# 赤霞珠酿酒葡萄总酚含量的近红外光谱定量分析

罗一甲[1]，祝　赫[1]，李潇涵[1]，董　娟[1]，田　昊[1]，史学伟[1]，王文霞[2]，孙静涛[1*]

1. 石河子大学食品学院，新疆 石河子　832003
2. 石河子大学机械电气工程学院，新疆 石河子　832003

**摘　要**　酿酒葡萄中的总酚含量是影响葡萄品质的重要指标，也是影响葡萄酒质量的关键因素。为了快速准确地检测赤霞珠葡萄的总酚含量，利用近红外光谱技术结合GA-ELM预测模型对赤霞珠葡萄总酚含量进行预测研究。试验采用5个收获期(每期采集40串，每串取10个)的赤霞珠葡萄，采集200组葡萄的12 500～4 000 $cm^{-1}$波段范围内的近红外光谱。基于福林酚比色法原理对赤霞珠葡萄的总酚含量进行测定，使用SPXY算法将样品按照3∶1比例分为校正集和预测集，共计150个校正集和50个预测集。分别采用多元散射(MSC)、标准正态变换(SNV)、数据中心化(MC)、移动窗口平滑(MA)和一阶导数＋SG方法对原始光谱进行预处理，优选出最佳的预处理方法为MSC。并进一步采用竞争性自适应重加权算法(CARS)、遗传算法(GA)、联合区间偏最小二乘算法(si-PLS)和连续投影算法(SPA)分别对光谱波段进行提取，经对比分析发现CARS提取的69个特征波长数据能有效提高模型的稳定性和预测结果。在MSC预处理和特征波长提取的基础上，引入极限学习机(ELM)算法，建立赤霞珠葡萄总酚含量的预测模型，在总酚含量预测过程中，采用遗传算法(GA)对ELM模型进行优化，并探究了不同的激活函数和隐含层神经元个数对GA-ELM模型预测能力的影响，确定最优的激活函数为Sigmoidal，最优的神经元个数为50个。最后，将ELM和GA-ELM模型的预测能力进行对比，结果显示GA-ELM模型的预测能力高于ELM模型的预测能力，其中MSC＋CARS＋GA-ELM模型预测能力最好，校正相关系数($R_c$)为0.901 7，预测相关系数($R_p$)为0.901 3，校正均方根误差(RMSEC)为2.112 4，预测均方根误差(RMSEP)为1.686 8，剩余预测偏差(RPD)为2.308 0。研究结果表明：利用近红外光谱技术结合变量优选建立的GA-ELM模型可实现对赤霞珠葡萄的总酚含量的预测，为赤霞珠葡萄品质的检测奠定了理论基础。

**关键词**　变量优选；赤霞珠葡萄；总酚；极限学习机；近红外光谱

**中图分类号**：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)07-2036-07

## 引　言

葡萄酒的品质与葡萄的质量密切相关。葡萄中的总酚含量决定了果实中酚成熟状态，影响着葡萄酒的颜色、味道和口感，是生产优质红葡萄酒的一个关键标准。目前，葡萄总酚含量的测定常用福林酚法，检测过程繁琐费时，大样本操作困难，检测后的样品组织遭到破坏，商品价值丧失。因此，寻求一种高效、快速的检测葡萄总酚含量的方法十分必要。

近红外光谱技术是一种新型分析检测技术，具有快速、无损、便捷等优势。在水果品质检测方面，近红外光谱技术应用广泛，其中在葡萄品质无损检测方面，国内外学者进行了一定的研究[1-3]。有时近红外光谱包含的数据量大、信息冗杂，需采用合适的波长提取方法使数据信息更具特征性。许峰等[4]利用近红外光谱技术结合CARS算法对红提葡萄的糖度和酸度进行预测，建立随机森林预测模型，预测相关系数分别达到0.956 8和0.940 5。Michael等[5]利用近红外光谱技术对不同年份、不同收获期的赤霞珠和西拉葡萄的可溶性固形物(SSC)、pH、可滴定酸、花青素和总酚含量进行预测，采用不同的预处理方法结合递归特征消除(REF)算法建立PLS预测模型，但模型的总酚含量预测结果较低。章林忠等[6]利用近红外光谱技术对不同品种的鲜食葡萄的总酚、总糖、果糖、蔗糖和SSC进行预测，建立PLS预测模型，但部分葡萄总酚含量的预测结果出现过拟合现象。当前，国内

