

基于不同算法对桂枝茯苓胶囊内容物吸湿性预测建模研究

陈 琪¹, 徐芳芳^{2,3*}, 张 欣^{2,3}, 徐 冰⁴, 吴 云^{2,3}, 王振中^{2,3}, 肖 伟^{1,2,3*}

1. 南京中医药大学, 江苏 南京 210023

2. 江苏康缘药业股份有限公司, 江苏 连云港 222001

3. 中药制药过程新技术国家重点实验室, 江苏 连云港 222001

4. 北京中医药大学, 北京 100029

摘要: 目的 比较不同算法对桂枝茯苓胶囊内容物吸湿性预测模型性能的影响, 确定最优建模算法。方法 以 54 个物理性质参数为输入, 胶囊内容物吸湿性为输出, 对比偏最小二乘算法 (partial least squares, PLS)、决策树算法 (classification and regression tree, CART)、多元自适应回归样条算法 (multivariate adaptive regression splines, MARS) 和广义路径追踪算法 (generalized path seeker, GPS) 对建立吸湿性预测模型性能的影响。结果 MARS 算法建立的预测模型性能最佳, 预测能力最强, 模型的校正集决定系数 (R^2_c) 为 0.843, 预测集决定系数 (R^2_p) 为 0.808, 校正集均方根误差 (root mean square error of calibration, RMSEC) 为 0.391, 预测集均方根误差 (root mean square error of prediction, RMSEP) 为 0.472, 平均相对预测误差为 2.69%, 小于 5%。结论 MARS 算法建立的吸湿性预测模型更适合桂枝茯苓胶囊的生产应用, 该算法可嵌入在线控制系统, 为生产过程的质量控制智能化提供技术支持。

关键词: 桂枝茯苓胶囊; 吸湿性; 算法; 建模; 预测精度; 偏最小二乘算法; 决策树算法; 多元自适应回归样条算法; 广义路径追踪算法; 决定系数; 均方根误差; 在线控制; 质量控制; 智能化

中图分类号: R283.6 文献标志码: A 文章编号: 0253-2670(2021)11-3216-08

DOI: 10.7501/j.issn.0253-2670.2021.11.008

Research on predictive model for hygroscopicity of contents in Guizhi Fuling Capsules based on different algorithms

CHEN Qi¹, XU Fang-fang^{2,3}, ZHANG Xin^{2,3}, XU Bing⁴, WU Yun^{2,3}, WANG Zhen-zhong^{2,3}, XIAO Wei^{1,2,3}

1. Nanjing University of Chinese Medicine, Nanjing 210023, China

2. Jiangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

3. State Key Laboratory of New-tech for Chinese Medicine Pharmaceutical Process, Lianyungang 222001, China

4. Beijing University of Chinese Medicine, Beijing 100029, China

Abstract: Objective The purpose of this paper was to study the differences in the stability and prediction accuracy of the moisture absorption prediction models of the contents of Guizhi Fuling Capsules (桂枝茯苓胶囊, GFC) established by different algorithms and choose the optimal modeling algorithm. **Methods** In this paper, 54 physical property parameters of the measured powder properties were taken as input, and the hygroscopicity of the content was taken as output. The traditional modeling algorithm partial least squares algorithm, classification and regression tree, multivariate adaptive regression splines and generalized path seeker algorithm were used to establish the moisture absorption prediction models, and the effect of different models on the prediction accuracy of the verification set was compared and studied. **Results** The best model was established by the multivariate adaptive regression splines (MARS) algorithm among the machine learning algorithm which had the best fitting effect and the strongest predictive ability. The calibration determination coefficient of model was 0.843, the prediction determination coefficient was 0.808, and the correction root mean square error of calibration was 0.391. The root mean square error of prediction was 0.472, and the mean

收稿日期: 2020-12-15

基金项目: 国家“重大新药创制”科技重大专项 (2018ZX09201010-004)

作者简介: 陈 琪, 硕士研究生。E-mail: 1654845280@qq.com

*通信作者: 肖 伟, 研究员级高级工程师, 博士生导师, 研究方向为中药新药的研究与开发。E-mail: kanionlunwen@163.com

徐芳芳, 女, 博士, 研究方向为过程分析技术。E-mail: 879164331@qq.com

relative prediction error was 2.69%. **Conclusion** The hygroscopicity prediction model established by the MARS algorithm was more suitable for production applications, which provided a basis for the quality control of the production process of GFC.

