近红外光谱结合 CARS-PLS 模型检测草莓可溶性固形物含量研究

蔡德玲[1],彭碧宁[1],曾 川[1],梁玉英[1],唐春华[2]
(1.中华人民共和国拱北海关技术中心,广东珠海 519000;2.珠海城市职业技术学院,广东珠海 519090)

摘要 为了实现对草莓内部可溶性固形物含量(soluble solids content,SSC)客观、准确、快速和无损检测,采用近红外光谱结合竞争性自适应重加权算法采样(CARS)变量选择以及多变量校正分析的测定方法。164 个草莓样本被分成校正集(123 个)和预测集(41 个)。基于全光谱数据,通过 CARS 算法获得了可以表征原始光谱信息的 117 个特征光谱变量。全光谱变量和特征光谱变量分别作为输入构建了偏最小二乘回归 PLS 和多元线性回归 MLR 模型,通过比较 3 类模型发现,基于特征光谱的 PLS 模型(即 CARS-PLS 模型)对草莓内部可溶性固形物含量测定性能最优,针对预测集样本,模型预测相关系数 r_p 和均方跟误差 RMSEP 分别为 0.950 9 和 0.335 2。

关键词 草莓;近红外光谱;CARS-PLS 模型;可溶性固形物含量;无损检测;光谱分析
中图分类号 TS 255.7 **文献标识码** A
文章编号 0517-6611(2020)08-0185-04
doi:10.3969/j.issn.0517-6611.2020.08.045
开放科学(资源服务)标识码(OSID):

Determination of Soluble Solids Content in Strawberry by Near Infrared Spectroscopy Combined with CARS-PLS Model

CAI De-ling, PENG Bi-ning, ZENG Chuan et al (Technical Center of Gongbei Customs District P.R., Zhuhai, Guangdong 519000)

Abstract In order to realize the objective, accurate, rapid and nondestructive detection of the soluble solids content (SSC) in strawberry, in this study, the near-infrared spectroscopy combined with competitive adaptive reweighted sampling (CARS) variable selection and multivariate calibration analysis method was proposed. 164 strawberry samples were divided into correction set (123 samples) and prediction set (41 samples). Based on the full spectral data, CARS algorithm selected 117 characteristic variables which could represent the original spectral information. Partial least square regression (PLS) model and multivariate linear regression (MLR) model were constructed by using full spectrum variables and characteristic spectral variables, respectively. By comparing three types of models, it was found that PLS (CARS-PLS model) based on characteristic variables had the best performance for determination of soluble solid content in strawberry. For the samples in prediction set, the correlation coefficient (rP) and mean square error of prediction (RMSEP) of model were 0.950 9 and 0.335 2, respectively.

Key words Strawberry; Near infrared spectroscopy; CARS-PLS model; Soluble solids content; Nondestructive detection; Spectral analysis

水果是一种重要且非常受欢迎的农产品,水果中含有丰富的且有益于人体健康的营养元素。每年世界各地都要消耗大量的新鲜水果。目前,在水果品质方面,消费者不仅注重大小、颜色、形状等外在品质,更注重含糖量、酸度、硬度等内在品质[1-2]。基于水果内在品质的检测和分级一直是果实采摘后商品化处理的重要一环。适当的内部品质分级不仅可以延长水果的贮藏期,而且可以提高其市场竞争力和经济价值[3-5]。可溶性固形物含量(soluble solids content,SSC)是影响鲜果品质和价格的重要内在品质属性之一[6]。该参数也是决定果实成熟度和收获时间的关键因素[6-7]。水果可溶性固形物含量检测通常是采用数字折光仪的常规分析,该方式耗时且需对水果进行破坏性检测[8-9]。这种方式适用于抽样检测,它不能满足消费者对整批次水果一致性和高品质的要求。

近红外(near-infrared,NIR)光谱技术具有快速、易操作和无损的分析特征,是目前最广泛使用的水果内部质量无损检测技术[10],该技术已经成功用于水果中 SSC 的检测,涉及水果包括苹果[11]、梨[12]、葡萄[13]、枣[14]、猕猴桃[15]、橙子[16]、香蕉[17]、西瓜[18]等。草莓是最常见的一种水果,其营养价值丰富,被誉为是"水果皇后"。但是目前针对草莓内部品质检测的研究非常少,因此,草莓内部品质尤其是可溶性固形物含量的快速、无损检测技术研究对提升草莓质量和采后分级具有重要意义。该研究将采用近红外光谱技术,构建草莓内部可溶性固形物含量定量分析模型,并采用自适应重加权采样(competitive adaptive reweighted sampling,CARS)算法对模型进行优化。

