doi：10.19969/j.fxcsxb.21111803

近红外光谱结合无参数校正增强实现不同年份烟叶总糖含量模型更新

耿莹蕊[1]，沈欢超[1,2]，倪鸿飞[2]，王 辉[3]，吴继忠[3]，张立立[3]，李永生[3]，何文苗[3]，陈 勇[1]，刘雪松[1*]

（1. 浙江大学 药学院，浙江 杭州 310058；2. 浙江大学 智能创新药物研究院，浙江 杭州 310018；3. 浙江中烟工业有限责任公司技术中心，浙江 杭州 310008）

摘 要：近红外光谱技术因快速、无损等特点，已广泛应用于烟草行业质量快速分析。然而，由于采收时间、环境差异等因素的影响，建立的近红外定量模型在新批次样本中的预测性能通常变差，因此必须对原有模型进行维护和更新。该研究采用半监督无参数校正增强(SS－PFCE)方法，通过约束优化，对主模型的回归系数进行修正。首先建立了2016年烟叶样本总糖含量的原始定量模型，其预测相关系数(Rp)为0.997 8、预测均方根误差(RMSEP)为0.310 8。采用SS－PFCE方法对模型更新后，分别预测2017年、2018年和2020年样本的总糖含量，3个测试集的Rp值比未更新模型提高了0.13%、1.32%和4.29%，RMSEP分别下降了15.26%、58.69%和36.53%。与重新建立的定量分析模型相比，更新后的模型具有更优的预测性能，同时大大降低了建模成本。研究表明，SS－PFCE方法可高效地实现不同年份烟叶样本的模型维护，在实际生产中具有重要的应用价值。

关键词：近红外光谱技术；模型更新；烟叶；半监督无参数校正增强(SS－PFCE)

中图分类号：O657.33；O629.1 **文献标识码**：A **文章编号**：1004-4957(2022)07-1066-06

Model Update of Total Sugar Content in Tobacco Leaves of Different Years by Near-infrared Spectroscopy Combined with Parameter-free Calibration Enhancement

GENG Ying-rui[1]，SHEN Huan-chao[1,2]，NI Hong-fei[2]，WANG Hui[3]，WU Ji-zhong[3]，ZHANG Li-li[3]，LI Yong-sheng[3]，HE Wen-miao[3]，CHEN Yong[1]，LIU Xue-song[1*]

（1. College of Pharmaceutical Sciences，Zhejiang University，Hangzhou 310058，China；2. Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University，Hangzhou 310018，China；3. Technology Center，China Tobacco Zhejiang Industrial Co.，Ltd.，Hangzhou 310008，China）

Abstract：The near infrared spectroscopy technology has been widely applied to the quantitative analysis in tobacco industry because of its advantages of rapidness and non-destructiveness. However，the accuracy and robustness of the original model may deteriorate when predicting samples with new variations. When samples are obtained from different harvest years and different environmental conditions，new variations will be introduced. Therefore，there is a need to maintain the predictive performance of the original model when it works on a new batch. In this study，a method called semi-supervised parameter-free calibration enhancement(SS－PFCE) was used to update the original model. The regression coefficient of original model was corrected by constrained optimization. The original model for total sugar determination was firstly developed with the tobacco samples of 2016，which showed a predicted correlation coefficient(Rp) of 0.997 8 and a root mean square error of prediction(RMSEP) of 0.310 8. After updating the model by the SS－PFCE method，the total sugar contents in the samples of 2017，2018 and 2020 years were predicted，respectively. The Rp values of the three test sets were improved by 0.13%，1.32% and 4.29%，and the RMSEP were decreased by 15.26%，58.69% and 36.53%，respectively，compared with the non-updated model. Furthermore，the updated model by the SS－PFCE approach offered a better predictive

收稿日期：2021－11－18；**修回日期**：2022－01－20

基金项目：浙江大学－浙江中烟联合实验室项目资助

*通讯作者：刘雪松，博士，研究员，研究方向：现代制药工程与医药智能制造，E－mail：liuxuesong@zju.edu.cn

performance than re-modeling method, while significantly reduced modeling costs. The results of this study showed that the SS – PFCE method could maintain the prediction accuracy for tobacco samples of different years efficiently, and it is of great practical application value in industrial production.

