

• 药剂与工艺 •

基于决策树算法的热毒宁注射液金银花青蒿醇沉过程质量控制研究

丁 鸿^{1,2}, 徐芳芳^{2,3*}, 杜 慧², 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 中药信息学系, 北京 102400

摘要: 目的 基于决策树算法, 深入挖掘热毒宁注射液金银花青蒿醇沉过程(金青醇沉)数据并探究潜在生产规律, 提升该过程质量控制水平。方法 依托数字化中药提取工厂数据平台收集 205 批金银花和青蒿浸膏(金青浸膏)历史数据并整合成数据矩阵。将数据集随机划分为训练集和测试集后分别采用分类与回归树(classification and regression tree, CART)、随机森林(random forests, RF)和 TreeNet 算法建立金青醇沉过程模型, 比较各模型性能并基于历史数据划分关键变量控制范围。结果 RF 和 TreeNet 模型性能较好且性能接近, 综合各模型分析结果得出醇提罐料液比及金银花浓缩收率为重要的影响因素, 对重要变量进行依存度分析并优选批次, 并以优选批次的金银花分配浸膏质量及加醇量做控制图, 密度为 1.11 g/cm³ 的金银花浸膏分配控制范围为 557.92~639.62 kg, 加醇量的控制范围为 3.370~3.828 m³; 密度为 1.12 g/cm³ 的金银花浸膏的控制范围为 540.4~616.9 kg, 加醇量的控制范围为 3.317~3.859 m³。结论 决策树算法建立的金青浸膏醇沉过程模型能够有效地挖掘潜在的生产过程规律, 为生产过程的质量控制提升提供技术支撑。

关键词: 热毒宁注射液; 金银花; 青蒿; 醇沉过程; 决策树算法; 质量控制

中图分类号: R283.6 文献标志码: A 文章编号: 0253-2670(2021)19-5836-09

DOI: 10.7501/j.issn.0253-2670.2021.19.006

Research on quality control of alcohol precipitation process of *Lonicerae Japonicae Flos* and *Artemisiae Annuae Herba* for Reduning Injection based on decision tree

DING Hong^{1,2}, XU Fang-fang^{2,3}, DU Hui², 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. Department of Chinese Medicine Information Science, Beijing University of Chinese Medicine, Beijing 102400, China

Abstract: Objective Based on the decision tree algorithm, the data of Jinyinhua (*Lonicerae Japonicae Flos*) and Qinghao (*Artemisiae Annuae Herba*) (LA) alcohol precipitation process of Reduning Injection (RI, 热毒宁注射液) were deeply mined and the potential production rules was explored to improve the quality control level. **Methods** Based on the digital extraction factory of traditional Chinese medicine, 205 batches of RI extraction batch records and LA alcohol precipitation process parameters records were collected and integrated into a data matrix. After the data set randomly divided into training set and test set, classification and regression tree (CART), random forest (RF), and TreeNet were used to establish LA alcohol precipitation process models respectively. The performance of each model was compared, and the control range of key variables was divided based on historical data. **Results** RF

收稿日期: 2021-06-22

基金项目: 国家新药创制科技重大专项(2013ZX09402203)

作者简介: 丁 鸿, 硕士研究生, 研究方向为过程分析技术。E-mail: 20190351@njucm.edu.cn

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

and treenet models had good performance and close performance. Based on the analysis results of each model, it was concluded that the material liquid ratio of alcohol extraction tank and the concentration yield of honeysuckle were important factors. The dependence of important variables were analyzed and the batches were optimized, and the quality and alcohol amount of honeysuckle distribution extract of the optimal batch was taken as the control chart. The distribution control range of honeysuckle extract with density of 1.11 was 557.92—639.62 kg, and the control range of alcohol dosage was 3.370—3.828 m³; The control range of honeysuckle extract with density of 1.12 was 540.4—616.9 kg, and the control range of alcohol dosage was 3.317—3.859 m³. **Conclusion** The application of decision tree algorithm in alcohol precipitation of LA extract can effectively explore the potential production process rules, and provide technical support for the improvement of quality control in the production process.

Key words: Reduning Injection; *Lonicerae Japonicae Flos*; *Artemisiae Annuae Herba*; alcohol precipitation process; decision tree; quality control

热毒宁注射液是由金银花、青蒿和栀子 3 种中药提取精制而成^[1-2], 其前提取精制过程包括金银花提取浓缩、青蒿提取浓缩、栀子提取浓缩、金银花青蒿(金青)醇沉、金青萃取、栀子萃取等工序, 均在中药数字化提取精制工厂中完成^[3]。中药数字化提取工厂配备集散控制系统(distributed control system, DCS)、车间制造执行系统(manufacturing execution system, MES)等, 可实现对中药前处理设备传感器实时数据以及生产过程数据的抓取。在工业数据的挖掘与分析中, 机器学习是常用的方法之一^[4]。杜慧等^[5]以金银花和青蒿浓缩浸膏醇沉工序为研究对象, 采用偏最小二乘法(partial least squares, PLS)进行数据挖掘, PLS 需要对预测变量和目标变量进行降维投影处理^[6], 不利于直观解释生产规律, 该研究筛选了关键变量但未划分变量的控制范围, 不便于生产的实际调控。决策树算法是机器学习中一个树状预测模型, 适用于多阶段决策、表查找和最优化等问题, 能够直观解释决策过程, 目前已成功应用于医疗、广告、交通和金融等多个领域^[7-8]。本实验采用决策树及其相关算法对历史数据进行分析, 深入挖掘并寻找关键变量的优化控制范围, 期望提升热毒宁注射液的质量控制水平。

