近红外光谱结合共识模型快速检测果酒的总酚含量

叶 华[1,3]，袁雷明[2*]，张海宁[3,4]，李理敏[2]

1. 淮阴工学院生命科学与食品工程学院，江苏 淮安 223001
2. 温州大学电气与电子工程学院，浙江 温州 325035
3. 江苏大学食品与生物工程学院，江苏 镇江 212013
4. 洛阳师范学院食品与药品学院，河南 洛阳 471934

摘 要 果酒发酵中的多酚是引起果酒口感、颜色变化的重要因素。为保证果酒品质，有必要开发一种快速监测发酵过程中多酚含量变化的技术。收集不同批次成熟期的蓝莓、桑葚为原料，分别碾压成汁，同时按比例混合二者，于小型发酵罐进行发酵。通过离线收集不同发酵时段的发酵液于离心管，高速离心后取上清液置于棕色瓶保存，共计得到48个果酒发酵样本。将上清液置于三个平行样比色皿，以傅里叶快速变换近红外光谱仪(FT-NIR)采集其透射光谱，取平均值作为该样本的光谱信号。然后将棕色瓶内的发酵液以国标法(即以标准液的吸光度值制定标准曲线)测定各样品的总酚含量，以duplex法计算样本光谱之间的距离且按2∶1的比例划分为训练集和预测集。采用间隔偏最小二乘法(iPLS)将训练集样本的透射光谱与总酚含量之间构建定量模型，间隔数从2依次变化到60个。该研究创新之处是使用共识方法融合多个已构建好的iPLS成员模型，按一定的共识规则分配权系数。通过各成员模型交互验证的残差及其残差之间的相关性来优化各成员模型的线性组合，以拉格朗日乘数法求解各成员模型的权系数，使间隔偏最小二乘-共识模型(consensual iPLS，C_iPLS)的交互验证均方根误差最小。相比于全局PLS模型、划分不同间隔数量时的iPLS模型，C_iPLS均具有较小的预测误差。当划分39个间隔时由三个iPLS成员模型(即14^{th}，16^{th}，18^{th})组成的共识模型误差最小为124.2，交互验证相关系数为0.944，对预测集样本的预测均方根误差为163.4，预测相关系数为0.931，预测性能均优于PLS和iPLS模型。另外，作为对比选用连续投影算法与无信息变量剔除法来优化光谱模型，其预测性能均不及本文提出的共识模型。分析各iPLS模型预测残差之间的相关性，发现共识模型主要是融合那些具有较高预测性能且模型间较低相关性的成员模型。结果表明，光谱分析结合共识方法可提高回归模型的预测精度、减少建模所需变量数，能够用于果酒总酚含量的离线快速检测。

关键词 近红外光谱；间隔偏最小二乘法；共识模型；总酚含量

中图分类号：O675 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)03-0777-05

引 言

果酒是以成熟水果为原料发酵酿造而成的一种酒类，纯度较低，营养丰富，有着独特的风味口感，是各类酒中最具发展潜力的一种酒类[1]。蓝莓、桑葚等果品富含花青素、多酚、有机酸、酚酸、氨基酸等多种具有保健功能的营养物质[2]，尤其是多酚类含量丰富并具有多种生理功能，深受饮酒人士的喜爱。多酚通常是果酒品质的重要指标之一，检测方法包括高效液相色谱法和分光光度计法，但费时费力，因此在果酒发酵过程中急需一种快速检测多酚含量的分析技术。

近红外光谱(near-infrared spectroscopy，NIRS)作为一种近30年来快速兴起的分析技术，其快速、简单、低成本、可重复性高等优点在食品安全与营养[3-4]、药物组分及活性[5]、环境污染[6]、石油化工[7]等方面得到了广泛的应用。采用可见-近红外光谱测量苹果[8]、柑橘[9]、葡萄[10]等水果的内部品质，预测相关性一般高于0.8。常用化学计量学建