Key words: Guizhi Fuling Capsules; hygroscopicity; algorithm; modeling; prediction accuracy; partial least squares; classification and regression tree; multivariate adaptive regression splines; generalized path seeker; determination coefficient; root mean square error; online control; quality control; intelligent

桂枝茯苓胶囊 (Guizhi Fuling Capsules, GFC) 是由桂枝、茯苓、牡丹皮、桃仁和芍药 5 味药材组成, 具有活血、化瘀、消癥等功效, 用于治疗痛经、子宫肌瘤等疾病。GFC 内容物属于中药半浸膏粉, 具有吸湿性强、易结块等特点^[1-2], 会影响胶囊的装量差异和稳定性等关键质量属性。王晴等^[1]采用偏最小二乘算法 (partial least squares, PLS) 建立 GFC 内容物吸湿性预测模型, 该建模方法在生产应用中需手动筛选关键变量, 数据预处理工作量大, 所建模型的交叉验证决定系数 (determination coefficient of cross-validation, Q^2)、校正集决定系数 (determination coefficient of calibration, R^2_c) 和预测集决定系数 (determination coefficient of prediction, R^2_p) 较高且需进一步提升。本实验尝试从算法角度提高模型性能, 研究不同算法对模型性能的影响。

决策树算法 (classification and regression tree, CART)、多元自适应回归样条算法 (multivariate adaptive regression splines, MARS) 都可自动筛选变量, 处理速度快, 能够提升建模效率和模型性能, 还可利用预测建模技术自动化、智能化对缺失值进行处理。吴晓倩等^[3]引入决策树算法, 利用与成绩相关的数据实现对学生未来成绩的预测, 预测准确率达到 94%。仇文岗等^[4]基于大量可靠的岩土监测数据, 运用 MARS 算法对输入、输出参数进行相关性拟合, 得到各输入参数的相对重要性。2 个评估实例中测试集的 R^2_c 分别为 0.921、0.986。上述研究表明 CART 算法和 MARS 算法可以弥补 PLS 算法的缺陷和不足。广义路径追踪算法 (generalized pathseeker, GPS) 算法的应用未见报道, 但它也具备自动筛选变量的优点, 模型性能情况有待进一步研究。上述 3 种可自动筛选变量的算法都属于目前机器学习算法, 在胶囊内容物吸湿性方面的应用未见报道, 本实验基于现有数据运用不同算法建模, 优选出最佳算法后嵌入在线控制系统中进行后续生产验证, 以此来提升 GFC 生产过程中质量控制智能化水平。

1 仪器与数据来源

BT-1001 智能粉体特性测试仪, 丹东百特仪器

有限公司; Bettersize 2600 激光粒度分布仪, 丹东百特仪器有限公司; LHS-250HC-II 恒温恒湿箱, 上海一恒科学仪器有限公司。

本研究所使用的数据来自于前期研究获得的 80 批 GFC 生产过程中的 5 种中间体 (原料细粉、湿颗粒、干燥颗粒、整粒颗粒、总混颗粒) 和 GFC 成品的物性参数, 收集的样品为 2018 年 6 月份至 2018 年 11 月份生产的中间体样品, 样品的批次为 180610~181121 共 480 个样品, 均由江苏康缘药业股份有限公司提供^[1]。

2 方法

2.1 样本集的划分

本研究按照实际生产时间顺序将样本以 3:1 比例划分校正集和验证集, 前 60 批次为校正集, 后 20 批次为验证集, 校正集的范围覆盖了预测集的整个范围, 且校正集分布较广, 适宜建立定量校正模型。

2.2 物性表征方法

2.2.1 粒径 使用激光粒度分布仪测定颗粒粒度分布, 计算 D_{10} (累计粒度分布数达到 10% 时所对应的粒径)、中值粒径 (D_{50} , 累计粒度分布数达到 50% 时所对应的粒径) 和 D_{90} (累计粒度分布数达到 90% 时所对应的粒径), 以及粒径分布宽度 (span)。

$$\text{span} = (D_{90} - D_{10}) / D_{50}$$

2.2.2 粒径 < 50 μm 百分比 (%Pf) 和相对均齐度指数 (I_0) 以空气为分散媒介, 使用激光粒度分布仪测定颗粒粒度分布, 计算每个粒径范围内颗粒所占的比例, 平行 3 次试验。所选取的粒径分别为 355、212、100、50 μm 。

$$I_0 = F_m / [100 + (d_m - d_{m-1})F_{m-1} + (d_{m+1} - d_m)F_{m+1} + (d_m - d_{m-2})F_{m-2} + (d_{m+2} - d_m)F_{m+2} + \dots + (d_{m+n} - d_m)F_{m+n}]$$

F_m 为颗粒粒径在多数范围 (100~212 μm) 的质量百分比, F_{m-1} 为多数粒径范围下一层筛子 (50~100 μm) 截留颗粒的质量百分比, F_{m+1} 为多数粒径范围上一层筛子 (212~355 μm) 截留颗粒的质量百分比, d_m 为多数粒径范围的颗粒平均粒径, d_{m-1} 为多数粒径范围下一层筛子截留颗粒平均粒径, d_{m+1} 为多数粒径范围上一层筛子截留颗粒平均粒径, n 为所确定的粒径范围个数