1 材料与方法

1.1 试验样本及样本集划分 新鲜草莓采摘后迅速运至实验室,选择果形、大小相对一致且表面无损伤的草莓作为样本放入冷藏室。试验前,从冷藏室取出并在室温中(20±1℃)放置超过 24 h,以消除温度对预测模型精度的影响。共计样本数为 164 个。

164 个样本按照 SSC 浓度值进行从小到大排序;然后每 4 个样本中选取第 2 个样本作为预测集样本,这样预测集中包含 41 个样本用于校正模型的评估,剩余 123 个草莓作为校正集样本用于校正模型的构建。在模型开发的过程中,所有模型校正集样本和预测集样本保持不变。

1.2 仪器和近红外光谱获取 原始草莓样本近红外光谱采集采用 AntarisTM Ⅱ 傅立叶变换近红外光谱仪(Thermo Fisher Scientific Inc.,Madison,WI,USA)。采集模式为漫反射模式。每个样本采集并获取一条光谱曲线,该光谱曲线的范围为 12 000~3 800 cm^{-1},相邻之间的间隔为 1.928 cm^{-1},因此,每条光谱曲线包含 4 254 个变量点。采集完光谱之后,采用 Unscrambler V9.7 software (CAMO PRECESS AS, Oslo, Norway)软件将原始反射光谱转换为吸收光谱。

1.3 SSC 参考值破坏性测量 草莓样本 SSC 参考值通过破坏性检测获得。整个草莓样本去除果梗进行榨汁后使用数字显示手持型折射计(型号:PR-101α,Atago Co.,Ltd.,Toky-

基金项目 中华人民共和国拱北海关科研项目(ZH2017-29)。
作者简介 蔡德玲(1983—),女,四川大竹人,工程师,硕士,从事食品与农产品检测技术研究。
收稿日期 2019-10-22;**修回日期** 2019-12-18

o,Japan)进行测量,3 次测量并进行读数,3 次读数的均值即为该样本最终 SSC 参考值。

1.4 变量选择算法 由于该研究光谱变量非常多,太多的变量一方面会增加模型的复杂性,降低模型的定量预测性能,另一方面模型的构建和评估都需要花费较长的时间,这不利于快速预测模型的开发。因此,该研究基于原始光谱数据,采用竞争性自适应重加权算法采样(CARS)算法进行特征光谱变量选择。CARS 变量选择算法是建立在模仿达尔文进化理论中"适者生存"的原则基础上提出的变量选择方法[19]。每次利用指数衰减函数(exponentially decreasing function,EDP)和自适应重加权采样技术(adaptive reweighted sampling,ARS)结合的方法优选出 PLS 模型中回归系数绝对值大的变量点,去除权重值较小的变量点,利用十折交叉验证选出 N 个 PLS 子集模型中 RMSECV 最小的子集,该子集所包含的变量即为最优变量组合。

1.5 多变量校正模型构建 在该研究中,构建两类型线性模型即偏最小二乘(partial least squares,PLS)和多元线性回归(multiple linear regression,MLR)模型用于草莓可溶性固形物含量定量预测。

PLS 是目前光谱分析中广泛使用的回归方法。PLS 同时考虑了目标化学性质矩阵 Y(SSC 值)和变量矩阵 X(光谱数据),找出 Y 和 X 之间的基本关系。利用 PLS 作为回归方法提取潜在变量(LVs)。将 LVs 作为原始光谱的新特征向量,降低了原始光谱的维数,压缩了原始光谱数据。在 PLS 模型的开发过程中,采用全交叉验证的方法,通过交叉验证的均方根误差(RMSECV)来确定 LVs 的最优数目,以防止过拟合问题。偏最小二乘法特别适用于变量多于样本的情况,以及变量之间存在多重共线性的情况。

MLR 是另一种常用的多变量线性校正算法。该算法简单且容易解释,但其很容易受变量之间的共线性影响。此外,当变量多于样本时,它也会失效。该算法类似 PLS,也可以同时考虑了目标化学性质矩阵 Y(SSC 值)和变量矩阵 X(光谱数据),找出 Y 和 X 之间的基本关系。