Key words: near infrared spectroscopy; model update; tobacco; semi-supervised parameter-free calibration enhancement(SS – PFCE)

在过去几十年，近红外光谱(NIRS)技术因具有快速高效、无损、低成本的优势，已成功应用于各个领域[1-2]。烟草作为一种复杂的天然产物，利用其近红外光谱信息结合化学计量学方法可实现烟叶定量分析、品质分类、质量控制等，具有重要的应用价值[3-5]。

建立一个稳健准确且能在实际生产中应用的校正模型是NIRS技术的关键，目前常用的模型校正方法有偏最小二乘回归(PLSR)[6]、最小二乘支持向量机(LS – SVM)[7]、人工神经网络(ANN)[8]等，通过这些方法建立校正模型，可实现未知样品目标成分的定量分析。但检测条件、仪器状态以及烟叶培育环境的变化均会造成光谱特性与质量属性的差异[9-10]，这些新变化可导致原有模型预测能力下降。

为维持原始模型在新样本中的良好预测性能，目前已开发了较多的模型转移算法[11-12]。传统的模型转移方法侧重于对数据的调整和修正，如分段直接标准化(PDS)[13]、斜率/截距修正算法[14]等，此类方法对不同仪器间的模型转移效果显著，但其标准样品的选择和获取在实际应用中存在一定难度，因此有标样的模型转移算法应用存在局限性。模型转移的第二种途径是模型更新[15]，即添加新样本进行校正，优化现有的模型[16]，该方法往往需要挑选具有代表性的样本，考虑新样本权重以优化模型[17-18]。此外，还有一些算法可通过消除外部影响因素达到模型更新的目的，但这类方法涉及大量参数的调整和优化[19-21]，对日常使用而言复杂耗时。

为解决上述方法的不足，有学者提出一种无参数校正增强框架(PFCE)算法[22]，其通过对回归系数进行相关性约束，从而增强原始模型对新样本的预测能力[23]。该方法不仅减少了对标准样品的需求，还省去模型更新需要多参数优化的步骤，大大提高了模型的更新效率。本文旨在通过PFCE模型更新策略消除采收时间对烟叶总糖含量预测结果的影响，以期维持主模型在不同年份烟叶样本中定量分析的性能。

1　实验部分

1.1　数据采集及参考值的测定

本研究使用的烟叶样本分别采收于2016年、2017年、2018年以及2020年，均由浙江中烟工业有限责任公司提供。烟叶样本在相同测试条件下采用Antaris II FT – NIR(Thermo Fisher Scientific)分析仪进行光谱测量，光谱的采集范围为10 000 ~ 3 800 cm^{-1}，分辨率为8 cm^{-1}，每个光谱包含1 609个变量。本研究选择烟叶中总糖含量建立定量分析模型，样品的参考值由浙江中烟技术中心依照烟草标准YC/T159—2002测定[24]。

1.2　实验设计及软件

采用2016年烟叶样本建立总糖含量预测的PLSR主模型，以2017年、2018年和2020年样品的光谱用于校正和更新主模型。主模型样本使用基于$x-y$距离样本集划分(SPXY)算法划分为校正集(70%)和测试集(30%)，用于更新主模型的样本划分为模型更新集(30%)和测试集(70%)。在研究中，采用半监督无参数校正增强(SS – PFCE)方法对主模型进行更新，另外比较了2017、2018及2020年样本重新建模的效果。对于所有定量模型，使用校正相关系数(Rc)、预测相关系数(Rp)、校正均方根误差(RMSEC)、预测均方根误差(RMSEP)和残差预测偏差(RPD)对其性能进行评价[25]。

所有算法和画图操作均使用MATLAB R2018 b软件完成。

1.3　理论与算法

1.3.1　偏最小二乘回归(PLSR)　PLSR是一种经典的定量建模方法，它将m个样本在n个变量处的光谱X与m个样本的相关参考值Y投影到新空间中构建线性回归模型。在本文中，采用留一交叉验证方法确定PLS模型中的最佳潜在变量(Latent variables，LV)数[26]。

