油脂安全

DOI: 10.12166/j.zgyz.1003-7969/2020.02.017

优化基于近红外光谱的联合间隔偏最小二乘法建模检测芝麻油掺伪含量

陈洪亮[1],曾　山[2],王　斌[1]

(1.南京财经大学 信息工程学院,南京 210046; 2.武汉轻工大学 数学与计算机学院,武汉 430040)

摘要:应用近红外光谱(NIR)分析技术建立测定芝麻油中大豆油含量的定量分析模型。基于32个含量梯度共384个掺伪芝麻油样品的近红外光谱,首先采用标准正态变量变换(SNV)对光谱进行预处理,再采用无信息变量消除法(UVE)初步筛选波长变量,然后结合联合间隔偏最小二乘法(SiPLS)和带极值扰动的简化粒子群优化算法(tsPSO)建立芝麻油中大豆油掺伪含量预测模型,经特征波段选取后建立的模型变量减少,波长变量由451个减少到219个,训练集和测试集相关系数分别为0.999 8和0.991 9,均方根误差分别为4.39E-2和3.99E-2。结果表明,该方法能够作为芝麻油中大豆油掺伪含量的快速检测方法。此外,该方法也可应用到芝麻油中掺入其他低价值油的掺伪含量检测中。

关键词:近红外光谱;无信息变量消除法;联合间隔偏最小二乘法;带极值扰动的简化粒子群优化算法

中图分类号:TS225.1;Q657　　**文献标识码**:A　　**文章编号**:1003-7969(2020)02-0086-05

Detection of adulteration content in sesame oil by optimizing the model established by synergy interval partial least squares based on near infrared spectroscopy

CHEN Hongliang[1], ZENG Shan[2], WANG Bin[1]

(1. College of Information Engineering, Nanjing University of Finance and Economic, Nanjing 210046, China; 2. School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430040, China)

Abstract: A quantitative analysis model for determining the adulteration content of soybean oil in sesame oil was established by near infrared spectroscopy technique. Based on the near infrared spectroscopy of a total of 384 adulterated sesame oil samples from 32 content gradients, the spectrum was pre-processed by the standard normal variate transformation (SNV), then the wavelength variable was initially filtrated by the uninformative variables elimination (UVE) method firstly, and then the model for predicting the adulteration content of soybean oil in sesame oil was established by combining the synergy interval partial least squares (SiPLS) with extremum disturbed simple particle swarm optimization (tsPSO). After wavelengths extraction, the number of wavelengths was reduced from 451 to 219. The correlation coefficients of the train set and the test set were 0.999 8 and 0.991 9 respectively. The root-mean-square errors were 4.39E-2 and 3.99E-2 respectively. The experimental results showed that the method could be used as a rapid method to detect the adulteration content of soybean oil in sesame oil. In addition, the method could also be applied to the detection of adulteration content of other low-value oil in sesame oil.

Key words: near infrared spectroscopy; UVE; SiPLS; tsPSO

收稿日期:2019-01-15;**修回日期**:2019-09-22

基金项目:国家重点研发计划(2017YFD0700501);江苏省科技计划(BY2016009-03)

作者简介:陈洪亮(1994),男,硕士,研究方向为粮油光谱检测技术(E-mail)15380428551@163.com。

通信作者:王　斌,教授(E-mail)wangbin@njue.edu.cn。

芝麻油是一种营养丰富的植物油，富含油酸和亚油酸，其特有的香味和出色的氧化稳定性使其成为备受青睐的调味品[1]。近年来，芝麻油的营养价值日益受到各行业的广泛关注[2]。由于芝麻油具有较高的营养价值，其售价远高于大豆油、菜籽油等常见食用油，因此一些不法生产者和经营者将低价格油掺入芝麻油中销售，这种做法严重损害了广大消费者和商家的利益。

目前，国内外学者对食用油掺伪检测方法的研究已取得了一定的进展，一些物理和化学方法被应用于食用油掺伪检测[3-4]。常规理化方法操作简便，不需要昂贵的仪器，但耗时长，测定过程往往要多名实验人员配合完成，无法满足快速检测油脂掺伪的要求[5]。色谱法与核磁共振波谱法分析快速，适合大批量油脂样本的掺伪检测，但所用仪器价格昂贵，且需要专业人员操作[6-7]。