1.6 模型预测性能评估 所有模型的预测性能通过相关系数 correlation coefficient(r)、建模均方根误差(root mean square error of calibration,RMSEC)、预测均方根误差(root mean equare error of prediction,RMSEP)参数进行评估。一个好的模型通常具有低的 RMSEC 和 RMSEP 值,高的 r 值。评估参数计算公式:

$$r_C^2, r_P^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \quad (1)$$

$$\mathrm{RMSEC}, \mathrm{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

2 结果与分析

2.1 光谱和 SSC 实测值分析 图 1a 为所有样本原始近红外光谱曲线。尽管光谱曲线中存在着一些交叉与重叠,但所有样本光谱有着类似的变化趋势,表明所测样本光谱数据不存在异常样本。从光谱曲线中可以看出,波数较大时(如大于 7 000 cm^{-1}),光谱吸收强度更大,这主要原因在于在波数较大区域存在着明显的 H_2O 吸收(如位于 6 944 和 5 155 cm^{-1}的吸收峰)。另外光谱曲线中也存在一些小的吸收峰如 8 403 cm^{-1},这些吸收峰与 C-H 二级倍频有关系。所有的这些吸收特性均有助于草莓内部 SSC 的预测。同时,在图 1a 的光谱图中也可以看出原始光谱存在着明显的光谱散射,因此,在进一步模型构建之前,原始光谱首先进行多元散射校正(multiplicative scatter correction,MSC)校正预处理。图 1b 为预处理后的光谱曲线,可以看出,预处理后光谱散射得到了明显的改善。

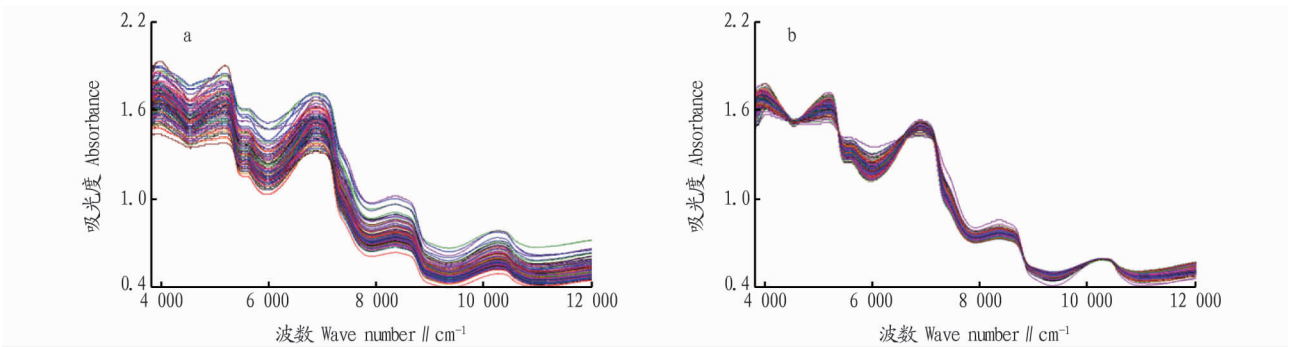

图 1 草莓样本原始光谱(a)和预处理光谱(b)

Fig.1 Raw spectrum (a) and pre-processed spectrum (b) of strawberry samples

从不同数据集中草莓样本可溶性固形物含量 SSC 统计值(表 1)可以看出,校正样本集 SSC 值为 6.18~13.57 Brix,预测样本集 SSC 值为 6.50~13.10 Brix,预测集样本其 SSC 值范围包含在校正集样本 SSC 值范围内,这有助于构建一个稳健的草莓可溶性固形物含量预测模型。

2.2 CARS 变量选择 图 2 表示基于全光谱变量数据,采用 CARS 算法(设置 MC 采样次数为 50 次)进行变量选择后获得的结果。图 2a、b 和 c 分别表示在 1 次 CARS 算法运行中随着 MC 采样次数的增加,变量数、十折交叉验证 RMSECV 值和每个变量回归系数的变化。从图 2a 可以看出,由于指数衰减函数 EDP 的作用,变量数在前 20 次 MC 采样中下降非常快,随后逐渐减缓慢并趋于平稳,表明 CARS 算法在特征变量选取中具有"粗选"和"精选"2 个过程。从图 2b 可以看出,起初阶段,由于大量与草莓内部 SSC 预测无关的变量