**收稿日期**：2020-07-02，**修订日期**：2020-11-19

**基金项目**：国家科技支撑项目(2015BAD19B03)，石河子大学高层次人才科研项目(RCSX2018B04)资助

**作者简介**：罗一甲，女，1994年生，石河子大学食品学院硕士研究生　　e-mail：luoyijiayj@126.com

＊通讯作者　　e-mail：sunjingtaovv@126.com

外利用近红外光谱技术对酿酒葡萄总酚含量的研究较少，研究多在某一固定采收期下进行，未在模型优化方面进行探讨，模型预测能力相对较低。

以酿酒葡萄赤霞珠为研究对象，利用近红外光谱技术结合遗传算法优化极限学习机(GA-ELM)，建立不同采收期赤霞珠葡萄总酚含量的预测模型。采用不同光谱预处理、特征变量筛选和建模方法，尝试提高模型的稳定性和预测能力，为实际生产中精准预测赤霞珠葡萄的总酚含量提供理论参考和依据。

## 1　实验部分

### 1.1　材料

试验使用的酿酒葡萄为赤霞珠，样品采集地点为新疆石河子市张裕酒庄。于 2019 年 9 月 16 日到 10 月 14 日期间进行，每间隔 7 天采集一次样品，共采集 5 次，每次采集从不同的植株上随机采取 40 穗葡萄，从每穗葡萄的上、中、下选取无病虫害、无机械损伤的葡萄果实，每 10 粒葡萄作为一个试验样品，共计 200 个样品。采收当天将样品运至 6 ℃冷藏室进行冷藏，试验前，需将样品在室温下静置 4 h，使其温度与室温基本一致。模型建立前，使用 SPXY 算法将样品按照 3∶1 比例分为校正集和预测集，共计 150 个校正集和 50 个预测集。

### 1.2　近红外光谱的采集

本研究使用德国布鲁克(BRUKER)公司生产的 TANGO-R 型傅里叶近红外光谱仪对赤霞珠葡萄的光谱进行采集。采集参数设置如下：波长范围为 12 500～4 000 $cm^{-1}$，分辨率为 8 $cm^{-1}$，扫描次数为 32 次，平滑为 3。仪器控制和初始光谱采集使用 OPUS 软件执行。将赤霞珠葡萄放在带有圆形石英底座(直径 3 cm)的光谱仪样品杯中，以反射模式收集赤霞珠的 NIR 光谱。每个葡萄平行测三次，每个样本 10 颗葡萄，共计 30 条光谱取其平均，每次测量后，应使用蒸馏水清洗光谱仪样品杯，并用纸巾擦干，以进行下一次样品的光谱采集。

### 1.3　赤霞珠葡萄总酚含量的测定

光谱采集完成后，采用 Ivanova 等[7]的方法对样品进行前处理，并将前处理后的样品液储存于－20 ℃冰箱中，用于后续总酚含量的测定。总酚含量基于福林酚比色法的原理，采用徐国前等[8]的方法进行测定，于 765 nm 波长下测量样品及不同质量浓度的没食子酸的吸光度值，并以没食子酸的质量浓度(mg·$g^{-1}$)为横坐标，吸光度为纵坐标制作标准曲线，根据标准曲线计算每个样品的总酚含量。

### 1.4　数据处理

采用多元散射(MSC)、标准正态变换(SNV)、数据中心化(MC)、移动窗口平滑(MA)和一阶导数＋SG 方法对原始光谱进行预处理，运用竞争性自适应重加权算法(CARS)、遗传算法(GA)、联合区间偏最小二乘算法(si-PLS)和连续投影算法(SPA)对全光谱进行特征波段的筛选，再结合极限学习机(ELM)和遗传算法优化极限学习机(GA-ELM)建立赤霞珠葡萄总酚含量的定量分析模型，以校正相关系数($R_c$)、预测相关系数($R_p$)、校正均方根误差(RMSEC)和预测均方根误差(RMSEP)为指标来评价模型的性能。

## 2　结果与讨论

### 2.1　赤霞珠葡萄近红外光谱分析

图 1 为不同采收期赤霞珠葡萄的原始光谱图，由图知光谱整体趋势基本一致，各谱线在反射强度上存在一定的差异，在波长为 10 200 和 8 340 $cm^{-1}$附近出现波峰，在 9 300 和 7 880 $cm^{-1}$附近出现波谷，这些波峰与波谷可能是由于赤霞珠葡萄总酚中的基团(C—H，O—H，N—H)对近红外光谱吸收率不同造成的，可为光谱与总酚建立关系提供理论基础。图 2 为不同采收期下赤霞珠葡萄的原始平均光谱，五个采收期的赤霞珠葡萄的光谱变化趋势也基本一致。由于处于不同采收阶段，葡萄内部的物质含量不同，光谱吸收的强度也有所不同。赤霞珠葡萄在第一个采收期(1WAV)即转色后的第 1 周的光谱值普遍高于其他采收阶段的光谱值，随着采收期的延长，赤霞珠葡萄的光谱值逐渐下降。

