潜变量机器学习方法在咖啡NIR定量分析中的应用

陈华舟[1,2]，许丽莉[3]，乔涵丽[1,2]，洪绍勇[4*]

1. 桂林理工大学理学院，广西 桂林 541004
2. 大数据处理与算法技术研究中心(桂林理工大学)，广西 桂林 541004
3. 北部湾大学海洋学院，广西 钦州 535011
4. 广州华商学院数据科学学院，广东 广州 511300

摘 要 采用近红外(NIR)光谱快检技术实现对咖啡蛋白质的定量检测，研究支持向量机(SVM)和极限学习机(ELM)等机器学习方法在建模分析中的实用性。结合潜变量分析技术，建立潜变量SVM(LV-SVM)模型和潜变量ELM(LV-ELM)模型，通过调试潜变量个数和机器学习关键参数的联合优选，实现数据降维和机器学习关键参数的同过程优化。运用定标—验证—测试机制，利用定标集样本建立咖啡蛋白质的NIR分析模型，随参数变动形成三维随动优选结构的建模预测结果，结合验证集样本对模型进行联合优选，然后将优化模型应用于测试集样本进行模型评价。LV-SVM建模优选的验证集预测均方根误差为6.797，对应的测试集预测均方根误差为8.384。LV-ELM建模优选的验证集预测均方根误差为6.118，对应的测试集预测均方根误差为7.837。与常规偏最小二乘(PLS)方法相比较，LV-SVM和LV-ELM方法均取得更好的预测结果，验证了潜变量机器学习方法在近红外定量分析中的应用优势，该方法有望应用于不同类型的咖啡各成分含量检测。

关键词 NIR光谱；咖啡；蛋白质；SVM；ELM；潜变量技术

中图分类号：O433.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)05-1441-05

引 言

随着生活质量的提高，食品的生产质量和品质安全直接关系到人们的健康，越来越多地受到人们的密切关注。咖啡是最流行的非酒精饮料之一，咖啡成分复杂，包含多种化合物，矿物质含量丰富，其中蛋白质是咖啡为人类提供能量的主要成分[1-3]。由于生长环境和加工方法的影响，不同种类的咖啡中的蛋白质含量存在一定差异，对于咖啡蛋白质含量的检测已经有比较成熟的实验室方法[4]，然而化学检测技术成本高、耗时长，需要化学试剂，容易造成污染，不能满足当今社会快节奏的生活和高质量的检测需要，同时，速溶咖啡粉末成品的制备和包装过程中不可避免地添加了一些食品添加剂，这对于蛋白质成分的提纯和检测增加了复杂度。因此，寻求一种快速检测技术来完成对咖啡蛋白质的检测具有重要的社会意义。

随着计算机和信息技术的发展，光谱快检技术广泛应用于农业、食品、生态环境、生物医学等领域[5-7]。近红外(NIR)光谱以其快速无损、无试剂、实时在线、多组分同时分析的特点得到相关行业认可[8-10]。而近红外光谱的分析过程是多变量定标校正过程，需要结合化学计量学方法的研究和应用。近些年，NIR分析技术在食品行业的应用逐渐趋于成熟，利用近红外光谱分析进行食品安全和品质检测的精度要求越来越高，如多元回归(MLR)、偏最小二乘法(PLS)等常规的线性分析方法已经不能满足建模定标需求[11-12]；大数据和智能计算技术的不断更新，涌现出一系列非线性计量学分析方法，如支持向量机(SVM)、神经网络(ANN)、极限学习机(ELM)等，用于NIR光谱建模，在定量分析方面取得良好的预测效果，能够提高模型预测精度的同时还肯定了机器学习方法在NIR分析中的可行性[13-15]。

针对速溶咖啡粉末的蛋白质快速定量检测的NIR光谱建模分析，提出利用SVM和ELM方法结合潜变量技术进

收稿日期：2020-06-23，**修订日期**：2020-10-08

基金项目：国家自然科学基金项目(61505037)，广西自然科学基金项目(2018GXNSFAA050045)，广东省普通高校青年创新人才类项目(2019KQNCX213)，广东省普通高校创新团队项目(2020WCXTD008)资助