2.2.3 休止角 (α) 采用固定底面积法, 使用粉末流动性测定仪对样品进行测试, 底面圆盘直径 100 mm, 喷嘴直径为 10 mm。取待测中间体颗粒约 50.0 g, 将仪器校零后, 从特殊喷嘴漏斗的上方缓慢加入, 使其逐渐堆积在圆盘上, 形成锥体, 测定锥体高 (h) 和底面圆盘半径 (d), 平行 3 次试验。

$$\alpha = \arctan(h/d)$$

2.2.4 松装密度 (D_a) 将约 50.0 g (质量记为 m) 待测颗粒缓慢地加入到 100 mL 量筒中, 轻轻抹平表面, 读取待测颗粒的体积 (V_a), 平行 3 次试验。

$$D_a = m/V_a$$

2.2.5 振实密度 (D_c) 将上述盛有待测颗粒的量筒固定在振实密度仪上, 设定频率 70 Hz, 振动 1250 次后读取待测颗粒的体积 (V_c), 平行 3 次试验。

$$D_c = m/V_c$$

2.2.6 豪斯纳比 (Hausner ratio, HR) 由松装密度和振实密度计算而得。

$$HR = D_c/D_a$$

2.2.7 卡尔指数 (Carr index, CI) 由松装密度和振实密度计算而得。

$$CI = (D_a - D_c)/D_c$$

2.2.8 颗粒间孔隙率 (I_e) 由松装密度和振实密度计算而得。

$$I_e = (D_a - D_c)/D_a D_c$$

2.2.9 吸湿性 (H) 取干燥的具塞称量瓶置于恒温恒湿箱 (设置温度为 25 °C, 相对湿度为 75%) 中, 12 h 后精密称定质量 (m_1)。取待测颗粒适量, 置上述称量瓶中并平铺于称量瓶内, 厚度约为 1 mm, 精密称定质量 (m_2)。将称量瓶敞口, 并与瓶盖同置于上述恒温恒湿条件下。24 h 后盖好称量瓶盖, 精密称定质量 (m_3), 平行 3 次试验。

$$H = (m_3 - m_2)/(m_2 - m_1)$$

本实验建模所用的输入参数 X 为软材细粉 (RC)、干颗粒 (GZ)、整粒颗粒 (ZL)、总混颗粒 (ZH) 的 D_{10} 、 D_{50} 、 D_{90} 、 D_a 、 D_c 、span、%Pf、 I_0 、HR、 α 、 I_e 、CI 以及湿颗粒 (SKL) 的 D_a 、 D_c 、HR、 α 、 I_e 、CI 共 54 个参数, 输出参数 Y 为胶囊内容物 H 值。

2.3 数据分析软件

采用 SIMCA-P 软件 (version 12.0, 瑞典 MKS Umetrics 公司) 建立 PLS 模型, 采用 SPSS25.0 进行自变量共线性评价, 采用 SPM8.3 软件 (美国 Salford Systems 公司) 中不同机器学习算法建立吸

湿性预测模型。

2.4 建模原理与方法

2.4.1 PLS 算法 PLS 是经多次迭代映射分解, 组合成 1 个 Y 与 X 的线性回归 ($Y = bX + f$), 将每次迭代过程中的映射变换参数进行线性运算可以得到最终模型的参数 b 和 f 。PLS 采用成分分解的思路来进行特征降维; PLS 关注各个变量对目标变量的解释, 会保留共线性的变量; PLS 先拟合大致轮廓, 再逐步拟合细节, 考虑当前的最优拟合决策进行局部最优化。

2.4.2 CART CART 具有高速、精确和容易使用的特点, 并自动对数据提供深入的探索研究, 产生高度可理解的预测模型; 仅使用排序信息因而不需要进行变量变形, 不会受到孤立点的影响, 同时使用替代分裂变量智能的处理缺失值。CART 是非参数的, 没有预设的函数形式, 它通过局域性的总结来逼近数据模式, 动态地决定使用哪个变量, 动态地决定关注哪个区域, 所有的一切都由数据自动决定^[5-10]。

2.4.3 MARS MARS 是一种非线性和多维关系建模的算法, 可以高度自动化的回归分析, 可以看成是逐步线性回归的推广, 也可以看成是为了提高 CART 在回归中的效果而进行的改进。

MARS 算法分为前向选择、后向剪枝、模型选择 3 个步骤, 通过自适应的选取节点对数据进行分割, 每选取 1 个节点就生成 2 个新的基函数, 前向过程结束后生成 1 个过拟合的模型, 后向剪枝过程是删除导致模型过度拟合的基函数, 便于选择最佳模型, 这一过程遵循广义交叉验证 (generalized cross-validation, GCV) 原则, GCV 值最小的模型为最终的 MARS 模型^[11-13]。GCV 是根据误差模型和正则化解构造 1 个广义交叉 G 函数, GCV 的计算方法如下式所示。

$$GCV = 1/N \sum_{i=1}^N [y_i - f(x_i)]^2 / [1 - (H + 1 + dH)/N]^2$$

N 为样本数据总数, H 为基函数的个数, d 为惩罚系数

MARS 的主要优点是可以进行变量选择, 将相关变量与不相关的预测变量分开; 能够估计基函数的贡献, 允许预测变量的加性和交互影响来确定相应变量; 使用新的嵌套变量技术处理缺失值; 进行广泛的自我测试, 以防止过度拟合^[14-17]。