近红外光谱（NIR）目前逐渐被应用到食用油的定性定量分析领域[8-9]，相较传统的食用油分析方法，近红外光谱分析技术具有灵敏度高、稳定、能实现快速在线分析等优点。近红外技术在食用油掺伪检测方面已有研究。涂斌等[10]以激光近红外光谱分析技术结合化学计量学方法对稻米油掺伪进行定性－定量分析，对比了偏最小二乘法（PLS）和支持向量机回归（SVR）两种方法，二者均有较高的预测精确度。冼瑞仪等[11]采用可见和近红外透射光谱分析技术结合区间偏最小二乘法（iPLS）、联合间隔偏最小二乘法（SiPLS）和反向区间偏最小二乘法（BiPLS）对掺杂不同含量煎炸老油的橄榄油建模分析，SiPLS 和 BiPLS 所建模型均取得了较好的预测效果，为合格植物油中掺杂其他不良油品的检测提供了参考。丁轻针等[12]采用标准正态变量变换（SNV）和偏最小二乘法（PLS）建立了芝麻油掺伪定量分析模型，当掺入量达到 10% 以上时，可以准确、可靠地实现快速检测。

虽然基于近红外光谱的食用油掺伪检测方法已有研究，但均未对预测食用油掺伪含量的最优特征波段进行探索，本研究应用近红外光谱分析技术结合无信息变量消除法、联合间隔偏最小二乘法和带极值扰动的简化粒子群优化算法优选特征波段建立芝麻油－大豆油掺伪含量分析模型，以期对波长变量做充分筛选后建立芝麻油掺伪含量预测模型取得相较单一 SiPLS 模型更好的预测效果。

1　材料与方法

1.1　试验材料

食用油掺伪定量鉴别试验样品：为配制具有代表性的掺伪样本，购买市售不同品牌、原料品种、加工工艺的芝麻油和大豆油，将大豆油以一定比例掺入芝麻油中，共配制 32 种掺伪含量。其中每种掺伪含量配制 12 份样品，共 384 个掺伪样品，每份样品约 10 g，充分振荡混合均匀后，在实验室静置 12 h 待测。具体芝麻油掺伪样本中大豆油的掺伪含量见表 1。

表 1　芝麻油掺伪样本中大豆油的掺伪含量

序号	掺伪含量/%	数量	序号	掺伪含量/%	数量
1	0	12	17	32	12
2	2	12	18	34	12
3	4	12	19	35	12
4	6	12	20	40	12
5	8	12	21	45	12
6	10	12	22	50	12
7	12	12	23	55	12
8	14	12	24	60	12
9	16	12	25	65	12
10	18	12	26	70	12
11	20	12	27	75	12
12	22	12	28	80	12
13	24	12	29	85	12
14	26	12	30	90	12
15	28	12	31	95	12
16	30	12	32	100	12

1.2　试验方法

1.2.1　试验流程

基于近红外光谱的无信息变量消除法－联合间隔偏最小二乘法－带极值扰动的简化粒子群优化算法（UVE－SiPLS－tsPSO）对芝麻油掺伪含量定量分析的具体实现流程如图 1 所示。由图 1 可知，芝麻油掺伪含量快速检测方法的步骤主要可概括划分为近红外光谱数据采集、光谱预处理、波长变量初步筛选和选择最优特征波段建立掺伪含量预测模型。

1.2.2　光谱采集

掺伪油样品光谱的采集采用激光近红外植物油品质快速检测仪，其主机为 Axsun XL410 型激光近红外光谱仪。Axsun XL410 型光谱仪以新型的超辐射发光二极管（SLED）作为光源，光谱测定范围 1 350 ~ 1 800 nm，扫描次数 32 次，分辨率 3.5 cm^{-1}，波长重复性 0.01 nm，信噪比（250 ms，RMS）大于 5 500∶1，温控范围 20 ~ 100 ℃。将光谱仪与工业电脑主板连接，方便待测样品光谱数据的采集和保存。试验中选用 2 mm 光程的比色皿，首先将油样滴在比色皿中，然后将比色皿放入光谱仪内便可开始采集光谱。