收稿日期：2019-01-09，**修订日期**：2019-03-27

基金项目：淮安市重点研发现代农业计划项目(HAN201627)，国家重点研发专项计划项目(22017YFD0401300)，国家自然科学基金项目(61705168)，温州市公益计划项目(S20170003，G20180009)资助

作者简介：叶 华，1978年生，江苏大学博士研究生，淮阴工学院讲师 e-mail：hgyehua@126.com

*通讯联系人 e-mail：yuan@wzu.edu.cn

模方法为偏最小二乘法(partial least square，PLS)，它是一种矩阵投影建模技术，通常是采用全波段的光谱构建回归模型。但全波段的光谱数量比较大，一般在几百甚至几千以上，且获取的光谱信息易受如水吸收峰[11]、仪器性能等因素干扰，因此变量筛选就显得尤为重要[8, 12-13]。

通常，待测物的理化指标与特定光谱区域的信号存在较高相关度，一些用于光谱变量的筛选方法被提出来，如间隔偏最小二乘法(interval PLS，iPLS)、联合区间法、遗传算法、连续投影算法等。但这些方法仅是筛选其中的一个或几个区间或多个变量来构建单个 PLS 模型，而忽略了其他光谱区间对构建模型的贡献，使信息损失。为此，一些多模型的共识方法被提出[14-16]，如基于不同样本子集建模的均值共识策略[14]，或基于不同模型分配权重的共识策略[15-16]来提高光谱预测精度，本文则提出一种基于光谱区间模型来分配各成员模型 iPLS 权重的共识策略。

1 实验部分

1.1 样本准备

采用江苏省溧水县生产的兔眼蓝莓、桑葚(品种为镇椹 1 号、台湾 1 号、射阳 3 号)，于成熟期收集各批次于果园的不同位置。按照蓝莓、桑葚果酒的酿制工艺[17-18]，将每批次进行单独打浆、加酶、灭酶等操作获取蓝莓汁、桑葚汁；按比例混合不同批次蓝莓汁、桑葚汁，共计得到 16 份果汁液，分别装入各小型发酵罐进行发酵。

以离线采样法分别收集处于发酵中后期的发酵液三次，将其 1 000 r·min^{-1}离心 10 min 后，取上清液置于棕色瓶，共计 48 个样，用于后续的光谱采集及多酚含量测量。

1.2 近红外光谱的数据采集

针对每个发酵液样本，装载于 3 个 10 mm 光程的比色皿，采用 Antaris Ⅱ傅里叶变换近红外(FT-NIR)光谱仪(美国 Thermo Fisher 公司)收集透射光谱，扫描范围为 10 000～4 000 cm^{-1}。以仪器内置背景为参比，扫描次数 16 次，分辨率为 8 cm^{-1}，共计 1 577 个波数点。图 1 为各测试样的平均透射光谱曲线。

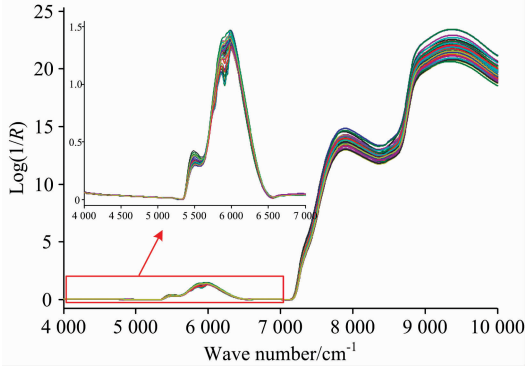

图 1 果酒的平均光谱透射曲线

Fig. 1 The averaged transmission spectra of fruit-wine

1.3 总酚含量的国标检测

根据国家标准 GB/T8313—2018 总酚含量的测定方法，以分光光度计法在波长 765 nm 处测量不同浓度没食子酸标准溶液 Y 的吸光度 A，绘制标准曲线($A=0.489\ 4Y+0.012\ 5$，$R^2=0.996\ 9$)，再对各发酵液标准处理得到的吸光度值计算出总酚浓度。

