网络首发时间：2020-02-28 14:47:59
网络首发地址：http://kns.cnki.net/kcms/detail/11.2272.r.20200226.1154.005.html

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

杜慧[1,3]，徐冰[2*]，徐芳芳[3,4,5,6]，张欣[3,4,5,6]，王晴[1,3]，夏春燕[1,3]，包乐伟[3,4,5,6]，王振中[3,4,5,6]，乔延江[2]，肖伟[1,3,4,5,6*]

（1. 南京中医药大学，江苏 南京 210023；2. 北京中医药大学 中药信息学系，北京 102400；3. 江苏康缘药业股份有限公司，江苏 连云港 222001；4. 中药制药过程新技术国家重点实验室，江苏 连云港 222001；5. 中成药智能制造国家地方联合工程研究中心，江苏 连云港 222001；6. 中药提取精制新技术重点研究室，江苏 连云港 222001）

[摘要] 金青醇沉是热毒宁注射液关键工艺单元之一，具有工艺参数多样、过程机制复杂的特点。为辨识影响金青醇沉过程的关键工艺参数，该文以热毒宁注射液数字化工厂为基础，采集热毒宁注射液金青醇沉工段 2017—2018 年的历史生产数据 259 批，共计 829 318 数据点，呈现出数据量大、价值密度低、来源多样等大数据部分特征。以金青醇沉浓缩制得浸膏质量为响应变量，通过数据清洗和特征提取，数据点减少为 9 936 个。采用 Pearson 相关分析和灰色关联度分析进行综合决策，从 48 个特征变量中筛选出 15 个潜在关键工艺参数（*p*CPPs）。进一步通过偏最小二乘（PLS）回归进行定量预测建模，证明基于 15 个 *p*CPPs 建立的预测模型与基于 48 个特征变量的建立的预测模型性能相当。通过变量重要性排序，辨识出影响金青醇沉浓缩浸膏质量的 9 个关键工艺参数（CPPs），包括 4 个初始输入浸膏质量参数、3 个加醇量参数和 2 个醇沉上清液体积参数，至此数据点为 1 863 个，占原始数据的 0.28%。从全局数据出发，采用大数据分析的方法可有效提高数据的价值密度，筛选得到的关键工艺参数有助于解析金青醇沉生产过程质量传递规律。

[关键词] 金青醇沉；热毒宁注射液；大数据；关键工艺参数；质量传递规律

Identification of critical process parameters of Jinqing alcohol precipitation of Reduning Injection by big data

DU Hui[1,3], XU Bing[2*], XU Fang-fang[3,4,5,6], ZHANG Xin[3,4,5,6], WANG Qing[1,3], XIA Chun-yan[1,3], BAO Le-wei[3,4,5,6], WANG Zhen-zhong[3,4,5,6], QIAO Yan-jiang[2], XIAO Wei[1,3,4,5,6*]

(1. *Nanjing University of Chinese Medicine*, *Nanjing* 210023, *China*; 2. *Department of Chinese Medicine Information Science*, *Beijing University of Chinese Medicine*, *Beijing* 102400, *China*; 3. *Jangsu Kanion Pharmaceutical Co., Ltd.*, *Lianyungang* 222001, *China*; 4. *State Key Laboratory of New-tech for Chinese Medicine Pharmaceutical Process*, *Lianyungang* 222001, *China*; 5. *National & Local Joint Engineering Research Center on Intelligent Manufacturing of Traditional Chinese Medicine*, *Lianyungang* 222001, *China*; 6. *Key Laboratory of New Technology for Extraction and Refining of Traditional Chinese Medicine*, *Lianyungang* 222001, *China*)

[Abstract] Lonicerae Japonicae Flos and Artemisiae Annuae Herba (LA or Jinqing) alcohol precipitation has various process pa-