被剔除导致单个 PLS 模型的十折交叉验证 RMSECV 值随着 MC 采样次数的增加逐渐变小，当 RMSECV 达到最低值时，所对应变量 MC 采样次数为 24 次（图 2c 中星号垂线标示），随着采样次数的进一步增加，RMSECV 也增加，表明光谱中的某些重要变量被剔除，因此，第 24 次 MC 采样后获得的变量确定为预测草莓 SSC 含量的特征变量，共计 117 个变量。

2.3　模型预测结果　基于校正集草莓样本，117 个被选取的特征变量和校正集样本的 SSC 值作为输入构建 PLS 和 MLR 模型（即CARS-PLS模型和CARS-MLR模型），采用预测集 41 个样本对所构建的模型进行预测性能评估，评估结果如表 2 所示。为了与全光谱模型进行比较，全光谱 4 254 个变量也作为输入构建了 PLS 模型（Full-PLS 模型）由于全光谱变量多于建模样本数，所以无法构建全光谱 MLR 模型，模型的预测结果也列于表 2 中。从表 2 可以看出，针对校正集样本，Full-PLS 模型的校正相关系数 r_C 和建模均方根误差 RMSEC 分别为 0.954 2 和 0.344 2，针对预测集样本，Full-PLS 模型的预测相关系数 r_P 和校正均方根误差 RMSEP 分别为 0.752 3 和 0.862 1；对于 CARS-PLS 模型，r_C 和 RMSEC 分别为 0.974 7 和 0.306 9，r_P 和 RMSEP 分别为 0.950 9 和 0.335 2；对于 CARS-MLR 模型，r_C 和 RMSEC 分别为 0.970 8 和 0.235 2，r_P 和 RMSEP 分别为 0.822 7 和 0.788 4。

2.4　模型比较分析　从表 2 可以看出，Full-PLS 模型由于太多的光谱变量参与模型的构建，可能使模型出现过拟合而降低了模型对外部样本预测精度，针对预测集 RMSEP 达到了 0.862 1，明显高于 CARS-PLS 和 CARS-MLR 模型。基于 117 个特征变量所构建的 CARS-PLS 和 CARS-MLR 模型，其预测能力明显高于全光谱 PLS 模型，表明 CARS 算法能够有效识别光谱中的有效变量。进一步比较 CARS-PLS 和 CARS-MLR 模型发现，前者预测性能优于后者，可能是由于 PLS 模型能够更好地处理光谱变量与草莓可溶性固形物之间的关系。综合来看，在所构建的 3 类模型中，CARS-PLS 模型对草莓内部 SSC 的评估性能最优，然而仅仅采用了原始光谱 2.75%的光谱变量，因此，相对全光谱模型，该模型是一个极简的预测模型，应该具有更快的光谱建模和预测速度。图 3 列出了 CARS-PLS 模型对校正集样本和预测集样本的预测散点图。从图 3 可看出，样本分布在回归曲线附近，且接近回归曲线，表明 CARS-PLS 模型能够准确预测草莓内部的可溶性固形物含量。

表 1　草莓样本不同数据集可溶性固形物含量统计

Table 1　Statistics of soluble solids content in different datasets of strawberry samples　Brix

数据集 Dataset	样本数 Number of samples	最小值 Minimum	最大值 Maximum	均值 Mean	标准差 Standard deviation
样本总数 Total number of samples	164	6.18	13.57	10.09	1.02
校正集 Correction set	123	6.18	13.57	10.39	1.11
预测集 Prediction set	41	6.50	13.10	10.25	1.07