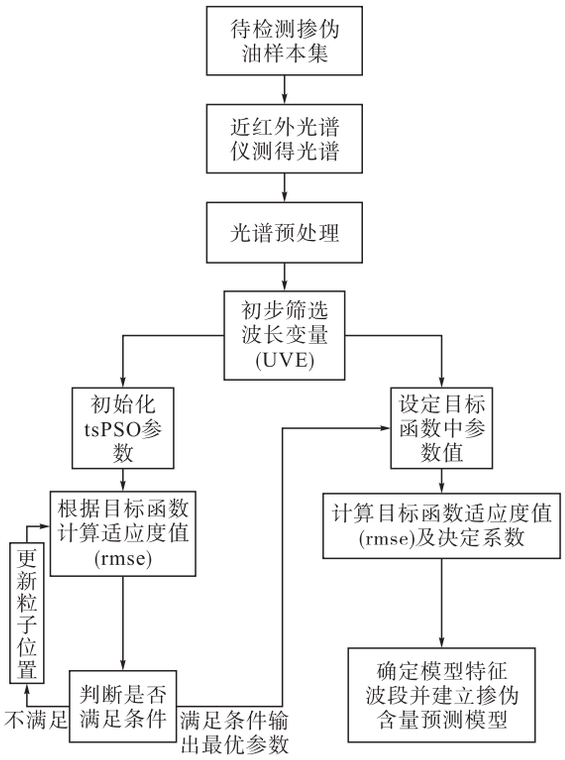

图 1 芝麻油掺伪含量快速检测方法流程图

1.2.3 光谱预处理及样本划分

近红外光谱技术属于二次分析技术,采集的光谱含有丰富的信息,但存在影响模型预测效果的因素,如谱带重叠严重、光谱信息专属性差、信噪比低等,建立模型前,为了去除光谱信号的高频随机噪声、比色皿对光程的影响及光线散射和杂散光影响,首先需要对光谱数据进行预处理,确保基于近红外光谱建立的定量检测模型具有较好的性能。采用标准正态变量变换(SNV)进行光谱预处理[13]。此外,本试验对样本光谱数据采用 SPXY 样本划分法[14],按 3∶1 的比例划分训练集和测试集,该方法能够覆盖多维向量空间,从而提升模型的预测能力。

1.2.4 建模及参数优化

首先采用无信息变量消除法(UVE)[15]排除与被测组分浓度无关的波长变量,利用 SPXY 样本划分法划分训练集和测试集后,采用联合间隔偏最小二乘法(SiPLS)结合带极值扰动的简化粒子群优化算法(tsPSO)优选最佳波长区间组合建立掺伪含量预测模型。

间隔偏最小二乘法(iPLS)是由 Norgaard 等[16]提出的,其原理是将整个光谱分成若干等宽子区间,对每个区间进行偏最小二乘回归,比较全光谱模型和每个子区间模型的性能,最终选择误差最小的子区间。联合间隔偏最小二乘法(SiPLS)[17]是间隔偏最小二乘法的拓展,它通过若干子区间的组合使误差最小。

粒子群优化算法(PSO)由 Kennedy 和 Eberhart 在 1995 年提出[18],该算法通过模拟鸟群、鱼群等生物捕食行为中相互合作机制寻找问题最优解。但是粒子群优化算法在进化后期收敛速度变慢,同时算法收敛精度不高,在多极值的复杂优化问题中易陷入局部最优解。本文采用带极值扰动的简化粒子群优化算法(tsPSO)[19],首先去掉了 PSO 进化方程的粒子速度项,避免由粒子速度项引起的后期收敛速度慢和精度低的问题,同时增加极值扰动算子用于使粒子跳出局部极值点继续优化。

2 结果与讨论

2.1 光谱预处理及特征波长选择

采用标准正态变量变换(SNV)对 32 种芝麻油 - 大豆油掺伪样品的近红外光谱进行预处理,能够有效去除光谱噪声。芝麻油 - 大豆油掺伪样本的近红外光谱如图 2 所示。