1 材料

建模数据由江苏康缘药业股份有限公司数字化提取工厂提供, 金青中间体批次为 170105 至 180905 共 205 批; 数据的训练集和测试集的划分应用 Python 的 Scikit-Learn 工具实现。建模算法均在 Salford Predictive Modeler 8.3 (美国 Minitab 公司) 实现, 单值过程控制图在 Minitab 19 (美国 Minitab 公司) 绘制。

2 方法

2.1 目标变量的确定

金青醇沉工段由金青浸膏分配、金青浸膏醇沉、

上清液暂存和上清液回收 4 个过程组成。金银花和青蒿的水提浓缩浸膏通过称量罐称定质量后分配至 2 个醇沉罐(醇沉罐 1 和醇沉罐 2), 在醇沉过程中, 向分配浸膏的 2 个醇沉罐中分 3 个阶段按不同的体积流量加入一定量的乙醇并不断搅拌, 搅拌停止静置一段时间并将醇沉上清液通过真空泵转入浓缩罐, 刮板浓缩后得金青醇沉浸膏。金青醇沉工段简图如图 1 所示, 各过程主要参数如表 1 所示。前期辨识出金青醇沉工段的关键因素为金青醇沉上清液和上清液传料体积、金青总质量、醇沉罐 1 及醇沉罐 2 金银花浸膏质量^[5]。上清液传料体积为上清液暂存罐传送至浓缩罐的上清液体积; 醇沉上清液为上清液暂存罐收集的上清液体积; 金青浸膏总质量为 2 个醇提罐加入金银花和青蒿总质量; 醇沉罐 1 和醇沉罐 2 金银花质量为分配至 2 个醇沉罐的金银花质量。所筛选的参数均为金青浸膏醇沉过程输入物料及输出物料的质量和体积参数, 醇沉过程工艺较复杂, 涉及影响因素较多, 为进一步挖掘潜在规律, 本研究选择金青浸膏醇沉过程为研究目标建立过程模型。

为消除过程物料质量体积相关性对模型的影响, 本研究拟以浸膏醇沉过程收率为目标变量。将金青浸膏醇沉收率和金青醇沉工段收率进行相关性分析, 相关图如图 2 所示, 相关系数(r)为 0.653。控制浸膏醇沉过程对提高金青醇沉工段的生产效率十分重要。综上, 本实验以金青浸膏醇沉收率作为目标变量, 建立浸膏醇沉过程模型。

2.2 建模数据准备

本实验建模数据主要来源于批记录及数字化提取工厂数据库集成数据。批记录为手动录入, 主要为原料药、中间体物料相关数据; 数据库收集的数据主要从 DCS 系统中收集各个设备传感器实时生产过程中抓取的时序数据, 抓取频率为每分钟 1 次。

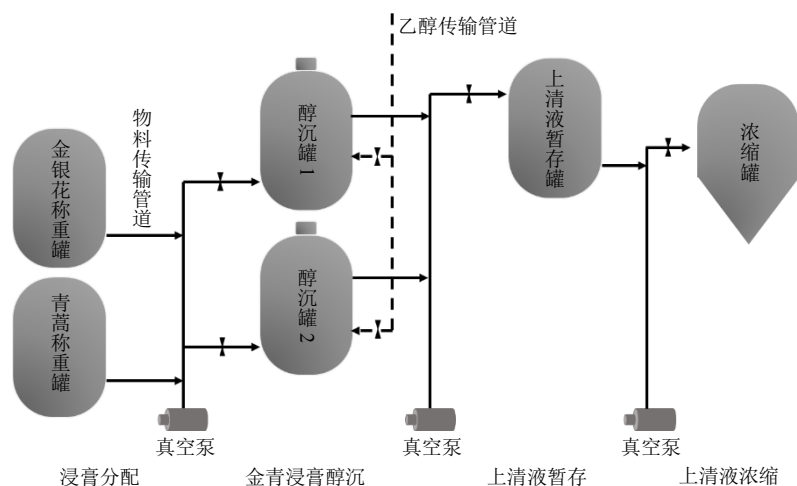


图 1 金青醇沉回收工段流程简化图

Fig. 1 Data map for LA alcohol precipitation process

表 1 金青醇沉 4 个过程的主要参数

Table 1 Main parameters of four processes of LA alcohol precipitation

工段	参数
金青浸膏分配	金银花浸膏质量、青蒿浸膏质量、分配上部温度、分配下部温度
金青浸膏醇沉	第 1 阶段加醇速度、第 2 阶段加醇速度、第 3 阶段加醇速度、加醇量、乙醇体积分数、乙醇温度
上清液暂存	金青醇沉上清液体积
醇沉液浓缩	金青上清液传料体积、单效回收时间、金青醇沉浸膏质量