1.3.2 半监督无参数校正增强(SS – PFCE) 用于光谱校正增强的无参数框架(PFCE)是Zhang等[22]最新提出的模型维护方法，其根据模型传递中标准品的有无分为非监督PFCE(NS – PFCE)、半监督PFCE(SS – PFCE)和全监督PFCE(FS – PFCE)。其中SS – PFCE方法仅需新样品的部分光谱和属性参考值对模型进行校正更新，无需额外挑选标准品，SS – PFCE的目标函数采用公式(1)进行计算。

$$\min_{b_{0,\ new},\ b_{new}} \left(\left\| y_{new} - [1\ X_{new}] \begin{bmatrix} b_{0,\ new} \\ b_{new} \end{bmatrix}^2 \right\| \right) \tag{1}$$

$$s.t.\ corr(b_{new},\ b_m) > r_{th} \tag{2}$$

公式(1)中，X_{new}代表新批次样本中被选为更新集的光谱，y_{new}表示参考值，$b_{0,new}$和b_{new}分别表示更新模型的截距和回归系数；公式(2)中，b_m代表主模型的回归系数，为约束新旧模型回归系数的阈值，保证更新模型获得适当的回归系数和截距，已有研究均将阈值设定为0.98[22]。

使用SS – PFCE方法实现不同年份烟叶模型的更新可概括为以下3个步骤：

(1) 选择某一年份样本的光谱，构建PLSR主模型，从中获得主模型回归系数b_m。

(2) 使用新年份样本的部分光谱和参考值对主模型进行维护和校正，从主模型回归系数b_m中得到新模型的b_{new}。

(3) 用新样本测试集的光谱验证更新后的模型，以RMSEP和R_p对模型更新效果进行评估。

2 结果与讨论

2.1 不同年份烟叶样本的近红外平均光谱

不同年份烟叶样本的近红外平均光谱如图1所示。不同年份烟叶样本具有相似的吸收峰趋势，但吸收强度存在差异，说明烟叶的光谱信息很大程度上受采收年份的影响。

表1数据表明，不同年份烟叶中总糖含量差异较大，除2020年外，2017年和2018年烟叶样本的总糖含量均超出2016年总糖含量的覆盖范围。结合图1可知，不同采收年份造成样本的化学信息和光谱特征产生差异，这些差异可能严重影响主模型预测新样本的准确性，因此需进行模型维护以提高主模型的稳健性。

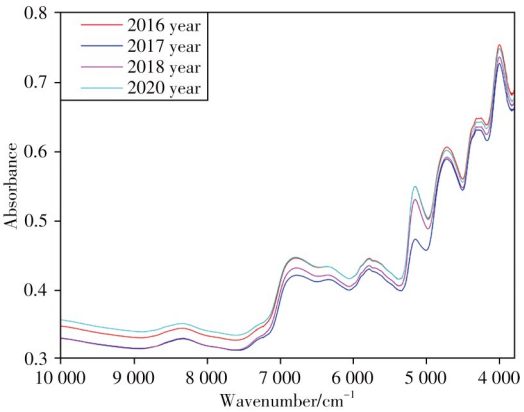

图1 不同年份烟叶样本的近红外平均光谱图

Fig. 1 The average raw NIR spectra of tobacco samples in different years

2.2 烟叶样本主模型的建立

表1 不同年份烟叶样本汇总

Table 1 Summary of tobacco samples in different years

Year	2016	2017	2018	2020
Total of samples	193	180	451	122
Total sugar content range(%)	12.9 ~ 42.9	17.2 ~ 44.5	9.21 ~ 44.0	17.0 ~ 38.6

采用SPXY方法将2016年193个烟叶样本按照7∶3的比例划分为校正集和测试集，划分结果及总糖含量汇总于表2。通过内部交叉验证，以最小的交叉验证均方根误差(RMSECV)为指标，确定最优潜在变量数(LV)，建立2016年烟叶的PLSR主模型。模型预测性能如表3所示，可以看出，主模型R_p值接近1，说明模型预测结果与参考值相关性很高，RMSEP值较小，RPD大于15，证明主模型性能较优，可实现相同年份间烟叶总糖含量的准确预测。

表2 主模型样本的划分结果

Table 2 Statistics of reference quality measurements for tobacco samples

Year	Training set			Validation set		
	Number of samples	Minimum value(%)	Maximum value(%)	Number of samples	Minimum value(%)	Maximum value(%)
2016	145	12.9	42.9	48	21.1	41.0

表3　主模型总糖含量的预测性能

Table 3　Total sugar content prediction performance obtained by original PLSR model

Year	Training set			Validation set		
	LV	*R*c	RMSEC	*R*p	RMSEP	RPD
2016	25	0. 999 5	0. 197 1	0. 997 8	0. 310 8	15. 25