2.4.4 GPS GPS 提供了完整的正则化策略, 利用更为广谱的弹性系数建立更多候选模型, 开始时模

型里没有任何预测变量,在每一步中添加 1 个新的变量或者更新现有变量的 1 个系数,通过建立若干步数不同的路径模型,并自动选择出其中最优的线性模型。GPS 从速度和覆盖率 2 个方面对正则化回归进行了显著提升。模型效果会比传统回归更好、更稳定,同时提供了强大的变量筛选能力,更好应对大数据高纬度降维的挑战。

以上机器学习算法都可以对变量重要性列表进行建模精炼,筛选重要变量。

2.5 模型性能评价

2.5.1 PLS 模型评价 使用 R^2_c 、 R^2_p 、 Q^2 评价拟合效果,这些指标越接近 1,表示 PLS 模型拟合数据效果越好,校正或预测能力越好;用决定系数(R^2)、交叉验证均方根误差(root mean square errors of cross validation, RMSECV)、预测均方根误差(root mean square errors of prediction, RMSEP)评价模型的性能, R^2 越大, RMSECV 和 RMSEP 越小,模型的性能越好。

2.5.2 机器学习算法模型评价 以 R^2 、均方根误差(root mean square error, RMSE)、平均绝对百分比误差(mean absolute percent error, MAPE)等指标来评价模型的性能, R^2 越接近 1, MAPE 越接近 0,表明模型的性能越好;RMSE 越接近 0,表明模型的预测性能越强^[18-19]。

$$RMSE = [1/n \sum_{t=1}^n (y - y_t)^2]^{1/2}$$

$$R^2 = 1 - \sum_{i=1}^n (y - y_i)^2 / \sum_{i=1}^n (y - \bar{y})^2$$

$$MAPE = \sum_{i=1}^n |(y - y_i)/y| \times 100/n$$

n 为校正集或验证集的样本数, $t \in [1, n]$, y 为参考值, y_t 为预测值, $\bar{y}$ 为所有样品参考值的平均值

2.5.3 模型精度评价 PLS 和机器学习算法都是根据验证集中的相对预测误差来评价模型预测精度的影响。

3 结果与分析

3.1 多算法模型的建立

3.1.1 PLS 建模结果 以胶囊内容物吸湿性(Y)为因变量,54 个物性参数为自变量(X),建立的 PLS 预测模型的 Q^2 为 0.399, R^2_c 和 R^2_p 分别为 0.216、0.656。变量个数为 8 时, Q^2 达到最大值 0.549, R^2_X 和 R^2_Y 为 0.616、0.602。筛选出的物性参数包括 R_C-D_{10} 、 R_C-D_{50} 、 R_C-span 、 $R_C-\alpha$ 、 $R_C-\%Pf$ 、 $ZH-D_c$ 、

$SKL-D_a$ 、 $SKL-D_c$ 。绘制的变量投影重要性(variable importance in the projection, VIP)图如图 1 所示,自变量回归系数如图 2 所示。使用方差膨胀因子(variance inflation factor, VIF)评价以上 8 个物性参数的共线性程度, $R_C-\%Pf$ 、 $R_C-\alpha$ 、 R_C-D_{10} 、 R_C-D_{50} 、 R_C-span 、 $ZH-D_c$ 、 $SKL-D_a$ 、 $SKL-D_c$ 物性参数共线性检验结果的 VIF 值分别为 25.839、1.303、7.181、21.223、2.461、1.259、7.054、7.109。由图 1、2 和 VIF 值考察结果可知, $VIP > 1$ 、回归系数 > 0.2 、 $VIF < 10$ 的变量有 $R_C-\alpha$ 、 R_C-D_{10} 、 R_C-span 、 $SKL-D_c$ 、 $ZH-D_c$, 可作为潜在关键物料属性。以 5 个变量建立的模型性能指标 Q^2 为 0.531, R^2_c 和 R^2_p 分别为 0.616、0.615。