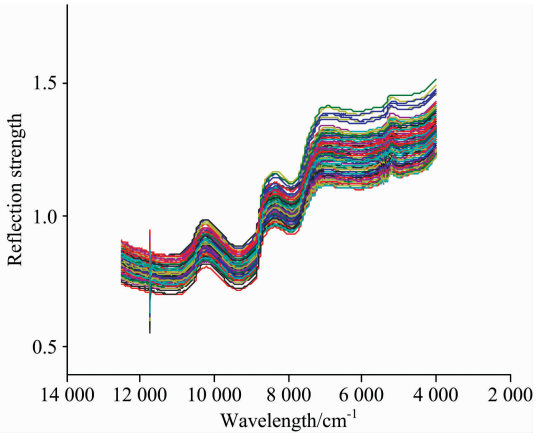


**图 1　不同采收期的赤霞珠葡萄的原始光谱曲线**

**Fig. 1　Raw spectra of Cabernet Sauvignon in different harvest stages**

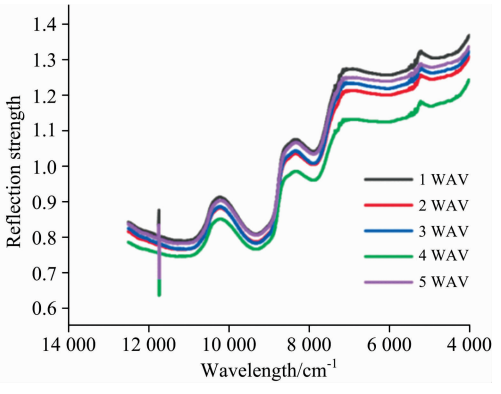


**图 2　不同采收期的赤霞珠葡萄的平均光谱曲线**

**Fig. 2　Mean raw spectra of Cabernet Sauvignon in different harvest stages**

### 2.2　不同采收期赤霞珠葡萄总酚含量统计

表 1 为五个不同采收期下赤霞珠葡萄总酚含量的统计结

果，从整体来看，赤霞珠葡萄在各个采收阶段总酚含量上升趋势明显。在 1 WAV 赤霞珠葡萄的总酚含量平均值为 18.99 mg·g$^{-1}$，随着赤霞珠葡萄采收期的延长，总酚含量持续上升，在 4 WAV 总酚含量的平均值达到最大为 28.10 mg·g$^{-1}$，到 5 WAV 总酚含量稍有下降，这与总酚在酿酒葡萄成熟期的变化趋势大体相同。不同采收期的赤霞珠葡萄总酚含量差异性比较明显，含量范围较大(11.23～34.85 mg·g$^{-1}$)。

**表 1 不同采收期的赤霞珠葡萄的总酚含量统计结果**

**Table 1 Statistical results of total phenol content of Cabernet Sauvignon in different harvest stages (mg·g$^{-1}$)**

| WAV | 样本数量 | 最大值 | 最小值 | 平均值 | 标准偏差 |
|---|---|---|---|---|---|
| 1 | 40 | 28.01 | 14.65 | 18.99 | 2.956 |
| 2 | 40 | 28.96 | 17.07 | 22.20 | 2.919 |
| 3 | 40 | 29.64 | 11.23 | 24.61 | 3.366 |
| 4 | 40 | 34.85 | 19.86 | 28.10 | 3.576 |
| 5 | 40 | 32.64 | 17.12 | 25.68 | 4.397 |

### 2.3 光谱数据预处理

在光谱采集过程中，由于葡萄表面的杂散光、采集环境、采集仪器等因素的影响使采集后的光谱包含大量噪声信息，影响了模型的预测结果[9]。因此，在建立预测模型前需对光谱进行预处理，本研究主要使用 MSC、SNV、MC、MA 和一阶导数＋SG方法对原始光谱进行预处理，并将预处理后的光谱建立 ELM 总酚含量预测模型，结果见表 2。

**表 2 不同预处理方法的总酚含量 ELM 模型建模结果**

**Table 2 Results of ELM modeling for the total phenol content with different pre-treatment methods**

| 预处理方法 | 校正集 | | 预测集 | | RPD |
|---|---|---|---|---|---|
| | $R_c$ | RMSEC | $R_p$ | RMSEP | |
| 原始 | 0.645 4 | 3.615 4 | 0.595 9 | 3.487 5 | 1.245 3 |
| MSC | 0.688 0 | 3.494 1 | 0.680 9 | 2.943 1 | 1.365 5 |
| SNV | 0.725 4 | 3.313 9 | 0.654 3 | 3.039 1 | 1.322 3 |
| MC | 0.674 5 | 3.494 7 | 0.610 1 | 3.441 0 | 1.262 1 |
| MA | 0.646 4 | 3.591 3 | 0.606 9 | 3.506 8 | 1.258 2 |
| 一阶导数＋SG | 0.634 2 | 3.649 2 | 0.623 1 | 3.501 3 | 1.278 6 |

由表 2 可知，经预处理后的光谱模型精度较原始光谱均有提高，其中 MSC 处理过的模型预测能力明显提高，$R_p$ 为 0.680 9，RMSEP 为 2.943 1，RPD 为 1.365 5，故将 MSC 处理后的光谱进行后续的研究。图 3 为 MSC 处理后的反射光谱，光谱之间的紧密性显著增强。