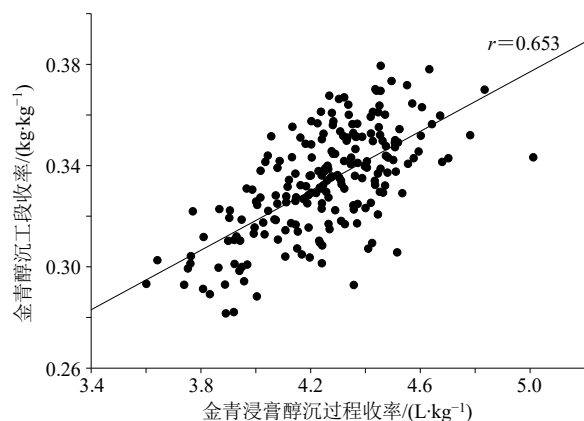


图 2 金青浸膏醇沉收率与金青醇沉工段收率相关图

Fig. 2 Correlation diagram between yield of LA extract alcohol precipitation and yield of LA alcohol precipitation section

中药生产环境复杂,不同时间点内时序参数相关性强,前期对 DCS 数据进行数据清洗和特征提取后得到建模特征参数^[6]。本实验在建模特征参数中选择 30 个醇沉罐传感器数据参数纳入建模数据集,考虑到浸膏质量和加醇体积间存在较强的相关性,本实验将金银花浸膏质量、青蒿浸膏质量和加醇量转化为金青浸膏质量比和加醇料液比。在中药制剂的过

程中,中药材的质量差异会传递至处方药味、中间体及成品,直接影响中药制剂批间质量的稳定^[9]。为考察药材及金银花浸膏及青蒿浸膏的各生产阶段可能对醇沉过程的影响,收集药材质量参数、药材厂商及药材至浓缩浸膏各工序的收率作为预测变量,样本收集时间为 2017 年 1 月—2018 年 9 月,共 205 批历史数据。以批号关联建模参数,建模参数共 49 个,最终得到 205×49 的建模数据集,建模变量如表 2 所示。

2.3 建模算法

决策树(decision tree)是一种基本的分类和回归方法,该学习方法主要包括特征选择、决策树生成和修剪,目前决策树学习常用的算法有 ID3、C4.5 和分类与回归树(classification and regression tree, CART)^[10]。当单一决策树学习器不能很好地处理问题时,可尝试集成学习方法(ensemble learning)提高模型预测性能。集成学习通过训练多个学习器并进行组合来解决同一问题^[11]。Bagging 算法和 Boosting 算法是 2 种经典的集成学习算法^[12]。本实验采用 CART 算法以及随机森林、TreeNet 2 种集成树算法进行建模。

表 2 金青浸膏醇沉过程 49 建模变量

Table 2 Forty-nine modeling variables in LA extract alcohol precipitation process

序号	参数	单位	序号	参数	单位
1	金银花药材厂商		26	罐 1 第 1 阶段 1~10 min 加醇平均速度	L·h ⁻¹
2	金银花绿原酸含量	%	27	罐 1 第 1 阶段 11~48 min 加醇平均速度	L·h ⁻¹
3	金银花药材水分	%	28	罐 1 第 2 阶段 1~10 min 加醇平均速度	L·h ⁻¹
4	金银花提取液收率	L·kg ⁻¹	29	罐 1 第 2 阶段 11~23 min 加醇平均速度	L·h ⁻¹
5	金银花提取液浓缩收率	kg·L ⁻¹	30	罐 1 第 2 阶段加醇阶段初速	L·h ⁻¹
6	青蒿药材供应商		31	罐 1 第 3 阶段 1~10 min 加醇平均速度	L·h ⁻¹
7	青蒿挥发油含量	%	32	罐 1 第 3 阶段 11~23 min 加醇平均速度	L·h ⁻¹
8	青蒿药材水分	%	33	罐 2 金青浸膏质量比	L·kg ⁻¹
9	挥发油转移率	%	34	罐 2 料液比	L·kg ⁻¹
10	青蒿提取率	L·kg ⁻¹	35	罐 2 浸膏分配上部初始温度	℃
11	青蒿三效回收率	kg·L ⁻¹	36	罐 2 浸膏分配上部结束温度	℃
12	青蒿醇沉收率	L·kg ⁻¹	37	罐 2 浸膏分配下部初始温度	℃
13	青蒿回收率	kg·L ⁻¹	38	罐 2 浸膏分配下部结束温度	℃
14	乙醇温度	℃	39	罐 2 浸膏分配前阶段上部平均温度	℃
15	乙醇浓度	%	40	罐 2 浸膏分配前阶段下部平均温度	℃
16	罐 1 料液比	L·kg ⁻¹	41	罐 2 浸膏分配后阶段上部平均温度	℃
17	罐 1 金青浸膏质量比		42	罐 2 浸膏分配后阶段下部平均温度	℃
18	罐 1 浸膏分配上部初始温度	℃	43	罐 2 第 1 阶段 1~10 min 加醇平均速度	L·h ⁻¹
19	罐 1 浸膏分配上部结束温度	℃	44	罐 2 第 1 阶段 11~48 min 加醇平均速度	L·h ⁻¹
20	罐 1 浸膏分配下部初始温度	℃	45	罐 2 第 2 阶段 1~10 min 加醇平均速度	L·h ⁻¹
21	罐 1 浸膏分配下部结束温度	℃	46	罐 2 第 2 阶段 11~23 min 加醇平均速度	L·h ⁻¹
22	罐 1 浸膏分配前阶段上部平均温度	℃	47	罐 2 第 2 阶段加醇阶段初速	L·h ⁻¹
23	罐 1 浸膏分配前阶段下部平均温度	℃	48	罐 2 第 3 阶段 1~10 min 加醇平均速度	L·h ⁻¹
24	罐 1 浸膏分配后阶段上部平均温度	℃	49	罐 2 第 3 阶段 11~23 min 加醇平均速度	L·h ⁻¹
25	罐 1 浸膏分配后阶段下部平均温度	℃			