作者简介：陈华舟，1983年生，桂林理工大学理学院教授 e-mail：hzchengut@foxmail.com

*通讯作者 e-mail：shy2002021@163.com

行建模，讨论两种方法的参数优选和潜变量提取的联合优化模式，结合简单的建模前预处理，以达到提高 NIR 光谱分析精度的目的。与常用的 PLS 方法进行对比，验证潜变量机器学习方法在近红外定量分析中的应用优势。

1 实验部分

1.1 样品采集与检测

收集 174 份咖啡粉末样品，采用常规食品蛋白质检测技术(GB/T 5009.5—2003)测定每个样品的蛋白质含量，作为 NIR 分析的参考化学值。所有样品的蛋白质百分比含量最小值为 46.55%，最大值为 73.35%，平均值为 60.00%，标准偏差值为 4.97%。使用 FOSS NIR Systems 5000 光栅型光谱仪采集咖啡粉末样本的近红外光谱，以空气作为背景，每测一个样品伴随着测量一次背景，用于光谱数据的基线校正。实验环境温度为(25±1) ℃，湿度为 45%±1%RH 的情况下，设置仪器内置光学系统对每个样品(包括背景测量)自动扫描 32 次，波长范围设置为 1 000～2 500 nm，光谱分辨率为 2 nm。光谱数据经过基线校正处理，消除光谱漂移影响，所得 174 个咖啡样本的 NIR 光谱如图 1 所示。

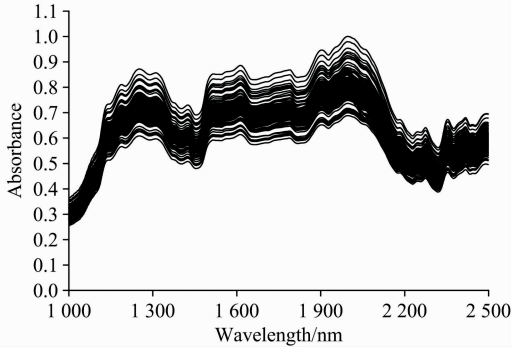


图 1 174 个咖啡粉末样本的 NIR 光谱

Fig. 1 NIR spectra of 174 coffee powder samples

1.2 潜变量机器学习方法

采用 SVM 和 ELM 两种机器学习方法，结合潜变量分析技术，对 174 个咖啡蛋白质的 NIR 光谱快速检测进行建模优化。潜变量是通过分析光谱数据的信号分布情况，提取出来的包含特定待测成分信息最大的综合变量[16]。潜变量分析常用的方法有因子分析(FA)、主成分分析(PCA)、隐马尔可夫模型(HMM)等；本工作利用 PCA 算法思想提取潜变量，并将潜变量提取过程与 SVM 和 ELM 进行联合优化，形成操作方便的新型数据优化分析模型。

(1)潜变量支持向量机(LV-SVM)模型

LV-SVM 的基本思路是采用 PCA 潜变量技术，将原光谱数据 X 通过潜变量提取形成光谱特征的潜变量特征数据 LX，进一步利用非线性映射核函数将潜变量 LX 映射到一个更高维的特征变量空间，使得原来变量之间的非线性对应关系转换成高维空间中的线性关系；加入松弛变量 ξ，在特征空间中基于线性最优化理论构建目标函数，

$$\min\left(\frac{1}{2}\|w\|^2+\gamma\sum_{j=1}^{p}\xi_j^2\right)$$

$$\text{s.t. } f=w^{\mathrm{T}}\phi(lx_j)+b+\xi_j,$$

$$lx_j\in \mathrm{LX},\ j=1,2,\cdots,p$$

其中 γ 为正则化参数，ξ_j 为松弛变量，lx_j 为潜变量矩阵 LX 的向量元素，b 为偏差因子。此为凸二次规划问题，可用 Lagrange 乘子法求解，经整理可以得到 LV-SVM 算法针对 NIR 光谱定量分析的预测模型为

$$y_i=\sum_{j=1}^{p}\alpha_j\phi(lx_j)+b_i,\quad i=1,2,\cdots,n$$

其中 y_i 为样本待测成分含量，α_j 是 Lagrange 乘子，lx_j 为潜变量变换之后的特征光谱，b_i 为基线校正偏差。

(2)潜变量极限学习机(LV-ELM)模型

ELM 算法是基于单一隐藏层的反馈式神经网络(SLFN)权值优化理论提出的一种机器学习方法，它可以为 SLFN 系统提供更优化的模型训练机制，以便更快速地确定最佳优化权值和最小训练误差，使其具有更好的泛化应用能力[17-18]。LV-ELM 的基本思想是将 PCA 提取的潜变量(LX)作为 SLFN 的输入变量，执行 ELM 算法过程，构建潜变量极限学习机模型，使得反馈式神经网络极限学习的模式完全作用于待测成分特征的光谱数据。