[收稿日期] 2019-09-01
[基金项目] 国家“重大新药创新”科技重大专项（2018ZX09201010）；国家工信部智能制造综合标准化与新模式应用项目（KYYY20170820）
[通信作者] * 肖伟，博士，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，Tel：（0518）81152367，E-mail：kanionlunwen@163.com；* 徐冰，副教授，硕士生导师，研究方向为中药质量和先进工艺控制，E-mail：xubing@bucm.edu.cn
[作者简介] 杜慧，硕士研究生，E-mail：2943586584@qq.com

rameters and complex process mechanism, and is one of the key units for manufacturing Reduning Injection. In order to identify the critical process parameters (CPPs) affecting the weight of the extract produced from the alcohol precipitation process, 259 batches of historical production data from 2017 to 2018 were collected, with a total of 829 318 data points. These data showed characteristics of large data, such as a large data volume, a low value density, and diverse sources. The data cleaning and feature extraction were first performed, and 48 feature variables were selected. The original data points were reduced to 9 936. Then, a combination of Pearson correlation analysis and grey correlation analysis were used to screen out 15 potential critical process parameters (*p*CPPs). After that, the partial least squares (PLS) was used in prediction of the weight of the extract, proving that the performance of predictive model based on 15 *p*CMAs is equivalent to that of predictive model based on 48 feature variables. The variable importance in projection (VIP) index was used to identify 9 CPPs, including 2 alcohol precipitation supernatant volume parameters, 4 initial extract weight parameters and 3 added alcohol volume parameters. As a result, the number of data points was 1 863, accounting for 0.28% of the original data. The big data analysis approach from a holistic point of view can effectively increase the value density of the original data. The critical process parameters obtained can help to accurately describe the quality transfer mechanism of the Jinqing alcohol precipitation process.

[Key words] Jinqing alcohol precipitation; Reduning Injection; big data; critical process parameter; quality transfer mechanism

doi: 10.19540/j.cnki.cjcmm.20191219.301

热毒宁注射液由金银花、青蒿和栀子 3 味中药提取精制而成，其提取、浓缩、醇沉、萃取、干燥等前处理工艺在中药数字化提取精制工厂完成[1]。在热毒宁注射液生产过程，提取、醇沉和萃取是对终产品物质基础影响较大的 3 个工段。其中金青醇沉过程受药液密度、药液温度、乙醇浓度、加醇操作方式和速度、乙醇用量、最终药液乙醇浓度、醇沉时间等多种因素的影响，过程监控的质量指标呈现出复杂的波动规律。对金青醇沉工段的工艺参数和质量参数进行关联分析，明确生产过程中金青醇沉工艺参数对质量指标的影响，辨识金青醇沉过程的关键物料属性(critical material attributes，CMAs)和关键工艺参数(critical process parameters，CPPs)，是指导金青醇沉过程关键质量属性(critical quality attributes，CQAs)在线监控、预测性调控和质量持续改进的前提，也符合 ICH 质量源于设计(quality by design，QbD)和产品生命周期管理(product lifecycle management，PLM)的要求[2-4]。

目前，中药制药过程 CMAs 和 CPPs 的筛选主要依赖经验评估，如失败模式和效应分析(FMEA)等风险评估法[5]；或在实验室规模实验或生产过程数据基础上，进行知识组织[6]、回归分析[7]、灰色关联分析[8]等，通过回归系数、灰色关联度等指标评价变量的关键性。在上述报道中，涉及的供筛选的潜在关键物料属性(potential critical quality attributes，*p*CMAs)和潜在关键工艺参数(potential critical process parameters，*p*CPPs)数目有限(≤16 个)[6-10]。然而金青醇沉生产过程包含多个工序，涉及参数众多，且在线仪表可记录醇沉过程温度、流速等参数随时间的变化，已具备了工业大数据(big data)的部分特征。本文结合大数据分析思路和金青醇沉的生产操作，综合采用数据地图、数据整理、数据清洗、特征提取、以及过程建模等方法，从生产大数据中辨识金青醇沉过程的关键工艺参数，理解过程质量传递规律，为实施智能调控奠定基础。