### 2.4 特征波长提取

全波段光谱信息中，存在数据冗杂和共线性变量等现象，为了精简模型、提高模型预测能力需要对光谱进行特征波长提取。本研究使用的特征波长提取方法有：CARS，GA，si-PLS 和 SPA。

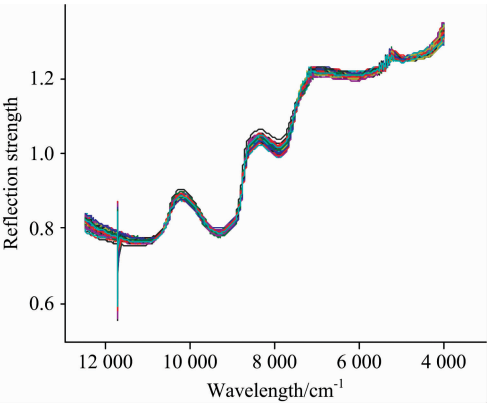


**图 3 MSC 处理后的反射光谱**

**Fig. 3 Reflectance spectra obtained after MSC**

#### 2.4.1 竞争性自适应重加权算法(CARS)

CARS 基于达尔文的“适者生存”思想，利用自适应重加权采样技术和指数衰减函数在建立的模型中循环运行，根据交叉验证均方根误差(RMSECV)最小值，确定最优的变量子集[10]。图 4(a)显示 CARS 采样次数与筛选出来的波长变量个数之间的关系，随着采样次数的增加筛选出来的波长变量数呈现出由快到慢的趋势，这是 CARS 算法筛选波长变量数由粗选到细选的一个过程。图 4(b)显示的是采样次数与 RMSECV 之间的关系，随着采样次数的增加，RMSECV 值呈现先降低后升高的趋势，当采样为第 25 次的时候，对应的 RMSECV 最小为 2.333 3。图 4(c)是采样过程中各波长变量

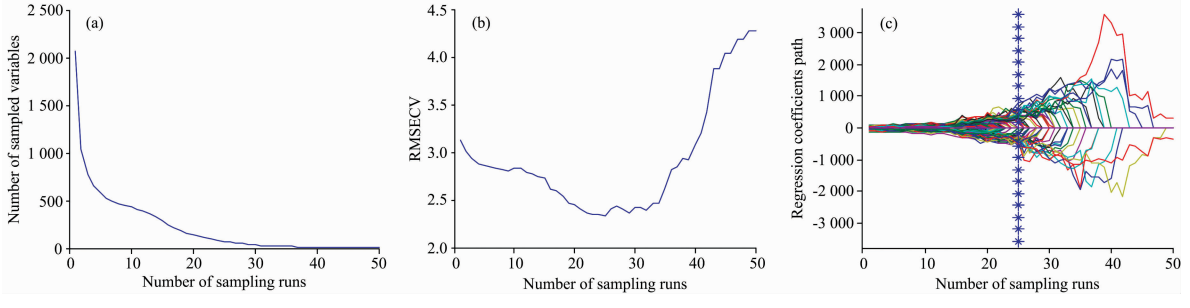


**图 4 CARS 特征波长提取**

(a)：波长变量变化；(b)：交叉验证均方根误差变化；(c)：变量回归系数的变化趋势

**Fig. 4 Characteristic wavebands selected by CARS algorithm**

(a): Variation of the number of selected wavelength variables; (b): Variation of RMSECV; (c): The changing trend of variable regression coefficients

的回归系数变化路径，图中的每条线描述了不同采样周期下所有波长变量的系数，星垂线表示采样次数为 25，结合图 4(b)可知此时 RMSECV 值最小，对应筛选出来的特征波长数为 69，仅占全波段的 3.34%。

2.4.2　遗传算法(GA)

GA 是基于自然选择原理，通过不断迭代全局寻优的一种算法[11]。该算法的参数设置为：初始种群 50，变异率 0.005，遗传迭代次数为 100 和收敛率为 0.5。图 5(a)为赤霞珠葡萄总酚变量选取频率图，超出绿色实线以上的变量为选中的波长变量。图 5(b)为所选波长变量数与 RMSECV 之间的关系图，随着所选波长数的逐渐变大，RMSECV 值呈现逐渐减小的趋势，减小趋势由快到慢，当筛选出来的波长数为 168 个(绿色圆点表示)，RMSECV 最小为 2.636 9，所选的波长数占全波段的 8.19%。