2.3.1 CART CART 模型是应用广泛的决策树算法^[13],它是由特征选择、树的生成及剪枝组成,既可以用于分类,也可以用于回归^[10]。相较于 ID3, CART 算法选择基尼不纯度作为度量标准,并采用二分递归分割方法划分数据集,最终生成二叉决策树,二叉划分适用更加广泛,也更易处理数值型数据^[6]。

2.3.2 随机森林 (random forests, RF) RF 是由多个 Bagging 算法训练的决策树组合而成^[14],在解决回归问题时,其以多个决策树的平均值为输出结果^[12,15]。它是通过利用 bootstrap 重抽样方法从原始样本中抽取多个样本,没有抽到的样本被称为带外数据 (out-of-bag, OOB)^[16]。OOB 误差估计是 RF 算法的一个无偏估计,可替代数据集的交叉验证,明显降低计算复杂性^[12,17]。

2.3.3 TreeNet TreeNet 被称为随机梯度提升 (stochastic gradient boosting)^[18]。梯度提升是一种实现 Boosting 的方法,其主要思想是每次建模都在前一模型损失函数 (loss function) 梯度下降方向进行^[19]。TreeNet 在训练的过程中首先生成 1 棵小树并计算该模型的预测残差,之后建立第 2 棵树对第 1 棵树的残差进行预测,第 3 棵树对前 2 棵树的残差进行预测。上步骤递归,新产生的小树对前面的模型进行不断修正,最终得到最佳模型。

2.4 数据集划分与训练

数据按 3:1 的比例随机划分为训练集和测试集。通过训练集的数据来训练数据,用测试集来评估模型。在训练模型的过程中需要训练不同模型参数的模型来筛选最佳模型,将保留训练集的一部分来评估各个模型的性能,新的保留集称为验证集。

本实验采用 10 折交叉验证来评估 CART 和 TreeNet 的训练模型, RF 为 OOB 误差估计。本实验以均方根误差 (root mean square error, RMSE)、平均绝对偏差 (mean absolute difference, MAD) 以及决定系数 (R^2) 作为模型评价指标。RMSE 及 MAD 越接近 0, R^2 越接近 1 表示模型误差越小, 性能越好。

$$\text{RMSE} = [1/n \sum_{i=1}^n (y - y_i)^2]^{1/2}$$

$$\text{MAD} = 1/n \sum_{i=1}^n |y - y_i|$$

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

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

3 结果与分析

3.1 各模型性能评价

基于训练集数据分别建立 CART、RF 和 TreeNet 模型, 并采用 10 折交叉验证来评估不同超参数的 CART 和 TreeNet 训练模型性能, OOB 误差估计评估 RF 模型性能。最终得到各模型最佳的超参数设置如表 3 所示, 最优模型的性能评价指标如表 4 所示, 各模型的预测变量重要性如表 5 所示。由表 5 中可知, 与 CART 算法相比, RF 和 TreeNet 2 种集成树算法的验证集和预测集均方根更小, 决定系数更大, 模型效果更好。说明集成学习算法有效地提高了单一决策树算法模型的性能。