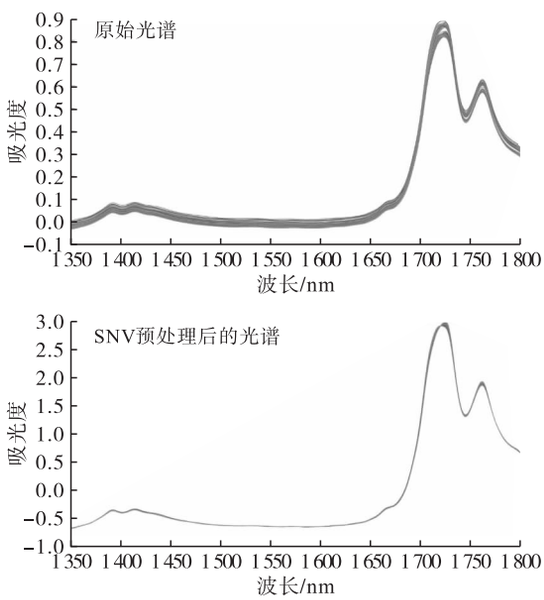

图 2 芝麻油 - 大豆油掺伪样品近红外光谱图

分析图 2,原始光谱和 SNV 预处理后光谱图中各有 384 条曲线,每条曲线代表一个样本在近红外波段各波长下的吸光度,对比可见原始光谱及经 SNV 预处理后的光谱,光谱的形状总体保持不变,只是排列更为紧凑。

对 SNV 预处理后的光谱数据采用 UVE 初步筛选特征波长,得到的光谱数据包含 436 个波长变量,经 UVE 筛选后的光谱图如图 3 所示。

图 3 中,横坐标表示经过 UVE 筛选后剩余波长变量由低到高排序,可见 UVE 在筛选掉经 SNV 预处理后的光谱中无信息波长变量的同时,没有破坏总体的光谱结构。

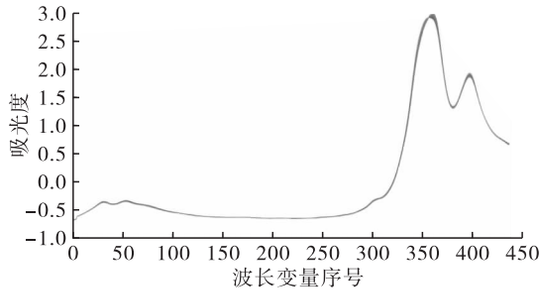

图 3　UVE 筛选后的光谱图

考虑到不同维度数据在分析中重要性不同,不同维度的信息具有一定的相关性,提供的信息有所重叠。经 UVE 筛选后的光谱数据仍具有较高的维度,因此需要对特征波长做进一步细选。由于光谱中重要信息往往分布较为集中,因此将光谱全波段划分为若干子区间进行分析能够取得更为理想的建模效果。试验中基于 1 350 ~ 1 800 nm 经 UVE 筛选后的光谱区域,采用 SiPLS 预测芝麻油中大豆油的掺伪含量,设子区间数为 M,所选取子区间数为 K,对于训练集光谱数据共有 C_M^K 种特征区间组合方式,逐一建立偏最小二乘回归模型并对测试集进行掺伪含量预测,最后选取训练集均方根误差最小的模型。

采用 SiPLS 方法建立模型时,波段的选择对模型的预测准确度存在一定的影响,波段分割过宽或选取波段数过多,会造成信息冗余;波段分割过窄或选取波段数过少,可能会丢失建模所需必要信息,因此选择合适的波段分割间隔和用于建模的波段数尤为重要。采用带极值扰动的简化粒子群优化算法(tsPSO)优化模型能很好地解决这一问题。

2.2　3 种模型的建立

为证明本研究方法的优越性,首先建立全波段掺伪含量预测模型和 UVE 模型作为对比试验。利用 SPXY 样本划分法对光谱经 SNV 预处理后的掺伪样本划分为训练集和测试集,采用 PLS 方法,将训练集光谱数据和掺伪含量数据作为输入量,建立全波段芝麻油掺伪含量预测模型。采用 UVE 对经 SNV 预处理后的光谱筛选波长变量,对降维后的光谱数据和掺伪含量数据采用 SPXY 样本划分法划分为训练集和测试集,对所得训练集样本利用 PLS 方法建立芝麻油掺伪含量预测模型,3 种模型预测结果见表 2。

表 2　3 种模型测试结果

建模方法	变量个数	参数		训练集		测试集	
		M	K	R^2	*RMSE*	R^2	*RMSE*
SNV – PLS	451	0	0	0.999 4	7.34E – 2	0.997 8	6.69E – 2
SNV – UVE – PLS	436	0	0	0.999 4	7.32E – 2	0.998 3	6.50E – 2
SNV – UVE – SiPLS – tsPSO	219	14	7	0.999 8	4.39E – 2	0.991 9	3.99E – 2