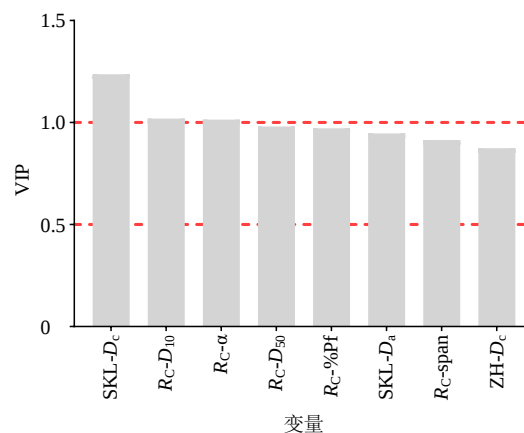


图 1 模型的 VIP 分布

Fig. 1 VIP indexes for independent variables of PLS model

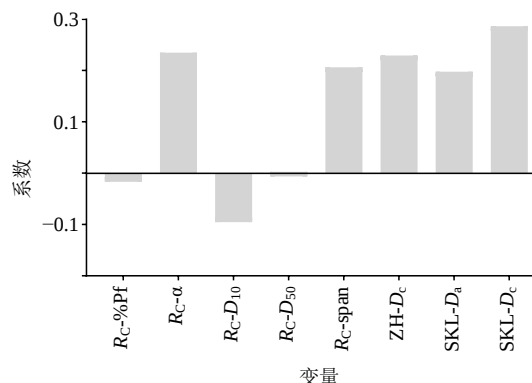


图 2 模型自变量回归系数

Fig. 2 Regression coefficients of PLS model

用 20 批样品测试所建模型的稳健性,验证集真实值与预测值相关性如图 3 所示,校正集 R^2_c 为 0.602 5,交叉验证均方根误差 RMSECV 为 0.630 9,预测集 R^2_p 为 0.348 7,预测均方根误差 RMSEP 为 0.408,验证集的相对预测误差为 3.69%。

3.1.2 CART 算法建模结果 节点和树的深度对算

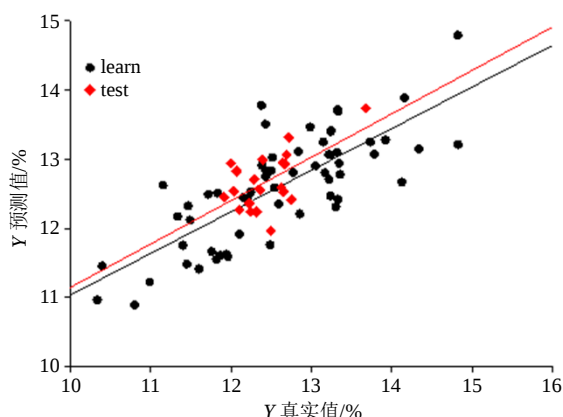


图3 PLS模型拟合效果

Fig. 3 Comparison between target and predicted hygroscopicity

法的影响较大,开始建模前通过 AUTOMATE 自动建模选项筛选最优 max nodes 参数为 8,树的深度为 30 时,性能最佳。根据变量重要性表来逐步筛选关键变量并结合验证集预测误差来确定最优模型。由 54 个物性参数建模,随机选择 20% 的样本作为验证集,所得变量重要性如图 4 所示,不同变量个数建立的 CART 模型性能和预测性能比较如表 2 所示,用 20 批样品对模型进行外部验证的结果如表 1 所示。由表 1 可知当变量数为 14 时, CART 模型的

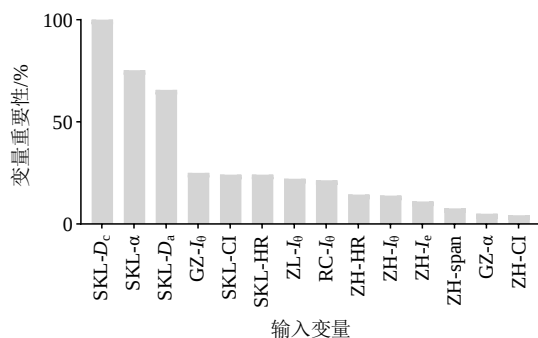


图4 变量重要性

Fig. 4 Importance of input variables

表 1 不同变量个数建立的 CART 模型的性能和预测性能比较

Table 1 Comparison of CART model performance and prediction performance established with different numbers of variables

自变 量数	校正集			预测集			平均相对预 测误差/%
	R^2_c	RMSE	MAPE	R^2_p	RMSE	MAPE	
14	0.662	0.576	0.039	0.712	0.579	0.037	5.67
13	0.542	0.260	0.018	0.034	0.391	0.023	5.72
4	0.704	0.539	0.036	0.766	0.522	0.029	6.56
3	0.759	0.486	0.030	0.845	0.425	0.024	7.47

性能不是最好但预测效果较好,变量数目减少,预测相对误差增大,可能损失了一些重要的信息。

3.1.3 MARS 算法建模结果 基函数的个数影响着预测模型的选择,可由 GCV 性能曲线得到最优基函数,因此得到最佳模型尺寸^[20-21]。由图 5 可知 GCV 值随基函数的变化规律, GCV 值最小时对应的基函数个数为 7,由表 2 可知基函数个数对 GCV 函数值和校正集验证集决定系数 R^2 和均方误差 (mean square errors, MSE) 的影响,拟合效果不随基函数增长而优化,应不断调试预设基函数的数目,以此选择最优基函数,建立最佳模型^[11]。