由表 5 可知, 3 种模型的共同重要变量为 X_{16} (罐 1 料液比)、 X_{34} (罐 2 料液比) 和 X_5 (金银花提

表 3 3 种模型的超参数设置

Table 3 Hyperparameter setting table of three models

模型	模型超参数
CART	最小父结点的值设置为 10, 最小子结点大小为 1, 缺失值惩罚设置为 1, 最大结点数为 32, 其他设置均为默认
RF	预测变量数设置为 11, 生成树的棵树为 200, 缺失值惩罚设置为 1, 其他设置均为默认
TreeNet	学习率设置为 0.1, 生成树的棵树为 200, 每棵树的最大结点为 6, 树的最小结点为 1, 缺失值惩罚设置为 1, 其他设置均为默认

表 4 3 种最优模型的性能指标评价

Table 4 Performance index evaluation of three models

模型	训练集			验证集			预测集		
	RMSE	MAD	R^2	RMSE	MAD	R^2	RMSE	MAD	R^2
CART	0.109 3	0.085 4	0.775	0.134 1	0.107 3	0.662	0.131 6	0.093 2	0.664
RF	—	—	—	0.115 1	0.086 9	0.751	0.105 6	0.080 2	0.783
TreeNet	0.079 0	0.047 6	0.883	0.112 6	0.082 1	0.757	0.112 2	0.081 6	0.755

RF 为 OOB 检验, “—” 代表无训练集, 表 6 同

RF be estimated by OOB, “—” means no training set, same as table 6

表 5 CART、RF 和 TreeNet 模型变量重要性

Table 5 Importance chart of predicted table of CART, RF and TreeNet

算法	预测变量	变量重要性	算法	预测变量	变量重要性	算法	预测变量	变量重要性
RF	X_{16}	94	RF	X_{14}	2	CART	X_{33}	43.13
	X_{34}	29		X_{49}	2		X_{38}	34.45
	X_5	13		X_{47}	1		X_{49}	33.95
	X_{21}	4		X_{23}	1	TreeNet	X_{16}	100.00
	X_{11}	4	CART	X_{33}	1		X_{34}	94.97
	X_{40}	3		X_{16}	100.00		X_4	87.91
	X_{42}	3		X_5	92.21		X_5	84.23
	X_{48}	3		X_{34}	90.95		X_{40}	82.98
	X_{38}	3		X_{17}	64.41		X_1	82.19

取液浓缩收率)。以这 3 个变量建立模型,其模型效果如表 6 所示。

由变量筛选后模型性能观察比较可知,变量筛选后测试集和预测集的模型误差及 R^2 变化较小,均保持较好的模型性能。说明醇沉罐 1 料液比、醇沉罐 2 料液比以及金银花提取液浓缩收率为金青浸膏

醇沉过程的关键变量。

3.2 关键变量依存性分析

为进一步描述 3 个关键变量与目标变量的潜在规律。对 TreeNet 模型的 3 个关键变量的变量依存度图 (partial dependency plots, PDPs) 进行分析,单变量依存性散点图如图 3 所示。

表 6 筛选后 3 种模型的性能指标评价

Table 6 Performance index evaluation of three models after screening

模型	训练集			验证集			预测集		
	RMSE	MAD	R^2	RMSE	MAD	R^2	RMSE	MAD	R^2
CART	0.094 4	0.071 3	0.832	0.133 2	0.100 7	0.666	0.142 6	0.111 3	0.605
RF	—	—	—	0.130 3	0.098 3	0.681	0.115 4	0.086 4	0.742
TreeNet	0.110 5	0.075 6	0.770	0.129 8	0.097 9	0.675	0.124 3	0.092 6	0.704