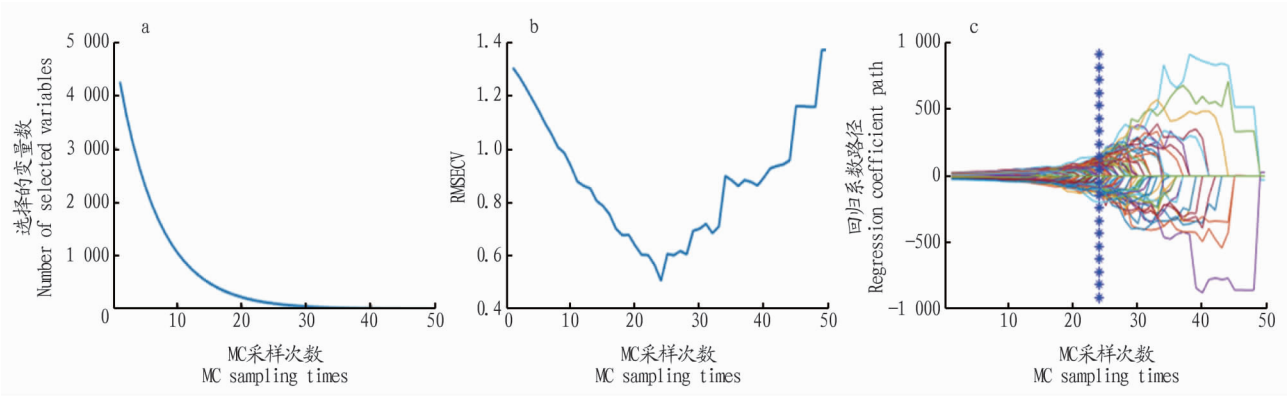

图 2　CARS 特征变量选择结果

Fig.2　Characteristic variable selection by CARS

表 2　基于不同变量的草莓 SSC 含量 PLS 模型预测结果

Table 2　Prediction results for SSC in strawberry by PLS models developed based on different variables

建模方法 Modeling method	变量数 Number of variables	LVs	校正集 Correction set		预测集 Prediction set	
			r_C	RMSEC	r_P	RMSEC
Full-PLS	4 254	15	0.954 2	0.344 2	0.752 3	0.862 1
CARS-PLS	117	15	0.974 7	0.306 9	0.950 9	0.335 2
CARS-MLR	117	—	0.970 8	0.235 2	0.822 7	0.788 4

3　小结

该研究采用近红外光谱技术结合竞争性自适应重加权算法采样（CARS）变量选择算法以及 PLS 建模分析方法成功实现了对草莓内部可溶性固形物含量 SSC 的有效定量分析。基于 CARS 算法获得了可以表征全部光谱分析的 117 个特征变量，并构建了基于特征变量的 CARS-PLS 模型和CARS-MLR 模型，结果表明特征变量模型性能明显优于全光谱 PLS 模型，一方面说明了 CARS 算法能够有效用于草莓近红外光谱变量的选择，另一方面也说明了通过合适变量选择能够有效提高模型的预测性能。通过比较所有模型的预测结果，最终确定 CARS-PLS 模型为草莓内部 SSC 预测的最佳模型。后续需要进一步增加样本量，提升模型在实际应用中的稳健性。

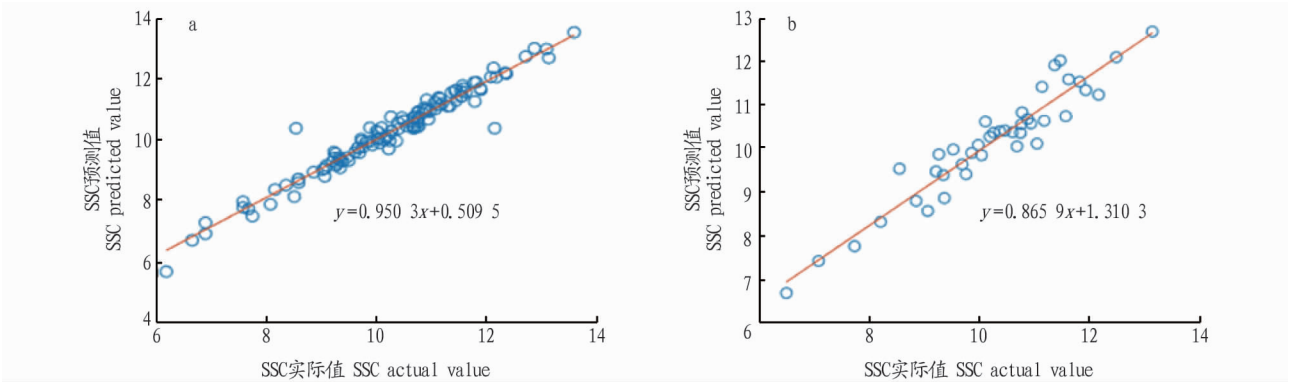

图 3 CARS-PLS 模型对校正集样本(a)和预测集样本(b)的预测结果散点图

Fig.3 Scatter plots of the prediction results of the CARS-PLS model on the calibration set samples (a) and the prediction set samples (b)

参考文献

[1] XIE L J,WANG A C,XU H R,et al.Applications of near-infrared systems for quality evaluation of fruits:A Review [J].Transactions of the ASABE,2016,59(2):399-419.

[2] PARK E,LUO Y G,MARINE S C,et al.Consumer preference and physicochemical evaluation of organically grown melons [J].Postharvest biology and technology,2018,141:77-85.