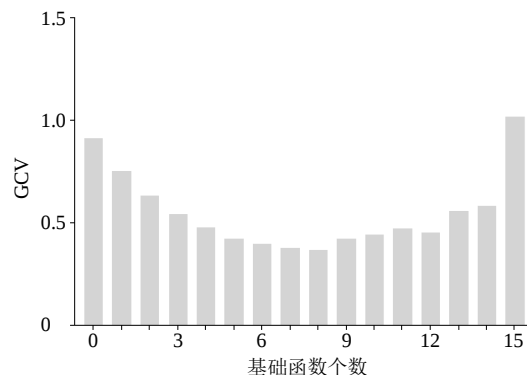


图5 不同 GCV 值对应基函数个数

Table 5 Different GCV values correspond to number of basis functions

MARS 算法中随机选择 20% 的样本做为验证集,由 54 个物性参数建立的 MARS 模型得到 8 个重要变量: SKL- D_c 、ZH- D_a 、RC- α 、GZ- D_a 、GZ- D_{90} 、SKL- D_a 、GZ- D_{10} 、GZ- α ,校正集的 R^2_c 为 0.843, RMSE 为 0.391, MAPE 为 0.025, 预测集的 R^2_p 为 0.808, RMSE 为 0.031, MAPE 为 0.031。

另取 20 批样品对 MARS 模型进行外部验证,参考值与预测值如表 3 所示,吸湿性平均相对预测误差为 2.69%,预测结果相对准确。

3.1.4 GPS 算法建模结果 MSE 表示测试集模型的性能,由来自于复杂的内部搜索结果的弹性系数决定。图 6 显示了最佳模型是一种最小化测试样品的 MSE 的模型,蓝色线是校正集,红色线是验证集,对应的模型系数为 26。不同变量个数建立的 GPS 模型性能和预测性能比较如表 4 所示,用 20 批样品对模型进行外部验证的结果如表 4 所示。

由表 4 可知用 54 个物性参数建立的模型过拟合程度较大,当逐步剔除变量重要性小的变量重新建模时,所选变量数为 3 时,模型的稳定性和预测精度最好,这 3 个变量为 SKL- D_c 、RC- α 、ZH- D_c 。

表 2 GCV 函数值和评价指标随基函数数目变化情况

Table 2 GCV function value and variations of measuring indexes with basis function number

基函数个数	输入变量个数	有效参数	GCV	GCV R^2	校正集 MSE	验证集 MSE	校正集 R^2_c	验证集 R^2_p
15	9	34.0	0.916 05	0.100 49	0.133 55	0.402 36	0.863 96	0.654 22
14	9	31.8	0.752 44	0.261 14	0.133 88	0.377 89	0.863 62	0.675 25
13	9	29.6	0.635 26	0.376 21	0.135 49	0.339 47	0.861 99	0.708 27
12	9	27.4	0.545 52	0.464 33	0.137 37	0.290 57	0.860 07	0.750 30
11	9	25.2	0.476 75	0.531 86	0.139 96	0.250 33	0.857 43	0.784 88
10	9	23.0	0.423 90	0.583 76	0.143 49	0.267 82	0.853 83	0.769 84
9	8	20.8	0.396 80	0.610 37	0.153 42	0.223 10	0.843 71	0.808 27
8	8	18.6	0.377 99	0.628 83	0.165 56	0.348 94	0.831 35	0.700 13
7	7	16.4	0.369 97	0.636 71	0.182 23	0.376 79	0.814 37	0.676 20
6	6	14.2	0.424 22	0.583 43	0.233 45	0.394 83	0.762 20	0.660 69
5	5	12.0	0.444 33	0.563 69	0.271 59	0.445 18	0.723 34	0.617 42
4	4	9.8	0.475 74	0.532 85	0.321 31	0.319 71	0.672 70	0.725 25
3	3	7.6	0.455 37	0.552 85	0.338 22	0.262 40	0.655 47	0.774 50
2	2	5.4	0.556 81	0.453 24	0.452 84	0.375 71	0.538 71	0.677 12
1	1	3.2	0.582 88	0.427 64	0.517 03	0.676 59	0.473 33	0.418 56
0	0	1.0	1.018 38		0.981 69	1.221 69		

表 3 外部验证样本参考值和预测值结果

Table 3 Experimental values and predicted values for external validation samples

批次	参考值	预测值	相对预测 误差/%	批次	参考值	预测值	相对预测 误差/%	批次	参考值	预测值	相对预测 误差/%
181030	13.678 2	13.705 3	0.20	181102	12.037 7	12.089 1	0.43	181112	12.358 3	12.046 3	2.52
181031	12.716 7	12.631 8	0.67	181103	12.752 6	13.011 5	2.03	181113	12.655 9	12.378 2	2.19
181033	12.691 8	12.479 8	1.67	181104	12.394 1	12.825 9	3.48	181114	12.315 7	11.632 9	5.54
181035	12.070 0	11.300 3	6.38	181105	12.106 6	10.940 7	9.63	181117	12.641 9	12.648 1	0.05
181036	11.916 2	12.210 4	2.47	181108	12.289 4	12.478 6	1.54	181118	12.667 2	12.247 8	3.31
181038	12.000 4	12.386 2	3.21	181109	12.228 6	11.920 3	2.52	181121	12.627 3	12.716 5	0.71
181101	12.496 9	12.621 5	1.00	181111	12.242 3	11.723 9	4.23				