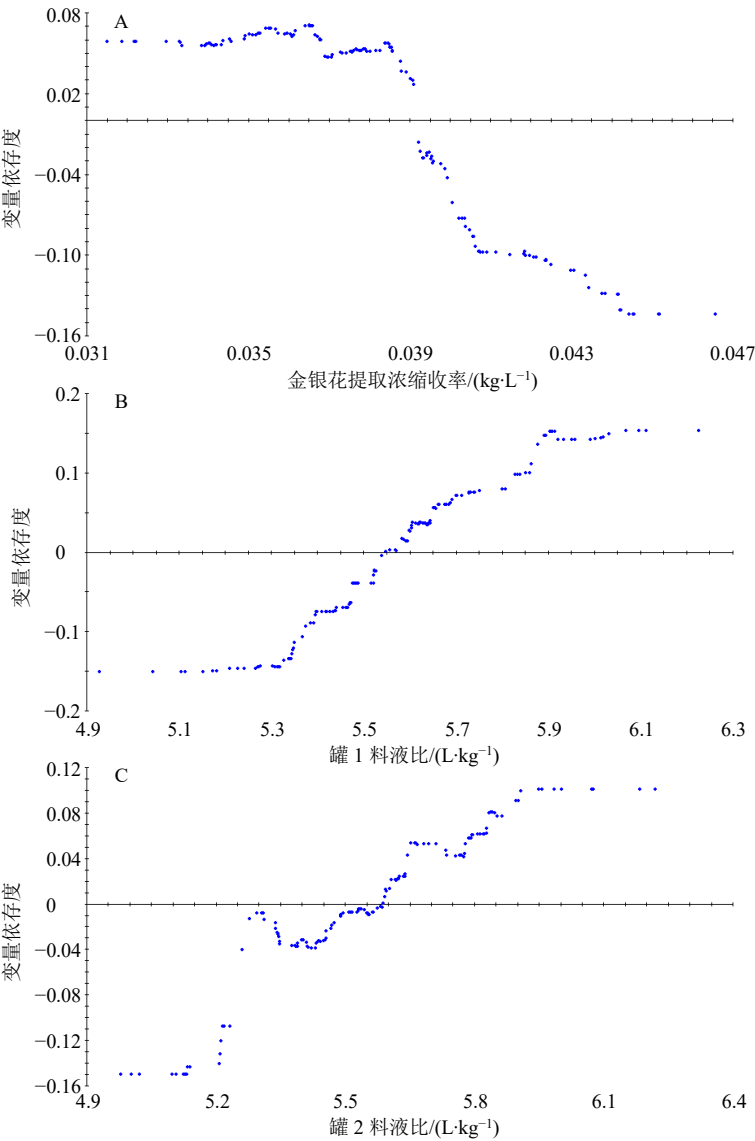


图 3 关键变量为金银花提取液浓缩收率 (A)、罐 1 料液比 (B) 及罐 2 料液比 (C) 单变量依存度散点图
Fig. 3 PDPs of univariate dependency with key variable were concentration yield of honeysuckle extract liquid (A) and solid-liquid ratios in extraction tank 1 (B) and that in tank 2 (C)

通过对 3 个关键变量的 PDPs 分析, 各变量与其部分依存度不是简单的线性关系。变量依存度分析以响应值的正负表示对转化率影响的好坏, 将工艺优化问题转化为计算问题。当金银花提取液浓缩收率小于 0.039 kg/L 时, 目标变量预测响应值为正值。醇沉罐 1 和醇沉罐 2 料液比的依存度随料液比的不断增加的变化可分为 3 个范围: 料液比较低, 当料液比较小时依存度为负值, 较小的料液比将降低目标变量的预测响应值; 料液比居中, 随着料液比的不断增加依存度也逐渐提高并转为正值; 料液比较高, 当增加到一定范围时, 料液比的增加对目标变量预测响应值的影响不变。将料液比控制在合

适区间, 在提高金青浸膏醇沉的收率同时可避免加入过量乙醇, 降低生产成本。

模型结果中醇沉罐 1 的料液比控制范围大约为 5.54~5.91 L/kg, 醇沉罐 2 的范围大约为 5.58~5.91 L/kg。对 2 罐的范围进行中位数检验, 分析 2 罐间是否存在差异, 以便于确定是否需要分罐控制。结果如表 7 所示。

醇沉罐 1 和醇沉罐 2 相比, P 值 >0.05 , 总体中位数相等, 醇沉罐 1 和醇沉罐 2 间料液比优化范围无显著差异。综上, 醇沉罐 1 和醇沉罐 2 可统一控制, 为控制金青浸膏醇沉过程的收率, 需主要对金银花提取液浓缩收率和各罐料液比进行控制。

表 7 不同提取罐料液比中位数

Table 7 Median of solid-liquid ratios in different extraction tanks

罐号	中位数	$N \leq$ 整体中位数	$N >$ 整体中位数	$Q_3 - Q_1$	95% 中位数置信区间
1	5.657 4	54	44	0.144 7	(5.637 5, 5.694 6)
2	5.709 2	36	45	0.177 2	(5.664 1, 5.771 2)
整体	5.678 8	—	—	—	—

N 表示观测值个数, $Q_3 - Q_1$ 表示四分位间距, Q_3 表示第 75 个百分点, Q_1 表示第 25 个百分点

N represents the number of observed values, $Q_3 - Q_1$ represents the interquartile range, Q_3 represents the 75th percentile, and Q_1 represents the 25th percentile

3.3 关键变量过程控制

上文筛选影响金青浸膏醇沉的关键变量为金银花提取液浓缩收率和料液比, 金银花醇沉上清液浓缩收率的影响因素较多, 涉及上一工段各工艺参数及设备过程参数, 需另外建模研究。本实验主要对料液比的过程控制进行研究。

控制料液比主要控制金银花浸膏、青蒿浸膏和加醇量。通过收集落入优化范围内的金银花浸膏、青蒿浸膏和加醇量, 进行中位数检验分析差异。结果如表 8 所示。统计分析表明, 影响料液比的变化主要是由于金银花浸膏质量和加醇量的波动引起的。企业内部控制标准中要求金银花浸膏浓缩至 1.10~1.12 g/cm³ (70~80 °C), 并分别对密度为

1.10、1.11、1.12 g/cm³ 的浸膏的出膏量给出不同的控制范围。依存度分析优化范围内的金银花浸膏出膏温度均符合要求, 浸膏密度分为 1.11、1.12 g/cm³。分别对密度为 1.11、1.12 g/cm³ 的金银花浸膏做控制图, 控制图如图 4 所示。