对任意 n 个训练样本的 SLFN 模型，$\{(lx_i,t_i)\}_{i=1}^{n}$，其中 $lx_i\in R^n$ 且 $t_i\in R^p$。具有随机 k 个隐含节点的系统输出为

$$o_i=\sum_{j=1}^{k}\beta_j g(a_j,b_j,lx_i)\quad i=1,2,\cdots,n$$

其中 $a_j\in R^n$ 和 $b_j\in R(1,2,\cdots,k)$表示第 j 个隐含节点的学习参数，$\beta_j\in R^p$ 表示隐含层的第 j 个节点到输出层的连接权值，$g(a_j,b_j,lx_i)$表示第 j 个隐含节点输出值与输入样本特征变量 lx_i 之间的关系。

具有 k 个隐含节点的 SLFN 若以零误差逼近于这 n 个样本，即 $\sum_{j=1}^{n}\|o_i-t_i\|=0$，则存在 β_j，a_j 和 b_j 满足

$$\boldsymbol{H\beta}=\boldsymbol{T}$$

其中 $\boldsymbol{H}=\{h_{ij}=g(a_j,b_j,lx_i)\}$为隐含层的输出矩阵，$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_k)$为输出权重矩阵，$\boldsymbol{T}=(t_1,t_2,\cdots,t_n)$为目标输出矩阵。于是，SLFN 系统方程转化为线性模型，则输出权重可通过最小二乘法来确定，即可以得到 $\boldsymbol{\beta}$ 的估计值为

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{-1}\boldsymbol{T}$$

其中 $\boldsymbol{H}^{-1}$ 为 $\boldsymbol{H}$ 的广义逆矩阵。利用 ELM 优化估计的值来预测样本待测成分的含量。

1.3 数据划分与模型评价指标

咖啡蛋白质定量检测的 NIR 建模采用定标—验证—测试的模式进行，将全部 174 个样本按照大约 2∶1∶1 的比例随机划分为定标集、验证集和测试集，其中定标集样本用于构建定量模型，验证样本用于对定标模型进行对比验证和参数优选，然后将优化模型应用于测试集样本进行模型评价。经过划分之后的三个样本集的统计数据如表 1 所示。

模型评价体系包括对验证集样品的评价和对预测集样品的评价，评价指标有均方根偏差(RMSE)和相关系数(r)，通

表 1　定标集、验证集和测试集样本的咖啡蛋白质含量基本统计数据

Table 1　The statistic data of coffee protein content for the calibrating, validating and testing sets

	样本个数	化学值			
		最大值	最小值	平均值	标准偏差值
定标集	84	72.80	46.55	60.431	4.977
验证集	45	73.35	49.01	59.811	4.453
测试集	45	71.37	48.58	59.400	5.476

过以下公式计算

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y'_i)^2}$$

$$r=\frac{\sum_{i=1}^{n}(y_i-y_m)(y'_i-y'_m)}{\sqrt{\sum_{i=1}^{n}(y_i-y_m)\sum_{i=1}^{n}(y'_i-y'_m)^2}}$$

其中 y_i 为样品 i 的化学检测值，y'_i 样品 i 的近红外预测值，y_m 为样品化学检测值的平均值，y'_m 为样品近红外预测值的平均值。定标过程的模型评价指标分别记为 RMSEV 和 R_V；预测过程的模型评价指标分别记为 RMSET 和 R_T。