从表 2 可看出,所建全波段模型训练集和测试集相关系数(R^2)均接近 1,均方根误差(*RMSE*)分别为 7.34E – 2 和 6.69E – 2,芝麻油掺伪含量预测精度一般。而利用 UVE 筛选后的光谱数据建立的 PLS 模型,均方根误差(*RMSE*)分别为 7.32E – 2 和 6.50E – 2,相较全波段模型,略微降低了预测误差,此外参与建模的波长变量由 451 个降低到 436 个,缩短了建模时间。

为提升模型预测性能,对 SNV – UVE – PLS 模型进行优化,在采用 UVE 初步筛选波长变量后,采用 SiPLS 结合 tsPSO 进一步选取建模所需特征波段并建立芝麻油掺伪含量预测模型。采用 tsPSO 对 SiPLS 中子区间数 M 和所选取子区间数 K 进行优选,分别将 M 和 K 设为 tsPSO 中粒子的第一维和第二维,M 变化范围设为[8,20],K 变化范围设为[4,$M/2$],且 M、$K \in Z$,粒子群规模为 5,最大迭代次数为 7 次,一组(M,K)参数决定 C_M^K 种波段组合,定义适应度函数 $fit = f(M,K)$ 用于计算在 C_M^K 种波段组合中最小训练集 *RMSE* 值,每次循环 M 和 K 在指定范围内搜寻最小训练集 *RMSE* 的值,该值对应波段即为用于建模的最优波段。

经过 tsPSO 优化得到参与芝麻油 – 大豆油掺伪定量分析模型建立的特征波段为 1 350 ~ 1 353 nm、1 364 ~ 1 366 nm、1 368 ~ 1 391 nm、1 485 ~ 1 515 nm、1 581 ~ 1 611 nm、1 643 ~ 1 673 nm、1 705 ~ 1 718 nm、1 720 ~ 1 800 nm,如图 4 所示。模型测试结果如表 2 所示。

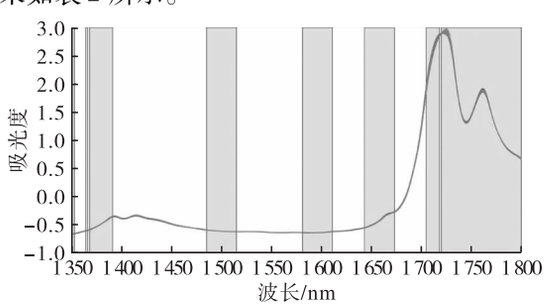

图 4　芝麻油 – 大豆油掺伪样品近红外光谱特征波段

由表 2 可知,参与建模的变量锐减到 219 个,训练集和测试集均方根误差(*RMSE*)分别为 4.39E – 2 和 3.99E – 2,并且相关系数(R^2)均接近 1,相较全

波段和 UVE 模型，显著降低了预测误差，缩短了建模时间。图 4 中灰色区域即为所选用于预测的最优特征波段组合，可见所选特征波段大部分集中在波峰、波谷附近，说明波峰、波谷位置的吸光度比其他波段的差异更为显著，附近范围内的波段更适合用于掺伪含量定量分析模型的建立。