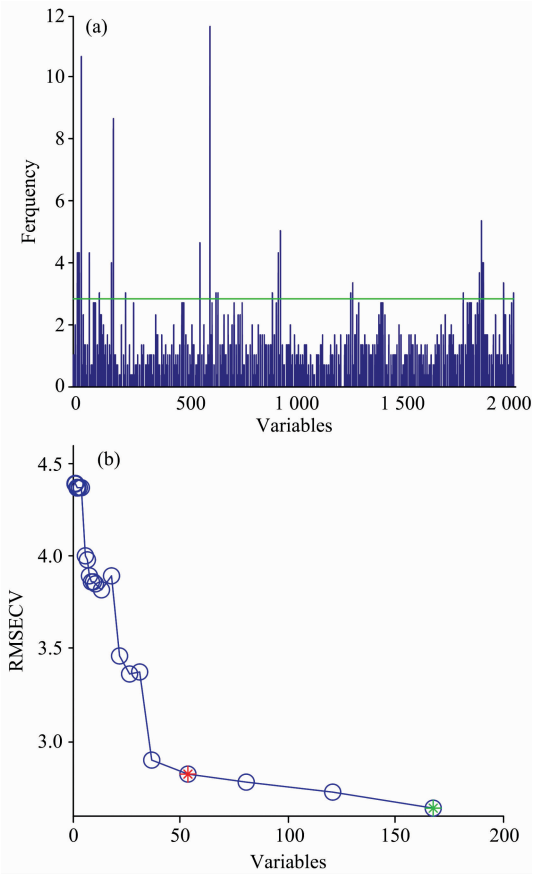


**图 5　(a)GA 筛选波段的频率分布；(b)GA 波长提取**

**Fig. 5　(a) Frequency distribution of bands selected by GA; (b) Selection of wavelengths by the GA method**

2.4.3　联合区间偏最小二乘算法(si-PLS)

si-PLS 是通过将全光谱区间进行划分，并组合区间中模型精度较高的子区间共同预测样品理化指标的一种算法[1]。将 MSC 算法处理过的全光谱划分为 20 个子区间，在主成分为 9 时，交叉验证均方根误差(RMSECV)最小为 2.979，并联合 1，15，18，20 四个子区间，共同预测赤霞珠葡萄总酚含量。四个子区间对应的波段区间为 12 492.46～12 068.08，6 538.77～6 118.50，5 265.62～4 845.36 和 4 416.86～3 996.60 $cm^{-1}$，共计 413 个特征波长变量，占全波段的 20%，如图 6 所示。

2.4.4　连续投影算法(SPA)

SPA 是一种向前循环算法，通过向量投影分析筛选出冗杂信息量少的波长变量，能够有效解决信息重叠、共线影响，提高模型的预测能力[12]。研究中设定 SPA 选取的特征波长数量为 50～100 个，采用 SPA 算法对 MSC 处理过的全光谱进行特征波长提取。当 RMSE 为 3.204 3 时，SPA 算法提取出来 50 个特征波长，占全光谱的 2.42%，其波长变量分布如图 7 所示。

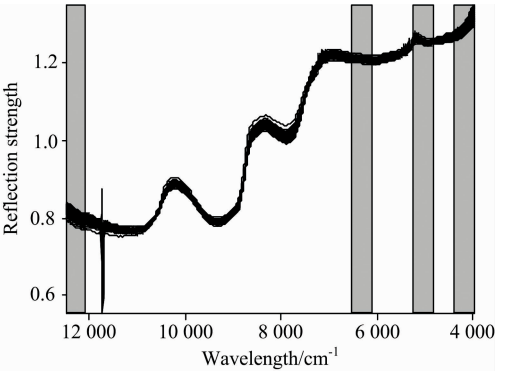


**图 6　si-PLS 特征波长提取**

**Fig. 6　Characteristic wavebands selected by si-PLS method**

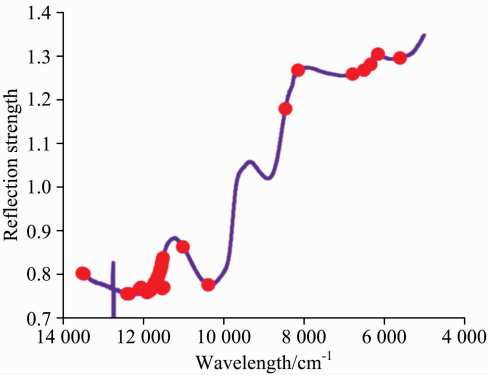


**图 7　通过 SPA 选择的 50 个特征波长分布图**

**Fig. 7　Plot of 50 wavelengths selected by SPA**

## 2.5　GA-ELM 模型建立

ELM 是一种单隐含层前馈神经网络学习算法，输入层、隐含层和输出层三部分组成，其中模型的阈值和权值可随机产生，无需迭代调整，因此 ELM 模型的训练速度和泛化能力较传统的模型具有速度快、性能好的优势。由于 ELM 模型的权值和阈值随机性，部分权值和阈值可能达不到最优状态，导致模型每次运行的结果不一致，而遗传算法(GA)是常用的全局优化算法，可以很好地寻找 ELM 神经网络的权值和阈值的最优值。