2　结果与讨论

分别采用 LV-SVM 和 LV-ELM 两种方法对咖啡粉末的 NIR 光谱建模，定量预测蛋白质含量，有利于人们选择咖啡蛋白能量的摄取。针对 84 个定标集样本建立 LV-SVM 模型进行训练，首先基于全谱段数据提取潜变量信息，由于不同潜变量个数将影响建模效果，调试前 30 个潜变量，结合 SVM 学习过程进行联合优化，设置正则化参数的调整范围为 $\gamma=1,2,\cdots,20$，将每一个参数组合所对应的模型应用于 45 个验证集样本蛋白质含量的预测，通过比较不同潜变量个数(LV)、不同正则化参数(γ)取值，依据模型评价指标(RMSEV)确定建模优化参数。双参数调试的 LV-SVM 建模验证结果如图 2 所示，其中图 2(a)为双参数联合调试任一参数组合的预测偏差，图 2(b)和图 2(c)分别为该预测结果分别对应 r 和 LV 两个变量方向的最小预测偏差投影。依图 2 可以选择优化的 r 为 14，LV 为 15，对应 LV-SVM 模型的优化 RMSEV 为 6.797，对应的 R_V 为 0.877。

利用 LV-ELM 模型针对定标集样本进行训练，基于全谱数据提取潜变量 LX，调试潜变量数量为 1，2，…，30，结合 ELM 的学习优化过程，设置 SLFN 网络的隐含层节点数量可变，调试取值为 $k\in\{5,10,15,20,25,30,35,40,45,50\}$，通过反馈式迭代确定各个隐含节点的参数，利用最小二乘回归计算 SLFN 隐含层至输出层的权值 β，进而完成对验证集样本的蛋白质含量预测。通过比较不同潜变量个数(LV)、不同隐含层节点个数(K)的取值，依据 RMSEV 确定建模优化参数。双参数调试的 LV-ELM 建模验证结果如图 3 所示，其中图 3(a)为双参数联合调试任一参数组合的预测偏差，图 3(b)和图 3(c)分别为该预测结果分别对应 K 和 LV 两个变量方向的最小预测偏差投影。依图 3 可以选择优化的 K 为 40，LV 为 18，对应 LV-ELM 模型的优化 RMSEV 为 6.118，对应的 R_V 为 0.908。

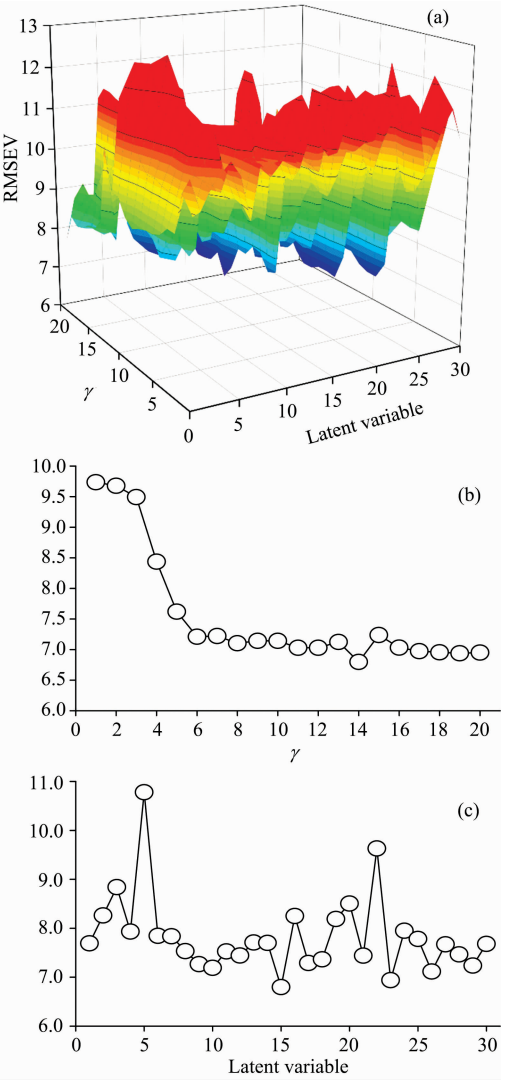
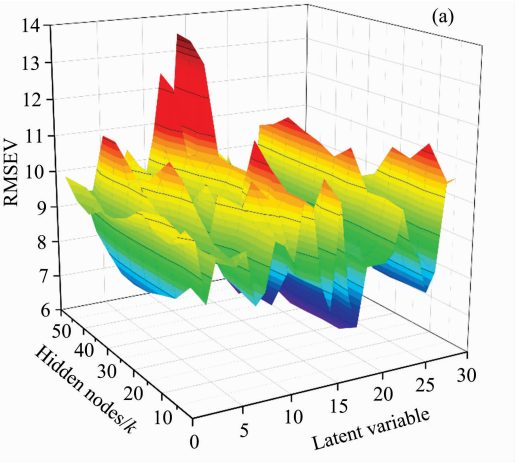