采用 duplex 算法以逐个拾取距离其余样本的光谱中心距离最远的那个样本依次按照 2∶1 的比例划分为校正集和预测集(如表 1 所示)，其中校正集用于构建、训练模型，预测集作为外来样本用于检验模型的预测性能。

表 1 果酒样本集的总酚含量的统计结果

Table 1 Statistic results of polyphenol content in fruit-wine

	Number	Range	Mean	SD	CV
Calibration	32	196～1 753	622.7	378.3	0.608
Prediction	16	265～1 762	679.2	430.2	0.633

Note：SD：standard deviation；CV：coefficient of variation

1.4 间隔偏最小二乘-共识模型的构建方法

间隔偏最小二乘模型是基于某一间隔区域的光谱信息建立的偏最小二乘模型。一般是将全区间光谱分割为 n 个等长度间隔，基于这些间隔内的光谱信息，分别构建偏最小二乘模型，并从 n 个 iPLS 模型中选择一个预测性能最好的作为最终模型。

共识模型 $f(x)$(consensus modeling)，并不是一种基于特征空间映射的建模方法，而是融合多个成员模型、按一定的共识规则分配权系数 c_i，从而达成共识。具体步骤为：(1)成员模型 $f_i(x_i)$是基于训练集的各特征集$\{x_1, \cdots, x_n\}$分别构建而成；(2)根据权系数 c_i 加权各成员模型[如式(1)]，同时计算各成员模型预测值 $\hat{y}_i$ 的残差向量 e_i 的标准差 δ_i 及各残差向量间的相关性 r_{ij}[如式(2)]；(3)为使残差均值(mean squared errors，MSE)最小，设定约束条件，通过拉格朗日乘数法求解式(3)，寻找最优解 c_i[17]。

$$F(x)=\sum_{i=1}^{n}c_i f_i(x_i) \tag{1}$$

$$\begin{cases}\sigma_i=\sqrt{\dfrac{1}{n}\|\hat{y}_i-y\|^2}\\ r_{ij}=\dfrac{\langle e_i, e_j\rangle}{n\sigma_i\sigma_j},\ i,j=0,1,\cdots,n\end{cases} \tag{2}$$

$$\begin{cases}\mathrm{MSE}=\mathrm{ARGmin}\langle\sum\limits_{i=1}^{n}(y_0-\sum\limits_{i}c_i f_i(x_i))^2/n\rangle\\ \quad=\sum\limits_{i=1}^{n}c_i^2\sigma_i^2+2\sum\limits_{i=1}^{n}\sum\limits_{j>i}^{n}c_i c_j r_{ij}\sigma_i\sigma_j\\ f_i(x)=\{f_1(x_1),\cdots。f_i(x_i),\cdots,f_n(x_n)\}\\ \text{s.t.}\begin{cases}0\leqslant c_i\leqslant 1\\ \sum c_i=1\end{cases}\end{cases} \tag{3}$$

图 2 为间隔偏最小二乘-共识模型(C_iPLS)构建流程：首先将训练集的光谱数据划分 n 个间隔，以每个间隔的光谱变量作为特征集，分别构建 n 个 iPLS 模型，作为成员模型 f_i 与权系数 c_i 向量相乘构建共识模型 $F(x)$。对式(3)的 MSE 进行开方计算，即为共识模型的交互验证均方根误差

(root MSE of cross validation，RMSECV)，此参数作为共识模型的主要评价指标。调整成员模型的间隔数量 n，根据 RMSECV 最小原则来选取最佳共识模型。