2.5.1　GA 参数设置

在 GA 对 ELM 模型的权值和阈值优化过程中，适宜的

模型参数设置尤为重要。GA 参数经多次运行确定，包括：种群规模为 50，最大遗传代数为 20，代沟为 0.95，交叉变异概率为 0.7，变异概率为 0.01。

2.5.2 ELM 参数设置

激活函数是影响 ELM 模型预测性能的关键因素。分别使用 Sigmoidal，Sine 和 Hardlim 三种不同的激活函数，在隐含层神经元个数为 40 时，利用 CARS，GA，si-PLS 和 SPA 算法对 MSC 处理后的全光谱进行特征波长的筛选，将筛选出来的波长作为输入变量建立 GA-ELM 模型，根据模型的预测结果来比较不同激活函数的性能。表 3 是经 MSC 处理后，用 CARS 提取特征波长的 GA-ELM 模型，由表 3 可知当激活函数为 Sigmoid 时，模型的性能相对较好，且运行时间与其他激活函数相差不大。利用 GA，si-PLS 和 SPA 这三种算法提取特征波长后建立的模型，当激活函数均为 Sigmoidal 时，模型的预测能力相对较高，因此后续的研究选用 Sigmoidal 激活函数。

**表 3 不同激活函数的性能比较**

**Table 3 Performance comparison of different activation functions**

| 激活函数类型 | RMSEC | $R_c$ | RMSEP | $R_p$ | RPD | 运行时间/s |
|---|---|---|---|---|---|---|
| Sigmoid | 2.196 0 | 0.893 3 | 1.861 5 | 0.878 3 | 2.091 4 | 34.39 |
| Sine | 2.602 7 | 0.846 4 | 1.946 6 | 0.866 0 | 1.999 9 | 33.84 |
| Hardlim | 3.432 4 | 0.710 8 | 3.632 4 | 0.374 4 | 1.078 5 | 33.46 |

隐含层神经元个数对 ELM 模型的预测能力也产生着重要的影响。隐含层神经元个数偏少会导致模型信息处理和学习能力降低，模型出现“欠拟合”状态，如果隐含层神经元个数偏多，模型结构将变得复杂、运行时间将变长，模型出现“过拟合”状态。因此需要通过多次尝试变换隐含层神经元个数，多次运行 GA-ELM 模型，从而筛选出最优的隐含层神经元个数。

设置 GA-ELM 的隐含层神经元个数初始化为 10，并以 10 为间隔逐步增加至 100，激活函数设为 Sigmoidal，利用 CARS，GA，si-PLS 和 SPA 算法对 MSC 处理后的光谱进行特征波长的筛选，将筛选出来的波长作为输入变量建立 GA-ELM模型，根据模型的预测结果选取合适的隐含层神经元个数。图 8 为不同隐含层神经元个数对 MSC＋CARS＋GA-ELM 模型预测能力的影响，当隐含层神经元个数小于 50 时，$R_c$ 小于 $R_p$，模型欠拟合；当隐含层神经元个数大于 50 时，$R_c$ 远大于 $R_p$，模型过拟合；当隐含层神经元个数为 50 时，$R_c$ 为 0.901 7，$R_p$ 为 0.901 3，模型适度拟合，故该模型组合最合适的隐含层神经元个数为 50。另外 MSC＋GA＋GA-ELM，MSC＋si-PLS＋GA-ELM，MSC＋SPA＋GA-ELM 这三种模型组合，在隐含层神经元个数同样设置为 50 时，预测能力相对较高，因此后续研究中模型的隐含层神经元个数设为 50。

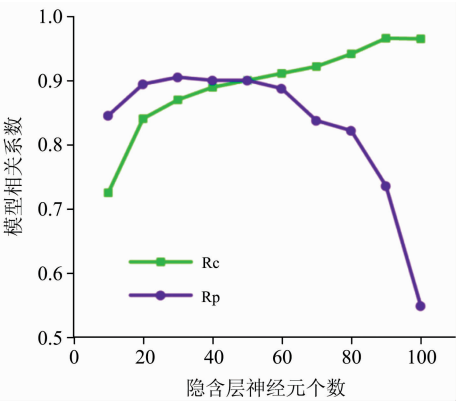


**图 8 不同隐含层神经元个数对模型性能的影响**

**Fig. 8 The influence of the number of neurons in different hidden layers on the performance of the model**

### 2.6 建模结果与分析

经 MSC 处理后，利用 CARS，GA，si-PLS 和 SPA 算法筛选出来的波长变量数分别为 69，168，413 和 50，将筛选出来的波长作为输入变量分别建立 ELM 和 GA-ELM 预测模型，预测结果如表 4 所示。在特征波长提取方面，经 SPA 算法筛选出来的波长变量数目最少，建立的模型预测结果不高，其原因可能是经 SPA 筛选出来的 50 个特征波长中与赤霞珠总酚相关的部分特征波长变量被剔除，造成模型预测能力的降低；经 si-PLS 算法筛选出来的特征波长变量数目最多，但是在这些有效区间内仍存在共线性变量，从而导致模型的预测能力降低；经 CARS 提取出来的 69 个变量建立的模型预测结果均高于其他模型，ELM 模型 $R_p$ 为 0.824 5，