图 2　LV-SVM 定标验证模型的 RMSEV 优选

Fig. 2　The optimization of RMSEV for the LV-SVM calibration models



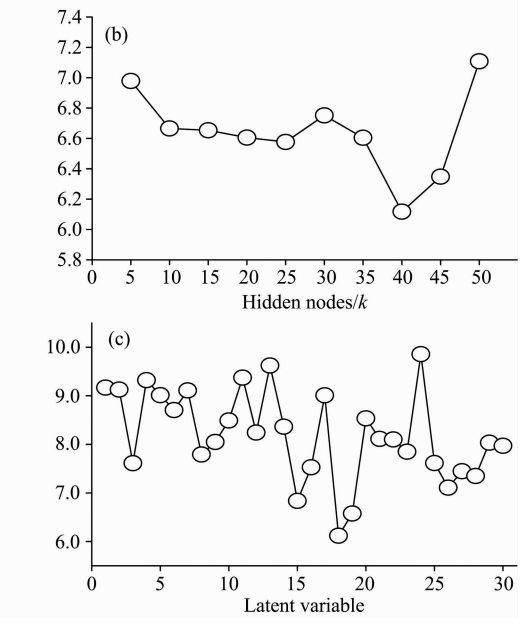


图 3 LV-ELM 定标验证模型的 RMSEV 优选

Fig. 3 The optimization of RMSEV for the LV-ELM calibration models

利用上述定标和验证过程得到的最优建模参数，即 15 个潜变量、正则化参数为 14 的 LV-SVM 模型和 18 潜变量、40 个隐含层节点的 LV-ELM 模型，分别对测试集的 45 个咖啡样本的蛋白质含量进行预测，计算对应的 RMSET 和 R_T，所得结果列于表 2 中；同时将常规 PLS 定标的优化模型预测结果也列于表中进行比较。对比可知，LV-SVM 和 LV-ELM 方法在咖啡蛋白的NIR光谱快速定量分析中能够取得比常规 PLS 方法更优的预测精度，且 LV-ELM 模型取得相对于 LV-SVM 模型更好的预测结果。经过定标—验证—测试结果可知，潜变量提取结合机器学习的方法在近红外定量分析中具有一定的应用优势，比常规的线性建模方法更有应用前景。

表 2 LV-SVM，LV-ELM 和 PLS 方法对咖啡蛋白质的 NIR 建模预测结果

Table 2 The NIR model prediction results for coffee protein based on the LV-SVM，LV-ELM and PLS methods

	潜变量个数	参数优选	RMSEV	R_V	RMSET	R_T
LV-SVM	15	$\gamma=14$	6.797	0.877	8.384	0.858
LV-ELM	18	$k=40$	6.118	0.908	7.837	0.861
PLS	24	—	8.719	0.842	9.037	0.829

3 结 论

采用 NIR 光谱快速检测技术实现对速溶咖啡样本中蛋白质含量的定量检测，在建模方法上采用潜变量结合机器学习的联合优化方法，建立 LV-SVM 和 LV-ELM 定标预测模型，形成 SVM 或 ELM 关键参数和潜变量优选的双参数联合调试模式，使建模预测偏差结果形成三维随动优选结构。该方法能够在实现变量降维的同时优选建模参数，对咖啡蛋白质的定量分析取得良好的预测效果，经过定标—验证—测试三个环节的建模对比，该方法普遍优于常规 PLS 的建模预测。结果表明，潜变量结合机器学习联合参数优化方法能够为 NIR 快速检测技术提供良好的建模分析手段，有望推广应用于其他类型的咖啡样本进行快速品质鉴定。