[3] BURDON J,PIDAKALA P,MARTIN P,et al.Fruit maturation and the soluble solids harvest index for 'Hayward' kiwifruit [J].Scientia horticulturae,2016,213:193-198.

[4] OH S B,MUNEER S,KWACK Y B,et al.Characteristic of fruit development for optimal harvest date and postharvest storability in 'Skinny Green' baby kiwifruit [J].Scientia horticulturae,2017,222:57-61.

[5] JIANG B,HE J R,YANG S Q,et al.Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues [J].Artificial intelligence in agriculture,2017,1:1-8.

[6] LI J L,SUN D W,CHENG J H.Recent advances in nondestructive analytical techniques for determining the total soluble solids in fruits:A review [J].Comprehensive reviews in food science and food safety,2016,15(5):897-911.

[7] 马本学,应义斌,饶秀勤,等.高光谱成像在水果内部品质无损检测中的研究进展[J].光谱学与光谱分析, 2009,29(6):1611-1615.

[8] KAWANO S,ABE H,IWAMOTO M.Development of a calibration equation with temperature compensation for determining the Brix value in intact peaches [J].Journal of near infrared spectroscopy,1995,3(4):211-218.

[9] 李江波,彭彦昆,陈立平,等.近红外高光谱图像结合 CARS 算法对鸭梨 SSC 含量定量测定[J].光谱学与光谱分析, 2014,34(5):1264-1269.

[10] 饶利波,陈晓燕,庞涛.基于光谱技术的 BiPLS 算法结合 CARS 算法的苹果可溶性固形物含量检测[J].发光学报,2019,40(3):389-395.

[11] TIAN X,FAN S X,LI J B,et al.Comparison and optimization of models for SSC on-line determination of intact apple using efficient spectrum optimization and variable selection algorithm [J].Infrared physics and technology,2019,102:1-11.

[12] PAZ P,SANCHEZ M T,PEREZ-MARIN D,et al.Instantaneous quantitative and qualitative assessment of pear quality using near infrared spectroscopy [J].Computers and electronics in agriculture,2009,69:24-32.

[13] CAO F,WU D,H E Y.Soluble solids content and pH prediction and varieties discrimination of grapes based on visible-near infrared spectroscopy [J].Computers and electronics in agriculture,2010,71S:S15-S18.

[14] WANG J,NAKANO K,OHASHI S.Nondestructive evaluation of jujube quality by visible and near-infrared spectroscopy[J].LWT- Food Science and Technology,2011,44:1119-1125.

[15] MOGHIMI A,AGHKHANI M H,SAZGARNIA A,et al.Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit[J].Biosystems engineering,2010,106(3):295-302.

[16] 刘燕德,施宇,蔡丽君,等.基于 CARS 算法的脐橙可溶性固形物近红外在线检测[J].农业机械学报,2013,44(9):138-144.

[17] JAISWAL P,JHA S N,BHARADWAJ R.Non-destructive prediction of quality of intact banana using spectroscopy [J].Scientia horticulturae,2012,135:14-22.

[18] 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

[19] 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,2009,648:77-84.

(上接第 174 页)

[9] 张健,刘少伟,张毅,等.仿刺参精酶解工艺条件优化及体外抗氧化[J].食品工业科技,2017,38(5):232-237.

[10] 原姣姣,陈锦璇,张帆,等.响应面优化超声_酶辅助强化油橄榄叶多糖的提取[J].中国油脂,2019,44(4):128-132.

[11] 王呈文,纪明慧,舒火明,等.牛大力总黄酮提取工艺及不同萃取物的抗氧化活性研究[J].化学研究与应用,2013,25(5):713-717.

[12] 王呈文,纪明慧,陈光英,等.热带莫氏兰根提取物的抗氧化活性及稳定性研究[J].食品工业科技,2013,34(5):209-211,217.

[13] 王国良,李建科,吴晓霞,等.水麻果多酚的提取纯化及其抗氧化、抗肿瘤活性作用[J].天然产物研究与开发,2019,31(1):1-9.

[14] GE Y,DUAN Y F,FANG G Z,et al.Polysaccharides from fruit calyx of *Physalis alkekengi* var.*francheti*:Isolation,purification,structural features and antioxidant activities [J].Carbohydrate polymers,2009,77:188-193.

[15] 楚秉泉,方若思,李玲,等.洋甘菊各萃取相抗氧化活性及其有效成分分析[J].食品工业科技,2019,40(8):1-6.