**表 4 ELM 和 GA-ELM 模型预测效果**

**Table 4 The prediction results of ELM and GA-ELM models**

| Prediction models | Selection methods | Variables number | Calibration sets | | Prediction sets | | RPD |
|---|---|---|---|---|---|---|---|
| | | | $R_c$ | RMSEC | $R_p$ | RMSEP | |
| ELM | CARS | 69 | 0.856 3 | 2.524 2 | 0.824 5 | 2.203 2 | 1.767 0 |
| | GA | 168 | 0.763 8 | 3.158 8 | 0.732 8 | 2.603 8 | 1.469 5 |
| | Si-PLS | 413 | 0.733 4 | 3.346 4 | 0.669 9 | 2.778 2 | 1.346 9 |
| | SPA | 50 | 0.742 7 | 3.216 4 | 0.731 4 | 2.847 1 | 1.466 4 |
| GA-ELM | CARS | 69 | 0.901 7 | 2.112 4 | 0.901 3 | 1.686 8 | 2.308 0 |
| | GA | 168 | 0.835 6 | 2.688 8 | 0.834 7 | 2.107 3 | 1.815 8 |
| | Si-PLS | 413 | 0.798 7 | 2.962 2 | 0.719 1 | 2.600 3 | 1.375 4 |
| | SPA | 50 | 0.854 7 | 2.493 7 | 0.844 9 | 2.233 4 | 1.869 3 |

GA-ELM 模型 $R_p$ 为 0.901 3，结果表明，CARS 算法可以明显提高总酚的预测能力。选择合适的特征波长提取方法可以有效剔除光谱中的冗杂信息，降低变量中的共线性，提高近红外光谱与赤霞珠葡萄总酚含量之间的相关性，从而提高模型的预测能力。在建模算法方面，将 ELM 和 GA-ELM 模型进行比较，经不同特征波长提取方法建立的 ELM 模型 $R$ 均为 0.669 9 以上，RMSE 均低于 3.346 4，其中 MSC+CARS+ELM 的模型预测结果相对较好，模型的 $R_p$ = 0.824 5，RMSEP=2.203 2，RPD=1.767 0；而利用不同特征波长优选方法建立的 GA-ELM 模型的 $R$ 均在 0.719 1 以上，RMSE 均低于 2.962 2，较传统的 ELM 模型，$R$ 得到提高，RMSE 相应降低，其中 MSC+CARS+GA-ELM 预测结果最好，预测模型的校正集和预测集散点图如图 9 所示，模型的 $R_c$ = 0.901 7，$R_p$ = 0.901 3，RMSEC = 2.112 4，RMSEP = 1.686 8，RPD=2.308 0。由结果可知 GA-ELM 模型的预测能力整体要高于 ELM 模型的预测能力，因此 GA 可以有效优化 ELM 算法。

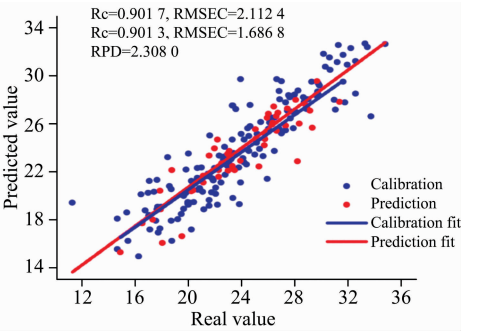


**图 9　总酚含量真实值与预测值的散点图**

**Fig. 9　Scatter plot of real value and predicted value of total phenol content**

## 3　结　论

利用近红外光谱结合 GA-ELM 网络建立赤霞珠葡萄总酚含量预测模型。采用 MSC 算法对光谱进行预处理，结合 CARS，GA，si-PLS 和 SPA 算法对波长变量进行优选，并基于所选的特征波长建立 ELM 模型，为了进一步提高模型的精度，采用 GA 算法来优化 ELM 神经网络。结果表明，利用 CARS 算法对 MSC 处理后的全光谱进行特征波长的筛选，将筛选出来的 69 个波长作为输入变量建立 GA-ELM 模型的预测能力最好，$R_c$ 为 0.901 7，$R_p$ 为 0.901 3，RMSEC 为 2.112 4，RMSEP 为 1.686 8，RPD 为 2.308 0。利用近红外光谱技术，采用变量优选和 GA-ELM 算法对赤霞珠葡萄总酚含量的预测是一种有效的方法。

## References


[1] WANG Wen-xia, MA Ben-xue, LUO Xiu-zhi, et al(王文霞, 马本学, 罗秀芝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(2): 543.

[2] GAO Sheng, WANG Qiao-hua, LI Qing-xu, et al(高　升, 王巧华, 李庆旭, 等). Analytical Chemistry(分析化学), 2019, 47(6): 156.