2. 3　SS – PFCE 模型更新

将2017年、2018年以及2020年的烟叶样本按照“1. 2”所述进行样本划分，更新集参与SS – PFCE方法对主模型回归系数的校正，划分结果汇总于表4。使用“2. 2”中2016年样本建立的主模型分别对2017、2018及2020年的样本进行总糖含量预测，图2展示了2016年主模型更新前对不同年份烟叶总糖的预测结果。

表4　用于模型更新的样本划分结果

Table 4　A summary of total sugar content range for model updating and testing sets for different years of tobacco

Year	Model updating			Testing set		
	Number of samples	Minimum value(%)	Maximum value(%)	Number of samples	Minimum value(%)	Maximum value(%)
2017	45	17. 2	44. 5	135	19. 4	43. 5
2018	113	9. 2	44. 0	338	13. 6	43. 9
2020	31	20. 7	38. 6	91	21. 5	37. 5

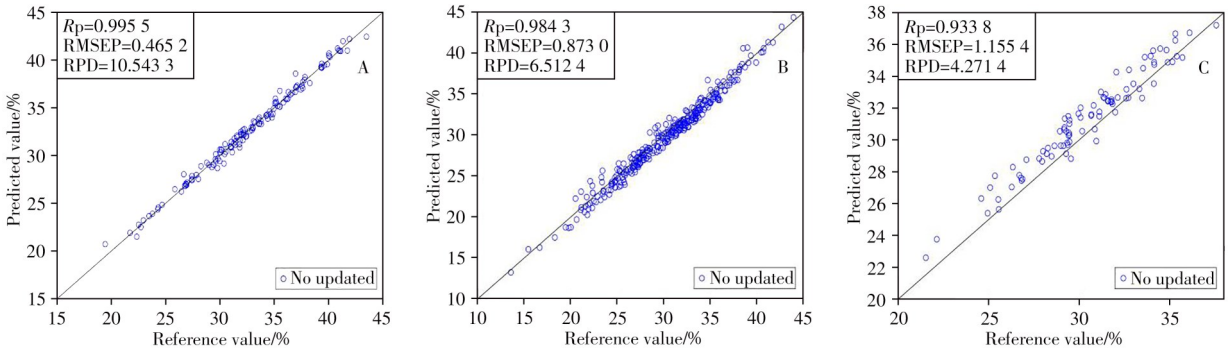

图2　采用2016年主模型预测2017年(A)、2018年(B)及2020年(C)烟叶样本的总糖含量

Fig. 2　Total sugar contents of tobacco made on samples in 2017(A), 2018(B) and 2020(C) years predicted with the master model for 2016 year

以2016年样本建立的主模型*R*p值为0. 997 8，RMSEP值为0. 310 8，而使用该模型直接预测其他年份样品时，*R*p值下降，RMSEP值升高，预测能力均下降(见表5)。结合表2和表4的数据，尽管2020年样本的总糖含量未超出2016年主模型的定量范围，但模型的预测效果下降，说明即使预测集样本含量在模型定量范围内，由于样本批次差异，其预测准确度难以保证，因此需对主模型进行模型更新以适用新批次样本的定量分析。

表5　采用SS – PFCE方法模型更新后对不同年份的预测效果

Table 5　Prediction effects of SS – PFCE method on model updating in different years

Original model	Testing set	Without update			Updated with SS – PFCE		
		*R*p	RMSEP	RPD	*R*p_ss	RMSEP_ss	RPD_ss
2016	2017	0. 995 5	0. 465 2	10. 54	0. 996 8	0. 394 2	12. 67
	2018	0. 984 3	0. 873 0	6. 51	0. 997 3	0. 360 6	13. 72
	2020	0. 933 8	1. 155 4	4. 27	0. 973 9	0. 733 3	4. 49

表5结果显示，使用SS – PFCE方法更新后，主模型对3个年份的预测结果均明显提高，2017年、2018年和2020年的*R*p值分别升高了0. 13%、1. 32%和4. 29%，RMSEP值分别下降了15. 26%、58. 69%和36. 53%，证明SS – PFCE方法对主模型进行更新后，可提高新批次样本的预测准确性。