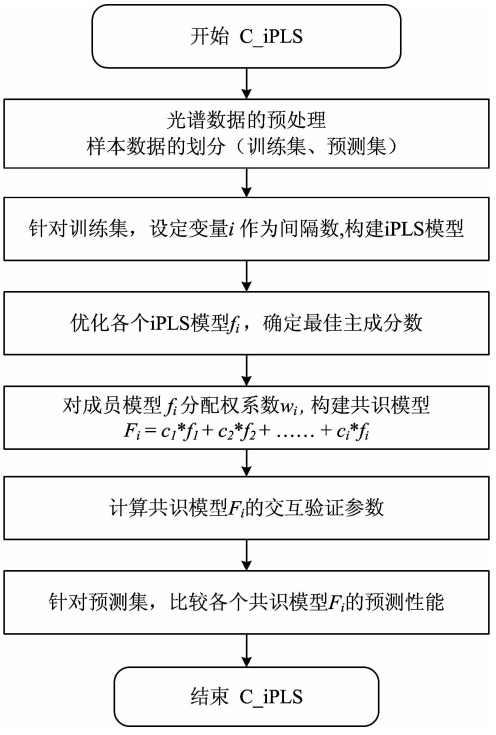

图 2　间隔偏最小二乘-共识模型的工作流程

Fig. 2　Workflow of the consensus modeling iPLS

2　结果与讨论

2.1　回归模型的构建与优化

基于校正集建立基于全光谱区间的 PLS 模型，以内部交叉验证避免所建模型的过拟合或欠拟合[9]。采用多元散射校正(MSC)、标准正态变换(SNV)等方式对果酒透射光谱数据进行预处理，但构建的 PLS 模型相比于无预处理时的预测性能，并未得到改善。因此，本工作所处理的光谱数据均为原平均光谱。

将全区间光谱划分的间隔数量从 2 依次增加到为 60；并基于这些间隔的光谱变量分别构建PLS模型，以RMSECV最小为原则筛选最佳的 iPLS 模型。作为对比，连续投影算法 (SPA)、无信息变量剔除(UVE)等变量筛选方法用于校正集的模型构建。

图 3 为不同间隔数 n 时的最佳 iPLS 及 C_iPLS 模型的交互验证均方根误差，随着 n 变大，RMSE 值均呈现先变大后逐渐变小，当 $n>12$ 时，误差在 125～155 内波动。当 $n=1$，此时 iPLS 模型、PLS 模型、C_iPLS 模型等价。当间隔数为 3 时，两者模型性能最差，RMSE 均达到 185 以上；当间隔数为 39 时，此时的 iPLS 模型、共识模型同时具有最小的 RMSECV，分别为 131.3 和 124.2，为全局($n\leqslant 60$)最低值。

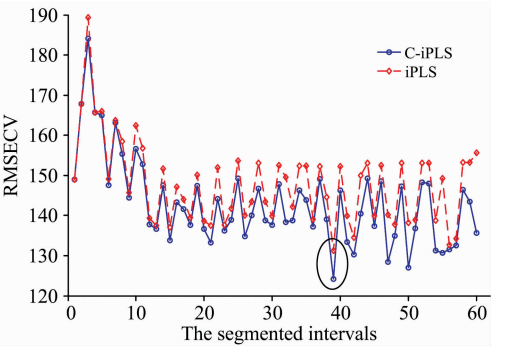

图 3　间隔偏最小二乘模型及其共识模型的交互均方根误差变化趋势

Fig. 3　RMSECV tendencies of the developed iPLS and C_iPLS models

2.2　模型预测及比较

表 2 为几种模型的预测结果。当 $n=39$ 时，以第 14 个间隔中的 40 个光谱变量构建的 iPLS 模型预测性能优于其他 iPLS 模型，但相比于全局 PLS 模型、共识模型，其预测效果最差，预测均方根误差(root mean squared errors of prediction，RMSEP)低至 178.7。结合以划分 39 个间隔时的 14^{th}，16^{th}，18^{th} 三个 iPLS 为成员模型构建的共识模型(如图 4 所示)，具有最好的预测性能，RMSEP 低至 163.4，优于 iPLS 和 PLS 模型，此时 C_iPLS 模型对训练集、预测集的样本预测分布如图 5 所示，样本分布越接近 45°对角线表示模型预测性能越好。与传统的 SPA 和 UVE 变量筛选后的模型对比，共识模型的预测性能具有一定的优越性。