[3] Yu J, Wang H, Sun X, et al. Journal of Food Measurement and Characterization, 2017, 11(6): 1.

[4] XU Feng, FU Dan-dan, WANG Qiao-hua, et al(许　锋, 付丹丹, 王巧华, 等). Food Science(食品科学), 2018, 39(8): 149.

[5] Michael F, Ralph B B, Andrew G R. American Journal of Enology and Viticulture, 2016, 67(1): 38.

[6] ZHANG Lin-zhong, CAI Xue-zhen, FANG Cong-bing(章林忠, 蔡雪珍, 方从兵). Acta Agriculterae Zhejiangensis(浙江农业学报), 2018, 30(2): 330.

[7] Ivanova V, Stefova M, Chinnici F. Journal of the Serbian Chemical Society, 2010, 75(1): 45.

[8] XU Guo-qian, ZHANG Zhen-wen, GUO An-que, et al(徐国前, 张振文, 郭安鹊, 等). Food Science(食品科学), 2010, 31(18): 275.

[9] GAO Tong, WU Jing-zhu, MAO Wen-hua, et al(高　彤, 吴静珠, 毛文华, 等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2019, 50(S1): 399.

[10] YANG Bao-hua, CHEN Jian-lin, CHEN Lin-hai, et al(杨宝华, 陈建林, 陈林海, 等). Transcations of the Chinese Society of Agricultural Engineering(农业工程学报), 2015, 274(22): 184.

[11] Qiu Y, Zhu R, Fan Z, et al. Spectroscopy Letters, 2018, 51(5): 226.

[12] Zhang C, Jiang H, Liu F, et al. Food and Bioprocess Technology, 2017, 10(1): 213.

# Quantitative Analysis of Total Phenol Content in Cabernet Sauvignon Grape Based on Near-Infrared Spectroscopy

LUO Yi-jia[1], ZHU He[1], LI Xiao-han[1], DONG Juan[1], TIAN Hao[1], SHI Xue-wei[1], WANG Wen-xia[2], SUN Jing-tao[1]*
1. College of Food Science, Shihezi University, Shihezi 832003, China
2. College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China

**Abstract** The contents of total phenol in wine grape are an important indicator of grape quality and also a key factor of wine quality directly. To detect the total phenol contents of the cabernet sauvignon grape quickly and accurately, this paper used near-infrared spectroscopy and GA-ELM prediction model to predict the total phenol content of Cabernet Sauvignon grapes. In the experiment, Cabernet Sauvignon grapes were collected in 5 harvest periods (40 bunches were collected in each harvest period, and 10 grapes were acquired in each cluster), and near-infrared spectra information in the range of 12 500～4 000 $cm^{-1}$ was collected for 200 groups of grapes. The total phenol content of Cabernet Sauvignon grapes was determined based on the principle of Folin-Ciocalteus colorimetry, SPXY algorithm was used to divide the samples into correction sets and prediction sets at a ratio of 3∶1, with a total of 150 correction sets and 50 prediction sets. Multiplicative Scatter Correction (MSC), Standard Normalized Variate (SNV), Mean Centering (MC), Moving Average (MA), and the First Derivative +SG was used to preprocess the raw spectra, MSC was compared as the best pretreatment method. And then, competitive adaptive reweighted sampling (CARS), genetic algorithm (GA), successive projections algorithm (SPA) and synergy interval partial least squares (si-PLS) were extracted the characteristic wavelengths, respectively. The comparative analysis found that the 69 characteristic wavelength variables extracted by CARS could effectively improve the model's stability and prediction ability. Based on the MSC and different variable optimization methods, the extreme learning machine (ELM) algorithm was introduced to establish the total phenol content prediction model. In predicting total phenol content, a genetic algorithm (GA) was used to optimize the ELM model and the influence of different kernel functions and the number of hidden layer neurons on the prediction ability of the GA-ELM model investigated. The optimal kernel function was Sigmoidal, and the optimal number of neurons was 50. Finally, the prediction capabilities of the ELM and GA-ELM models were compared. The results showed that GA-ELM models were more accurate in predicting than the ELM models, and the MSC+CARS+GA-ELM model was the best with a correlation coefficient of calibration ($R_c$) of 0.901 7, the correlation coefficient of prediction of 0.901 3, the root mean square error of calibration (RMSEC) of 2.112 4, the root mean square error of prediction (RMSEP) of 1.686 8 and residual prediction deviation (RPD) of 2.308 0. The combination of variable optimization methods and the GA-ELM model was an effective method, which provided a theoretical basis for detecting Cabernet Sauvignon grapes' quality.

**Keywords** Variable optimization; Cabernet sauvignon grapes; Total phenol; Extreme learning machine; Near infrared spectroscopy

(Received Jul. 2, 2020; accepted Nov. 19, 2020)

* Corresponding author