图 5 显示了试验建立的 UVE－SiPLS－tsPSO 芝麻油掺伪样本测试集预测结果与真实值的对比。由图 5 可知，此模型具有很高的预测准确度。

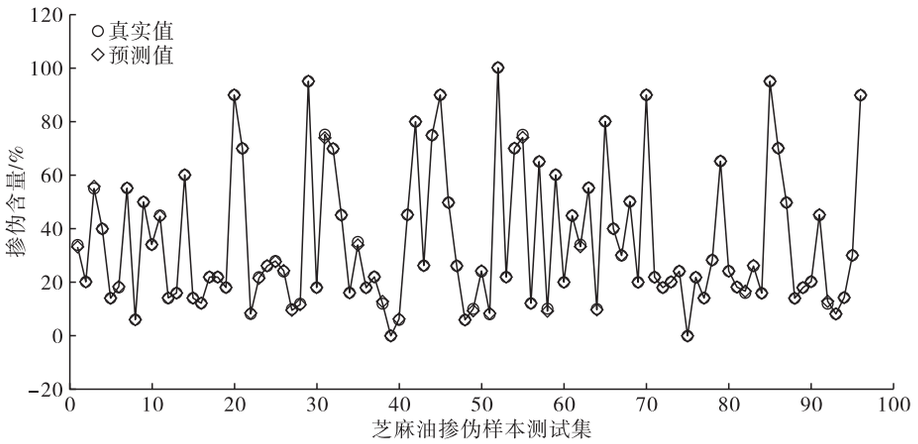

图 5 芝麻油中大豆油掺伪含量预测结果

3 结 论

本文基于芝麻油中掺伪大豆油样本的近红外光谱，首先用 SNV 对光谱进行预处理，再采用无信息变量消除法（UVE）对掺伪芝麻油近红外光谱变量进行初步筛选，然后采用 tsPSO 选取 SiPLS 中的最优特征波段组合建立芝麻油中大豆油掺伪含量快速检测模型。所建模型通过特征波长变量的粗选与细选相结合的方式显著降低了芝麻油中大豆油掺伪含量预测误差，同时减少了建模变量和建模时间。此外，本研究为其他食用植物油的掺伪检测提供了一种可供借鉴的方法，在食用油掺伪研究领域体现出良好的可行性和参考价值。

参考文献：

[1] 余东成. 芝麻的食品科学[J]. 中国油脂, 1990,15(6):2－5.
[2] 毕艳兰,任小娜,彭丹,等. 粒子群最小二乘支持向量机结合偏最小二乘法用于芝麻油质量的鉴别[J]. 分析化学,2013,41(9):1366－1372.
[3] 陈华松,邓素娥,周喜满,等. 芝麻油掺伪通用检验方法研究[J]. 郑州粮食学院学报,1996(1):24－27.
[4] 王乐,胡健华. 食用油掺伪餐饮业废油脂鉴别检测方法研究进展[J]. 中国油脂,2007,32(9):75－77.
[5] 王江蓉,周建平,张令夫,等. 植物油掺伪检测方法的应用与研究进展[J]. 中国油脂,2007,32(6):78－81.
[6] 林丽敏. 气相色谱法测定芝麻油掺伪的研究[J]. 粮食储藏,2006(3):43－45,50.
[7] 王乐,黎勇,胡健华. 核磁共振法鉴别食用植物油掺伪餐饮业废油脂[J]. 中国油脂,2008,33(10):75－77.
[8] GUILLEN M D, CABO N. Some of the most significant changes in the Fourier transform infrared spectra of edible oils under oxidative conditions[J]. J Sci Food Agric, 2000,80:2028－2036.
[9] 杨佳,武彦文,李冰宁,等. 傅里叶变换红外光谱技术在食用油脂分析领域的应用[J]. 中国油脂,2013,38(3):81－86.
[10] 涂斌,宋志强,郑晓,等. 基于激光近红外的稻米油掺伪定性－定量分析[J]. 光谱学与光谱分析,2015,35(6):1539－1545.
[11] 冼瑞仪,黄富荣,黎远鹏,等. 可见和近红外透射光谱结合区间偏最小二乘法(iPLS)用于橄榄油中掺杂煎炸老油的定量分析[J]. 光谱学与光谱分析,2016,36(8):2462－2467.
[12] 丁轻针,刘玲玲,武彦文,等. 基于 FTIR 的芝麻油真伪鉴别和掺伪定量分析模型[J]. 光谱学与光谱分析,2014,34(10):2690－2695.
[13] 尼珍,胡昌勤,冯芳. 近红外光谱分析中光谱预处理方法的作用及其发展[J]. 药物分析杂志,2008,28(5):824－829.
[14] 展晓日,朱向荣,史新元,等. SPXY 样本划分法及蒙特卡罗交叉验证结合近红外光谱用于橘叶中橙皮苷的含量测定[J]. 光谱学与光谱分析,2009,29(4):964－968.
[15] KOSHOUBU J, IWATA T, MINAMI S. Elimination of the uninformative calibration sample subset in the modified UVE－PLS method[J]. Anal Sci, 2001, 17(22):319－322.
[16] NORGAARD L, SAUDLAND A, WAGNER J, et al. Interval partial least－squares regression (iPLS): a comparative chemometric study with an example from near－infrared spectroscopy[J]. Appl Spectrosc, 2000, 54:413－419.
[17] 彭海根,彭云发,詹映,等. 近红外光谱技术结合联合区间间隔偏最小二乘法对南疆红枣糖度的测定[J]. 食品科技,2014,39(6):276－280.
[18] POLI R, KENNEDY J, BLACKWELL T. Particle swarm optimization[J]. Swarm Intell,2007(1):33－57.
[19] 胡旺,李志蜀. 一种更简化而高效的粒子群优化算法[J]. 软件学报,2007(4):861－868.