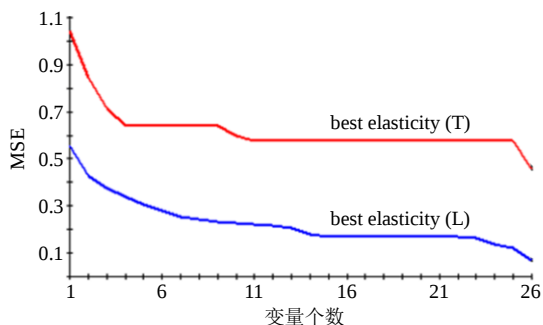


图 6 MSE 随弹性系数变化规律

Table 6 MSE changes with elasticity coefficient

表 4 不同变量个数建立的 GPS 模型性能和预测性能比较

Table 4 Comparison of GPS model performance and prediction performance established with different numbers of variables

自变 量数	校正集			预测集			平均相对预 测误差/%
	R^2_c	RMSE	MAPE	R^2_p	RMSE	MAPE	
54	0.738	0.486	0.030	0.611	0.675	0.040	4.63
4	0.566	0.633	0.004	0.599	0.745	0.045	4.07
3	0.514	0.622	0.042	0.596	0.748	0.046	3.05
2	0.402	0.297	0.018	0.196	0.345	0.021	1.82

3.2 不同算法模型性能和预测精度比较结果

综合对比分析以上不同算法建立的最优模型结

果如表 5 所示,可以发现 MARS 算法建立的预测模型在校正集和验证集上获得了较大的决定系数和较

表5 不同算法模型性能和预测精度比较

Table 5 Comparison of performance and prediction accuracy of different algorithm models

建模方法	自变量数	校正集			预测集			平均相对预测误差/%
		R^2_c	RMSE	MAPE	R^2_p	RMSE	MAPE	
PLS	5	0.616	0.678	—	0.615	0.408	—	3.69
CART	14	0.662	0.576	0.039	0.712	0.579	0.037	5.67
MARS	8	0.843	0.391	0.025	0.808	0.472	0.031	2.69
GPS	3	0.514	0.297	0.018	0.196	0.345	0.021	3.05

小的各项误差,模型的稳定性和预测精度更好。校正集和预测集的决定系数分别为 0.843、0.808, RMSE 分别为 0.391、0.472, MAPE 分别为 0.025、0.031, 验证样本的平均预测误差为 2.69%。

4 讨论

本研究运用不同算法对 GFC 内容物吸湿性进行建模,结果表明 MARS 算法建立的模型最好,模型平均相对预测误差为 2.69%。综合分析可知, MARS 算法优化了 PLS 算法需手动筛选关键变量的缺点,提高了模型运用的效率;标识了 PLS 算法和 CART 算法未显示的变量累加贡献值或变量之间的相互作用结果^[14,22-23];弥补了 GPS 算法不可处理非线性问题的缺陷。

本研究选用的 4 种算法都以特殊的方式来计算变量重要性, PLS 算法根据残差平方和的最小化选择变量^[24], CART 算法根据信息增益选择特征变量, GPS 算法利用正则化策略和弹性系数来选择变量, MARS 算法利用广义交叉原则和基函数来筛选变量。4 种算法筛选出的相同变量有 SKL- D_c 、 R_c - α , 表明 GFC 内容物吸湿性与软材细粉的休止角和湿颗粒的振实密度有关,与前期研究结果相同。休止角的大小是粉体粒子大小和粒径分布、粉体表面性质以及粒子间相互作用力的综合体现,休止角越大,吸湿性越强^[1]。软材细粉的休止角与制软材时浸膏的干燥粉碎方式、干燥工艺参数和浸膏的混合比例等有关,需收集更多的样本来进行实验研究确定影响吸湿性的显著因素。湿颗粒振实密度与颗粒的粒径分布、骨架密度和孔隙结构有关,因而影响颗粒吸湿性。前期研究结果显示湿法制粒过程中物料与黏合剂作用导致粉体表面黏性变化,湿颗粒振实密度与黏合剂加入量和浓度有关,提示生产过程中需及时调整黏合剂用量和浓度来保证颗粒的吸湿性处于一定的均匀范围内。湿颗粒振实密度还有可能与制粒过程中所加辅料的性质有关,尚需通过实验进