表 2　几种不同回归模型的预测比较

Table 2　Comparisons of different regression models for prediction

Methods	Variables	Intervals	PCs	Calibration set			Prediction set		
				RMSECV	R_{cv}	Bias	RMSEP	R_p	Bias
PLS	1557	1	5	187.9	0.869	−5.92	169.6	0.939	−44.05
iPLS[a]	40	14^{th}	4	131.3	0.936	0.25	178.7	0.906	−20.77
C_iPLS	120	14^{th}, 16^{th}, 18^{th}	—	124.2	0.944	−1.68	163.4	0.931	−23.53
SPA-MLR	12	—	—	158.6	0.905	−35.7	225.1	0.841	−26.6
UVE-PLS	68	1	2	171.7	0.888	−9.73	173.2	0.945	−55.01

Note：a：the full spectra were segmented into 39 intervals

2.3 间隔偏最小二乘共识模型的分析

分析共识模型中各成员模型的权系数，仅入选了三个iPLS模型，共识模型的表达式为 $F(x)=0.783f_{14}+0.092\ 7f_{16}+0.124\ 3f_{18}$。成员模型的权系数绝对值越大，表示该成员模型在共识模型占有的信息就越多。从构成来看，共识模型舍弃了绝大部分的iPLS模型，这是由于其他iPLS模型具有较大的预测误差才设置其权系数为0；从图4及权系数来看，共识模型不仅舍弃了部分具有较低RMSECV的iPLS模型，还分配了部分具有较差预测能力iPLS模型的权系数，如舍弃了 11^{th}～13^{th}，29^{th}等iPLS模型，以及对 18^{th} iPLS成员模型分配的权系数大于 16^{th} iPLS成员模型。分析式(3)，可以发现，除了要求各成员模型的残差平方和 $\sum_{i=1}^{n}c_i^2\sigma_i^2$ 尽可能小，并要求各成员模型之间的残差向量的相关性 r_{ij} 尽可能地低，这样才能使MSE值最小。

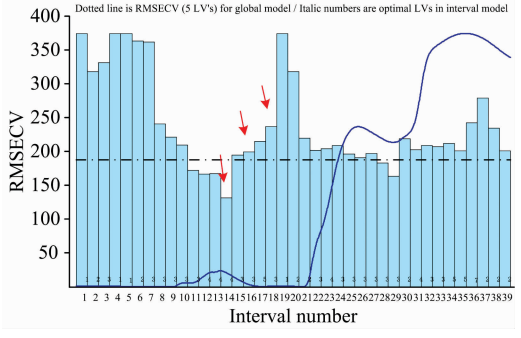

图4 间隔偏最小二乘模型的交互验证均方根误差柱形图

Fig. 4 Bar plot of RMSECV by the iPLS models

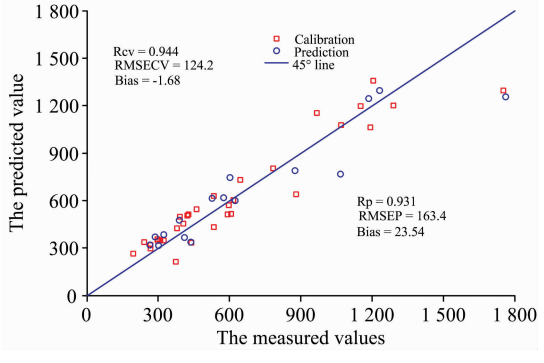

图5 间隔偏最小二乘的共识模型的预测散点图

Fig. 5 Scatter of prediction versus measurements for C_iPLS model

统计39个iPLS模型的交互验证残差向量(e_i，e_j)之间的相关性 r_{ij}，如图6所示，区域越白，表示相关性越高；反则表示相关性越低。观察间隔 11^{th}～13^{th}的iPLS模型，与其他间隔的iPLS存在较高的相关性；而部分黑色区域如间隔 1^{th}～7^{th}的iPLS模型之间虽然存在较低的相关性，但是这些模型的RMSECV值较大，易使公式3中的MSE数值变大。作为一个好的共识策略，应该筛选那些具有较低RMSECV的成员模型，以及这些成员模型残差之间较低的相关性，从而提升了共识模型的预测精度、稳定性。