2. 4　与重新建模方法的比较

为进一步验证SS – PFCE方法对主模型的更新效果，使用表4中2017、2018和2020年的更新集分别重新建立定量校正模型，模型预测性能与SS – PFCE更新结果的对比如表6所示。数据表明，相比于重新建模，采用SS – PFCE方法对主模型进行更新后，3个年份测试集的*R*p值分别升高3. 53%、0. 25%、3. 01%，RMSEP值分别下降70. 24%、28. 69%和30. 32%，表明模型预测性能有大幅提升。

表6 模型预测性能对比

Table 6 Total sugar content predictions for different years based on two different model updating approaches

Original model	Testing set	Rebuild models			Updated with SS – PFCE		
		Rp_re	RMSEP_re	RPD_re	Rp_ss	RMSEP_ss	RPD_ss
2016	2017	0. 962 8	1. 324 5	3. 72	0. 996 8	0. 394 2	12. 67
	2018	0. 994 8	0. 505 7	9. 83	0. 997 3	0. 360 6	13. 72
	2020	0. 945 4	1. 052 4	3. 11	0. 973 9	0. 733 3	4. 49

图3更直观地对比了两种方式对不同年份烟叶中总糖含量的预测结果，其中绿色“△”代表重新建模效果，红色“○”代表采用SS – PFCE方法对主模型进行更新后的预测效果，可明显看出红色“○”更加紧密地分布于拟合直线上。相比之下，SS – PFCE方法进行模型更新不仅可得到更好的模型预测性能，同时大大减少了重新建模所需的时间和计算成本，在实际应用中具有较大的价值和意义。

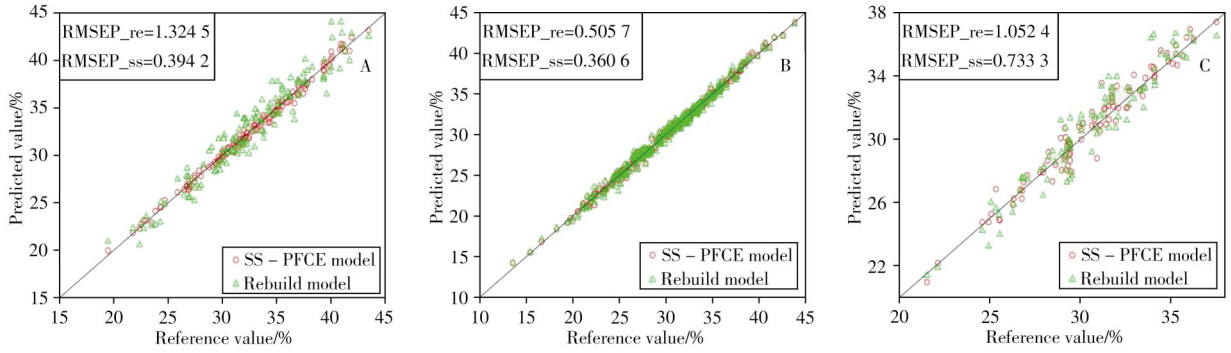

图3 重新建模和SS – PFCE对2017年(A)、2018年(B)及2020年(C)烟叶样本总糖含量的预测结果比较

Fig. 3 Comparison of prediction results for total sugar content of tobacco samples in 2017(A), 2018(B) and 2020(C) years by the rebuild model and SS – PFCE

3 结 论

针对定量模型应用于新场景导致模型性能下降的问题，本研究采用半监督无参数校正增强(SS – PFCE)的模型更新策略对3个不同年份的烟叶样本进行模型更新。结果表明SS – PFCE方法可以显著地改善主模型对新样本的预测结果，通过对回归系数的约束优化，可直接使用新样本的光谱数据进行预测。与重新建模方法相比，SS – PFCE方法可在更新过程中以更少的时间和成本达到较高的预测精度。此外，PFCE是一种基于历史数据量化的模型维护方法，不涉及模型中其他复杂参数的优化，也无需挑选具有代表性的标准样品，这种更新策略在消除外部影响时具有高效低成本的优势，对未来实际应用中多种变化场景的模型共享和模型更新均具有重要意义。

参考文献：

[1] Alam M A, Liu Y A, Dolph S, Pawliczek M, Peeters E, Palm A. *Int. J. Pharm.*, **2021**, 601: 120581.

[2] Lan Z W, Zhang Y, Sun Y, Ji D, Wang S M, Lu T L, Cao H, Meng J. *J. Pharm. Biomed. Anal.*, **2020**, 188: 113387.