References

[1] Janissen B, Huynh T. Resources, Conservation and Recycling, 2018, 128: 110.
[2] YANG Kai-zhou, ZHAI Xiao-na, DU Bing-jian, et al(杨凯舟，翟晓娜，杜秉健，等). Food Science(食品科学), 2014, 35(3): 243.
[3] Waters D M, Arendt E K, Moroni, A V. Critical Reviews in Food Science and Nutrition, 2017, 57(2): 259.
[4] CHEN Lei, AN Miao, YAN Hui-ying, et al(陈 雷，安 苗，闫会莹,等). Journal of Jilin Normal University·Natural Science Edition (吉林师范大学学报·自然科学版), 2017, 38(3): 79.
[5] HE Yong, PENG Ji-yu, LIU Fei, et al(何 勇，彭继宇，刘 飞,等). Transactioins of the Chinese Society of Agricultural Engineering (农业工程学报), 2015, 31(3): 174.
[6] Sakudo A. Clinica Chimica Acta, 2016, 455: 181.
[7] Jamshidi B, Mohajerani E, Jamshidi J. Measurement, 2016, 89: 1.
[8] Prieto N, Juarez M, Larsen I L, et al. Meat Science, 2015, 110: 76.
[9] LIANG Man, HUANG Fu-rong, HE Xue-jia, et al(梁 曼，黄富荣，何学佳，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2014, 34(8): 2132.
[10] CHEN Wan-chao, TAO Xin, FAN Chang-chun, et al(陈万超，陶 鑫，范长春，等). Chinese Journal of Analytical Chemistry(分析化学), 2019, 47(2): 315.
[11] Liu J, Chen N, Yang J, et al. Food Chemistry, 2018, 253: 284.
[12] Chen H, Xu L, Jia Z, et al. Analytical Letters, 2018, 51: 1564.
[13] Liu T, Li Z, Yu C, et al. Infrared Physics & Technology, 2017, 87: 124.
[14] Chen H, Liu X, Jia Z, et al. Chemometrics and Intelligent Laboratory Systems, 2018, 182: 101.
[15] BIN Jun, FAN Wei, ZHOU Ji-heng, et al(宾 俊，范 伟，周冀衡，等). Tobacco Science & Technology(烟草科技), 2016, 49(9): 50.

[16] Shao X, Du G, Jing M, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 114: 44.
[17] Henriquez P A, Ruz G A. Engineering Applications of Artificial Intelligence, 2019, 79: 13.
[18] Jin Y, Li J, Lang C Y,et al. Multidimensional Systems and Signal Processing, 2017, 28(3): 905.

Latent Variable Machine Learning Methods Applied for NIR Quantitative Analysis of Coffee

CHEN Hua-zhou[1,2], XU Li-li[3], QIAO Han-li[1,2], HONG Shao-yong[4*]
1. College of Science, Guilin University of Technology, Guilin 541004, China
2. Center for Data Analysis and Algorithm Technology, Guilin University of Technology, Guilin 541004, China
3. College of Marine Sciences, Beibu Gulf University, Qinzhou 535011, China
4. School of Data Science, Guangzhou Huashang College, Guangzhou 511300, China

Abstract Near-infrared (NIR) spectroscopy rapid detection technology was used to determine protein content in instant coffee. Support vector machine (SVM) and extreme learning machine (ELM) was applied for validating their practicality in modeling analysis. We proposed the latent variable SVM (LV-SVM) and latent variable ELM (LV-ELM) methods combined with latent variable analysis technique. Thetuning of latent variables and the optimization of the key parameters in machines were joint in-one so that the data dimension reduction and the selection of machine parameters can be both accomplished in one single modeling process. The calibrating-validating-testing mechanism was used for sample division. The NIR analytical models were trained based on the calibrating sample set. The model prediction results were generated and saved as a 3D box as they were determined by the simultaneous tuning of the latent variable and the machine parameter. Then the joint optimization of model parameters was selected in the way of predicting the validating samples. Further, the optimal model was evaluated by the testing samples. The optimal LV-SVM model gave the validating root mean square error as 6.797; the corresponding testing root mean square error as 8.384. The optimal LV-ELM model obtained the validating root mean square error as 6.118. The corresponding testing root means square error as 7.837. Compared with the common partial least square method, the LV-SVM and LV-ELM methods have better prediction results, which verified the application advantages of the latent variable machine learning method in near-infrared quantitative analysis. This proposed method is expected for further application in content detection of other kinds of coffee.

Keywords NIR spectroscopy; Coffee; Protein; SVM; ELM; Latent variable technique

(Received Jun. 23, 2020; accepted Oct. 8, 2020)

* Corresponding author