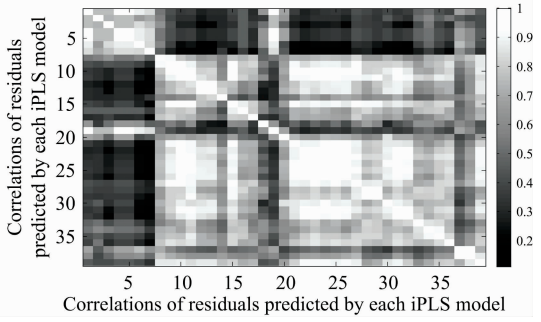

图6 间隔偏最小二乘模型的残差向量间的相关性

Fig. 6 Correlation mappings between the residual vectors of iPLS models

3 结 论

探讨傅里叶变换近红外光谱快速离线测量发酵过程中的果酒多酚含量。通过间隔偏最小二乘法挖掘不同光谱区间内的模型信息，并以iPLS模型作为成员模型，融合各成员模型之间的冗余信息，使共识模型同时考虑了各特征集的模型信息、各个成员模型的误差以及误差之间的相关性，使共识模型的预测结果更加稳定、可靠，有助于近红外光谱的应用推广。

References

[1] Kelly N P, Kelly A L, O'Mahony J A. Trends in Food Science & Technology, 2019, 83: 248.

[2] Sun X, Yan Z H, Zhu T, et al. Food Chemistry, 2019, 279: 63.

[3] Alishahi A, Farahmand H, Prieto N, et al. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 2010, 75(1): 1.

[4] Grassi S, Alamprese C. Current Opinion in Food Science, 2018, 22: 17.

[5] Mazivil A, Sarmento J, Olivieri A C. Trac Trends in Analytical Chemistry, 2018, 98: 50.

[6] Horta A, Malone B, Stockmann U, et al. Geoderma, 2016, 241-242: 180.

[7] Roman M, Balabina R Z, Safievab E I, et al. Analytica Chimica Acta, 2010, 671(1): 27.

[8] Yuan L M, Cai J R, Sun L, et al. Food Analytical Methods, 2016, 9(3): 785.

[9] Yuan L M, Sun L, Cai J R, et al. Journal of Food Process Engineering, 2015, 38(4): 309.

[10] YUAN Lei-ming, CAI Jian-rong, SUN Li, et al(袁雷明，蔡健荣，孙 力，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(4): 1220.

[11] Kumar G A, Riman R E, Brennan J G. Coordination Chemistry Reviews, 2014, 273-274: 111.

[12] Zou X B, Zhao J W, Povey J W, et al. Analytica Chimica Acta, 2010, 667(1-2): 14.

[13] Yun Y H, Li H D, Deng B C, et al. TRAC Trends in Analytical Chemistry, 2019, 113: 102.
[14] LI Yan-kun, SHAO Xue-guang, CAI Wen-sheng(李艳坤, 邵学广, 蔡文生). Chemical Journal of Chinese Universities(高等学校化学学报), 2007, (2): 246.
[15] Ji G L, Huang G Z, Chen X J, et al. Chemometrics and Intelligent Laboratory Systems, 2015, 144: 56.
[16] GUO Zhen-zhu, CHEN Xiao-jing, YUAN Lei-ming, et al(郭珍珠, 陈孝敬, 袁雷明, 等). Acta Photonica Sinica(光子学报), 2018, 47(8): 106.
[17] ZHANG Hai-ning, LI Xi, MA Yong-kun(张海宁, 李 希, 马永昆). Food Science(食品科学), 2017, 38(12): 190.
[18] WANG Hang, ZHANG Hai-ning, MA Yong-kun, et al(王 行, 张海宁, 马永昆, 等). Modern Food Science and Technology(现代食品科技), 2015, 31(1): 90.