如控制图所示, 通过筛选得到密度为 1.11 的金银花浸膏分配控制范围为 557.92~639.62 kg, 加醇量的控制范围为 3.370~3.828 m³; 密度为 1.12 的金银花浸膏的控制范围为 540.4~616.9 kg, 加醇量的控制范围为 3.317~3.859 m³。综上分析, 优化范围内加醇量的差异不大, 金银花浸膏质量的波动是影响金青醇沉过程的重要因素。需加强对金银花浸膏质量的监控手段和方法, 将金银花浸膏质量控制在稳定的优化范围内控制加醇量, 从而控制醇沉过程高效稳定进行。

4 总结与展望

醇沉工序是中药提取精制过程的关键工序之一, 在实际生产控制过程中对金青醇沉过程控制的稳定性还有待提高。本实验依托热毒宁提取精制生产大数据, 收集金青醇沉过程历史数据, 以金青浸膏醇沉上清液收率为目标变量, 基于决策树算法进行建模分析, 综合 3 个模型建模结果, 筛选出重要变量为金银花上清液浓缩收率和醇沉罐料液比; 通

表 8 不同料液比的相关变量中位数

Table 8 Median of related variables for different solid-liquid ratios

参数	中位数		
	范围内料液比	高料液比	低料液比
金银花浸膏质量	601	563**	660**
青蒿浸膏质量	31	30.75	33
加醇量	3.617	3.570 5	3.726**