一步验证^[25-27]。

本研究所选数据量远少于生产过程产生的大批量数据,在后续研究中还需要继续收集更多的数据来进一步验证、更新、维护和优化模型,提高模型预测精确度,以此提升 GFC 生产过程质量控制智能化水平。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 王晴,徐冰,王芬,等. 桂枝茯苓胶囊内容物吸湿性预测建模研究 [J]. 中国中药杂志, 2020, 45(2): 242-249.
- [2] Han P J, Xue Z F, Zhang L N, et al. Moisture sorption and diffusion determination of Chinese herbal granules: Moisture-resistant effects of fluidized bed granulation with dextrin [J]. *Chin Herb Med*, 2018, 10(3): 290-297.
- [3] 吴晓倩,权丽丽,陈诚,等. 基于大数据决策树算法的学生成绩分析与预测模型仿真 [J]. 电子设计工程, 2020, 28(24): 138-141.
- [4] 仇文岗,洪利,黎泳钦. 基于多元自适应回归样条的高维岩土工程问题分析 [J]. 河海大学学报: 自然科学版, 2019, 47(4): 359-365.
- [5] 吴邦华,黄海莹,姚强,等. 大数据及人工智能方法在妊娠期糖尿病预测的应用 [J]. 中国卫生信息管理杂志, 2017, 14(6): 832-837.
- [6] 朱梦琦,朱合华,王昕,等. 基于集成 CART 算法的 TBM 掘进参数与围岩等级预测 [J]. 岩石力学与工程学报, 2020, 39(9): 1860-1871.
- [7] Ghiasi M M, Zendehboudi S, Mohsenipour A A. Decision tree-based diagnosis of coronary artery disease: CART model [J]. *Comput Methods Programs Biomed*, 2020, 192: 105400.
- [8] Abellán J, Mantas C J, Castellano J G. A random forest approach using imprecise probabilities [J]. *Knowl-Based Syst*, 2017, 134: 72-84.
- [9] 焦莉萍,郭晶晶,杨云云,等. 决策树模型与 Logistic 回归模型在生活饮用水水质影响因素分析中的应用 [J]. 中国卫生统计, 2020, 37(6): 874-877.
- [10] 唐容. 基于特征选择的 CART 算法研究 [D]. 成都: 电

- 子科技大学, 2020.
- [11] 仇文岗, 洪利, 黎泳钦. 基于多元自适应回归样条的高维岩土工程问题分析 [J]. 河海大学学报: 自然科学版, 2019, 47(4): 359-365.
- [12] Zhang W G. *MARS Applications in Geotechnical Engineering Systems* [M]. Singapore: Springer Singapore, 2020.
- [13] Tien Bui D, Hoang N D, Samui P. Spatial pattern analysis and prediction of forest fire using new machine learning approach of multivariate adaptive regression splines and differential flower pollination optimization: A case study at Lao Cai Province (Viet Nam) [J]. *J Environ Manage*, 2019, 237: 476-487.
- [14] 任美娜. 基于 MARS 的助听器语音质量评价方法研究 [D]. 北京: 华北电力大学 (北京), 2019.
- [15] 张艺潇, 赵忠国, 郑江华. 利用 MARS 估算伊犁地区参考作物蒸散量 [J/OL]. 武汉大学学报: 信息科学版, doi: 10.13203/j.whugis20190337.
- [16] Cheng M Y, Cao M T. Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines [J]. *Appl Soft Comput*, 2014, 22: 178-188.
- [17] Safari M J S. Decision tree (DT), generalized regression neural network (GR) and multivariate adaptive regression splines (MARS) models for sediment transport in sewer pipes [J]. *Water Sci Technol*, 2019, 79(6): 1113-1122.
- [18] 刘元铭, 王振华, 王涛, 等. 热轧带钢出口凸度数据驱动建模及智能化预测分析 [J]. 中国机械工程, 2020, 31(22): 2728-2733.
- [19] 郑伟达, 张惠然, 胡红青, 等. 基于不同机器学习算法的钙钛矿材料性能预测 [J]. 中国有色金属学报, 2019, 29(4): 803-809.
- [20] 刘帆, 李正文, 种洋. 一种基于 GCV 法改进的频域内位场正则化向下延拓的参数选取方法 [J]. 物探化探计算技术, 2018, 40(1): 73-81.
- [21] Gan M, Zhu H T, Chen G Y, et al. Weighted generalized cross-validation-based regularization for broad learning system [J]. *IEEE Trans Cybern*, 2020, doi: 10.1109/TCYB.2020.3015749.
- [22] 郭亚周, 刘小川, 白春玉, 等. 闭孔泡沫金属几种不同建模方法的对比性研究 [J]. 航空材料学报, 2020, 40(4): 85-91.
- [23] 韩斐斐. 不同建模数据抽样方法对兴安落叶松立木材积预测精度的影响 [D]. 哈尔滨: 东北林业大学, 2017.
- [24] 赵永胜. 基于改进的线性和非线性变量选择方法研究 [D]. 太原: 山西大学, 2020.
- [25] 赵舒婷, 杨白雪, 谢佳蓉, 等. 物料属性研究在中药丸剂质量控制中的应用进展 [J/OL]. 中国实验方剂学杂志: 1-8 [2021-05-07].
- [26] 赵立杰, 冯怡, 徐德生, 等. 中药制剂原料吸湿特性与其物理特性相关性研究 [A] // 加快转变医药发展方式, 占领科学技术制高点——2011 年中国药学会暨第 11 届中国药师周论文集 [C]. 烟台: 中国药学会, 2011: 2691-2698.
- [27] 祝明利, 张芳语, 张瀚, 等. 中药大品种制造关键质量属性表征: 粉末与颗粒物理属性的苏黄止咳胶囊中间体过程质量控制方法研究 [J]. 中国中药杂志, 2021, 46(7): 1636-1643.

[责任编辑 郑礼胜]