[3] Huo J, Ma Y P, Lu C T, Li C G, Duan K, Li H Q. *Spectrochim. Acta Part A*, **2020**, 251: 119364.

[4] Soares F L F, Marcelo M C A, Porte L M F, Pontes O F S, Kaiser S. *Microchem. J.*, **2019**, 151: 104225.

[5] Bi Y M, Li S T, Zhang L L, Li Y S, He W M, Tie J X, Liao F, Hao X W, Tian Y N, Tang L, Wu J Z, Wang H, Xu Q Q. *Spectrochim. Acta Part A*, **2019**, 215: 398 – 404.

[6] Yuan L M, Mao F, Huang G Z, Chen X J, Wu D, Li S J, Zhou X Q, Jiang Q J, Lin D P, He R Y. *Postharvest Biol. Technol.*, **2020**, 169: 111308.

[7] Li W L, Yan X, Pan J C, Liu S Y, Xue D S, Qu H B. *Spectrochim. Acta Part A*, **2019**, 218: 271 – 280.

[8] Xue J T, Yang Q W, Li C Y, Liu X L, Niu B X. *Food Chem.*, **2021**, 342: 128386.

[9] Anderson N T, Walsh K B, Flynn J R, Walsh J P. *Postharvest Biol. Technol.*, **2021**, 171: 111358.

[10] Qin Y H, Gong H L. *Infrared Phys. Technol.*, **2016**, 77: 239 – 243.

[11] Shi Y Y, Li J Y, Chu X L. *Chin. J. Anal. Chem.*(史云颖, 李敬岩, 褚小立. 分析化学), **2019**, 47(4): 479 – 487.

[12] Mishra P, Nikzad – Langerodi R, Marini F, Roger J M, Biancolillo A, Rutledge D N, Lohumi S. *TrAC Trends Anal. Chem.*, **2021**, 143: 116331.

[13] Mou Y, Zhou L, Yu S J, Chen W Z, Zhao X, You X G. *Chemom. Intell. Lab. Syst.*, **2016**, 156: 62 – 71.

[14] Wang A D, Yang P, Chen J, Wu Z S, Jia Y F, Ma C H, Zhan X Y. *Infrared Phys. Technol.*, **2019**, 103: 103046.
[15] Feudale R N, Woody N A, Tan H W, Myles A J, Brown S D, Ferré J. *Chemom. Intell. Lab. Syst.*, **2002**, 64(2): 181 – 192.
[16] Xu B, Wu Z S, Lin Z Z, Sui C L, Shi X Y, Qiao Y J. *Anal. Chim. Acta*, **2012**, 720: 22 – 28.
[17] Farrell J A, Higgins K, Kalivas J H. *J. Pharm. Biomed. Anal.*, **2012**, 61: 114 – 121.
[18] Stork C L, Kowalski B R. *Chemom. Intell. Lab. Syst.*, **1999**, 48(2): 151 – 166.
[19] Zeaiter M, Roger J M, Bellon – Maurel V. *Chemom. Intell. Lab. Syst.*, **2006**, 80(2): 227 – 235.
[20] Guo D S, Zhu Q B, Huang M, Guo Y, Qin J W. *Comput. Electron. Arg.*, **2017**, 142: 1 – 8.
[21] Mishra P, Roger J M, Rutledge D N, Woltering E. *Postharvest Biol. Technol.*, **2020**, 170: 111326.
[22] Zhang J, Li B Y, Hu Y, Zhou L X, Wang G Z, Guo G, Zhang Q H, Lei S C, Zhang A H. *Anal. Chim. Acta*, **2021**, 1142: 169 – 178.
[23] Mishra P, Woltering E. *Anal. Chim. Acta*, **2021**, 1177: 338771.
[24] Sun W F, Zhou Z L, Li Y, Xu Z Q, Xia W S, Zhong F. *Eur. Food Res. Technol.*, **2012**, 235(4): 745 – 752.
[25] Ma H, Pan H Y, Pan D Y, Ni H F, Feng X J, Liu X S, Chen Y, Wu Y J, Luo N. *Spectrochim. Acta Part A*, **2020**, 242: 118792.
[26] Teh S L, Coggins J L, Kostick S A, Evans K M. *Postharvest Biol. Technol.*, **2020**, 166(9): 111125.

（责任编辑：龙秀芬）