与范围内料液比比较: ** $P < 0.01$

** $P < 0.01$ vs solid-liquid ratios within the range

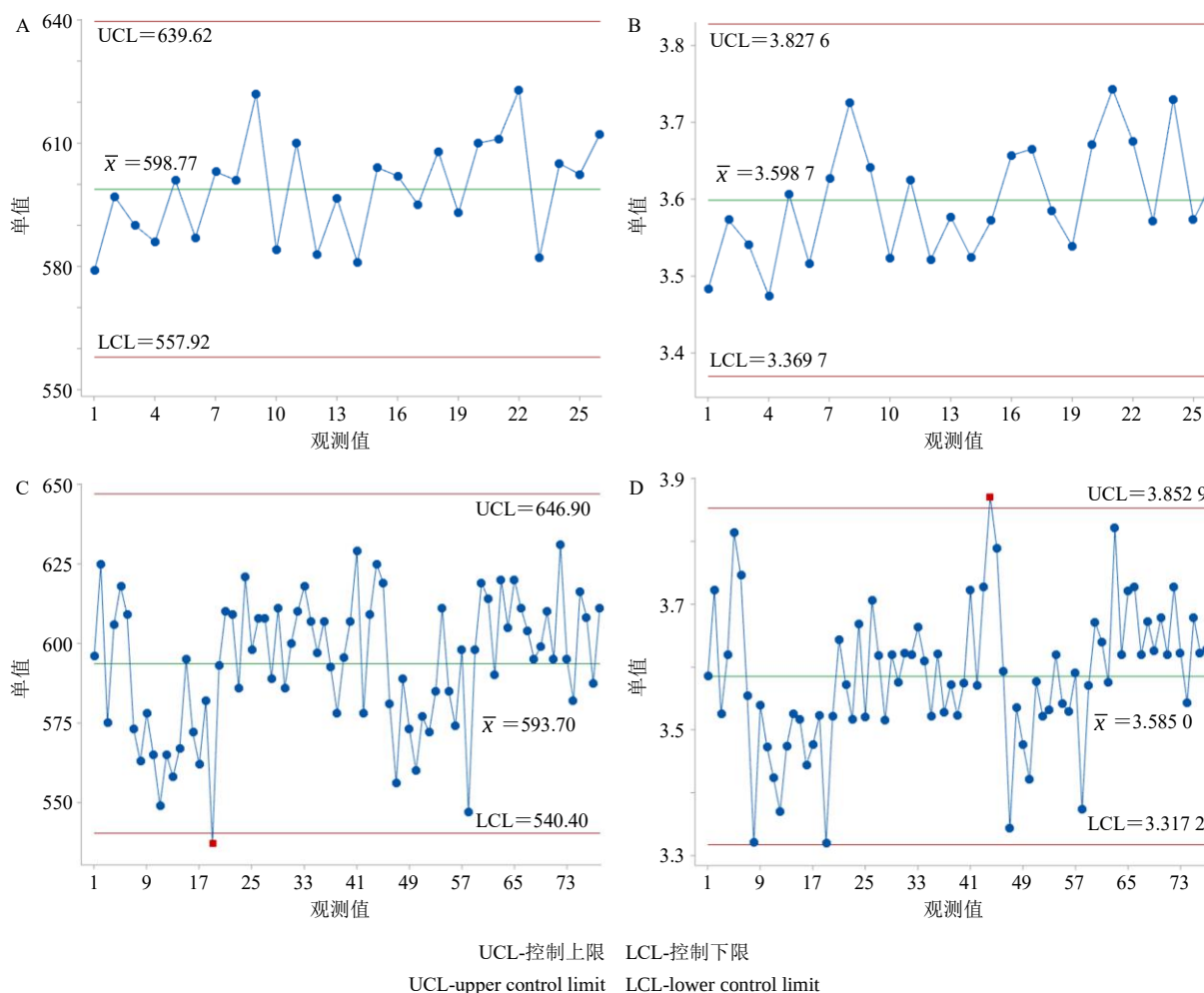


图 4 密度为 $1.11 \text{ g}\cdot\text{cm}^{-3}$ 金银花浸膏质量 (A) 与加醇量 (B) 及密度为 $1.12 \text{ g}\cdot\text{cm}^{-3}$ 金银花浸膏质量 (C) 与加醇量 (D) 的过程控制图

Fig. 4 Process control diagram of quality of honeysuckle extract with a density of $1.11 \text{ g}\cdot\text{cm}^{-3}$ (A), amount of alcohol (B), quality of honeysuckle extract with a density of $1.12 \text{ g}\cdot\text{cm}^{-3}$ (C), and amount of alcohol (D)

过变量依存度分析, 优选批次并作金银花质量及加醇量过程控制图优化控制范围。从热毒宁注射液工业数据中挖掘质量传递规律, 有利于该产品质量的追溯与控制, 促进制造过程的优化及持续改进^[17]。

决策树算法在中药工业大数据的建模应用报道较少, 本实验首次尝试应用于热毒宁注射液的醇沉过程, 相较于 PLS 算法, 决策树算法在缺失值处理、变量筛选及优化参数范围的划分等方面更具优势, 为热毒宁注射液的数据挖掘和质量控制提供新路径。本实验建模数据为生产历史数据, 在后续的工作中, 还需结合近红外快速检测技术收集中间体含量数据, 多目标决策优选工艺参数控制范围。此外, 还需持续收集生产数据, 不断优化及验证工艺参数控制范围, 提升金青醇沉过程的控制水平。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 徐芳芳, 杜慧, 张欣, 等. 在线中红外光谱监测热毒宁注射液金银花与青蒿醇沉过程 7 种指标成分研究 [J]. 中草药, 2021, 52(10): 2909-2917.
- [2] 王永香, 郑伟然, 米慧娟, 等. 热毒宁注射液青蒿金银花浓缩过程近红外快速定量检测方法的建立 [J]. 中草药, 2017, 48(1): 102-108.
- [3] 辛国斌. 智能制造探索与实践 [M]. 北京: 电子工业出版社, 2016: 150-161.
- [4] Ge Z Q, Song Z H, Ding S X, *et al.* Data mining and analytics in the process industry: The role of machine learning [J]. *IEEE Access*, 2017, 5: 20590-20616.
- [5] 杜慧, 徐冰, 徐芳芳, 等. 大数据驱动的热毒宁注射液金青醇沉关键工艺参数辨识研究 [J]. 中国中药杂志, 2020, 45(2): 233-241.
- [6] 张曦. 基于统计理论的工业过程综合性能监控、诊断及质量预测方法研究 [D]. 上海: 上海交通大学, 2008.

- [7] 张桢, 曹健. 面向大数据分析的决策树算法 [J]. 计算机科学, 2016, 43(S1): 374-379.
- [8] 陈琪, 徐芳芳, 张欣, 等. 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 [J]. 中草药, 2021, 52(11): 3216-3223.
- [9] 国家药品监督管理局药品评审中心. 中药均一化研究技术指导原则(试行). [2021-05-30]. <http://www.cde.org.cn/attachmentout.do?mothed=list&id=d44cbb17b0387bb7>, 2020.
- [10] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012: 80.
- [11] 连克强. 基于 Boosting 的集成树算法研究与分析 [D]. 北京: 中国地质大学(北京), 2018.
- [12] 王奕森, 夏树涛. 集成学习之随机森林算法综述 [J]. 信息通信技术, 2018, 12(1): 49-55.
- [13] Gordon A D, Breiman L, Friedman J H, *et al.* Classification and regression trees [J]. *Biometrics*, 1984, 40(3): 874.
- [14] Breiman L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5-32.
- [15] 马骊. 随机森林算法的优化改进研究 [D]. 广州: 暨南大学, 2016.
- [16] Breiman L. Out-of-bag estimation, CA 94708 [R]. Technical Report, Department of Statistics, University of California, Berkeley, 1996.
- [17] 张春霞, 郭高. Out-of-bag 样本的应用研究 [J]. 软件, 2011, 32(3): 1-4.
- [18] Friedman J H. Stochastic gradient boosting [J]. *Comput Stat Data An*, 2002, 38(4): 367-378.
- [19] 姚东. 基于流的大规模网络安全态势感知关键技术研究 [D]. 郑州: 解放军信息工程大学, 2013.

[责任编辑 郑礼胜]