Rapid Measurement of the Polyphenol Content in Fruit-Wine by Near Infrared Spectroscopy Combined with Consensus Modeling Approach

YE Hua[1,3], YUAN Lei-ming[2*], ZHANG Hai-ning[3,4], LI Li-min[2]
1. Department of Life Science & Food Engineering, Huaiyin Institute of Technology, Huaiyin 223001, China
2. College of Mathematics, Physics & Electronic Engineering Information, Wenzhou University, Wenzhou 325035, China
3. School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China
4. College of Food and Drug, Luoyang Normal University, Luoyang 471934, China

Abstract Polyphenol is one of important factors that cause the changes of taste and color in fruit-wine. To ensure the quality of fruit-wine, it is necessary to develop a fast measurement that monitors the change of polyphenol content during the fermentation. The ripe blueberry and mulberry were collected from different harvest batches. They were crushed respectively into juice, and their mixed juice was also mixed in certain ratio for fermentation in the small fermentation tanks. Those fermenting liquors from the different fermenting periods were collected through the off-line sampling access. The supernate was obtained by centrifugation pretreatment and totally 48 fermenting samples were preserved in the brown bottles for later use. The supernate were injected into three paralleled cuvettes, whose transmission spectrums were scanned by FT-NIR spectrometer, and their repeated readings were averaged for the spectral signals. Then, the total phenol content was measured by the national standard method (i.e. the standard curve was established between the absorbance value and the standard solution), and all samples were divided into the calibration and prediction set in a ratio of 2∶1 by duplex algorithm, which was used to calculate the spectral distance from the divided sample to the center of the rest samples. Interval partial least square (iPLS) was used to construct series of quantitative models between the transmission spectra and the total phenol contents in the training set, and the number of intervals was successively changed from 2 to 60. The innovate point in this work was that the consensual rule was used to integrate the calibrated member models (here referring to the iPLS model) into a consensus model and distribute the weighting coefficients. The linear combinations of member models were optimized to minimize the mean squared error (MSE) in the consensus model through the residual errors from the cross validation and their correlations. The weighted coefficient of each member model was solved by Lagrange multiplier method, so as to minimize the root mean square error of the consensus model. Compared with the global model of partial least squares (PLS), interval partial least-squares (iPLS) model with different number of spectral intervals, the consensual iPLS (C_iPLS) model commonly obtained a better performance. When the full spectra were divided into 39 intervals, the C_iPLS model, composed of three iPLS members models (those were 14^{th}, 16^{th}, 18^{th} iPLS model respectively), got the minimum root mean squared error of cross validation (RMSECV) of 124.2, as well as the correlation coefficient of cross validation (R_{cv}) of 0.944, and the samples in prediction set were tested well with root mean square of prediction (RMSEP) of 163.4, as well as the correlation coefficient (R_p) of prediction of 0.931. In addition, the successive projection algorithm and the uninformative variable elimination were used to optimize the spectral model, but the predictive performances were not better than the proposed consensus model. By analyzing the correlation between the predicted residuals of each iPLS model, it was found that the consensus model commonly screened these member models featured with high prediction performance and low correlation between member models. Results showed that the spectral analysis technology combined with the consensus method could improve the prediction accuracy of the regression model, reduce the modeling number of variables, and could be employed off-line for the rapid detection of total phenol content in fruit wine.

Keywords Near infrared spectroscopy; Interval partial least square; Consensus model; Total phenol content

* Corresponding author
(Received Jan. 9, 2019; accepted Mar. 27, 2019)