禄, 等. SPXY算法的西瓜可溶性固形物近红外光谱检测[J]. 光谱学与光谱分析, 2019, 39(3): 738-742. DOI:10.3964/j.issn.1000-0593(2019)03-0738-05.

[32] 朱荣光, 段宏伟, 王龙, 等. 不同分集方法对牛肉嫩度高光谱检测模型的比较[J]. 食品与发酵工业, 2016, 42(4): 189-192. DOI:10.13995/j.cnki.11-1802/ts.201604034.