1 数据来源和数据组织

在热毒宁注射液生产过程中，金银花和青蒿药材分别经提取、浓缩后，将浓缩液按一定比例合并为金青浸膏，再进行醇沉、萃取和干燥等生产处理，制得干浸膏中间体[10]。在金青醇沉过程中，首先由集散控制系统(distributed control system，DCS)调出金青醇沉程序，依据生产指令，将金青浸膏分配至 2 个相同规格的醇沉罐中，然后分别加入乙醇、不断搅拌至达到规定的含醇量，继续搅拌一段时间后，关闭搅拌桨，静置；静置结束后，将金青醇沉上清液吸入上清液暂存罐，随后转移至浓缩器；浓缩后称重得到醇沉浓缩浸膏质量。热毒宁注射液金青醇沉工段数据地图见图 1。其数据主要源自 2 个部分：第一部分为生产批记录，即通过密度计、温度计、乙醇计等工具测量后人工录入的数据，这些数据包括浸膏密度、浸膏温度、乙醇温度、乙醇浓度、单效回收时间、出膏密度、出膏温度；第二部分由生产设备上的检测仪表生成，传入 DCS 中，实时数据库软件通过 OPC(OLE for process control)接口从 DCS 中采集数据并存储，

这些数据包括出膏量、金银花质量、青蒿质量、分配温度、加醇流速、加醇量、醇沉上清液体积、醇沉上清液传料体积。本文收集了金青醇沉工段在 2017—2018 年的 259 批生产数据。

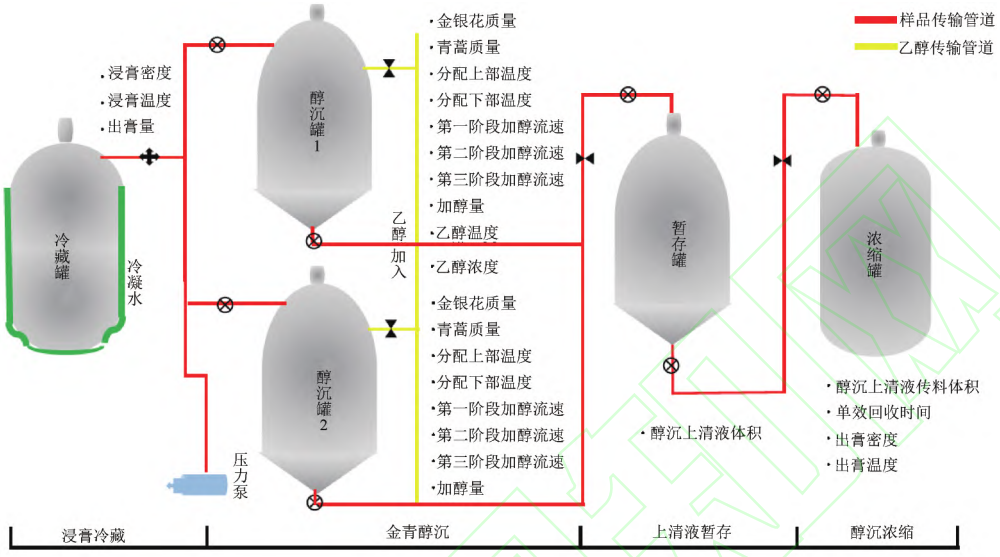


图 1　金青醇沉工段数据地图

Fig. 1　Data map for the Jinqing alcohol precipitation process

在每个醇沉罐的上部和下部分别安装了温度仪表，可分别记录浸膏分配至醇沉罐过程中的上部和下部温度随时间变化的数据曲线，记录频率为每分钟 1 数据点。加醇过程分为 3 个阶段，流量仪表可记录每个阶段加醇速度随时间变化的数据曲线。在生产过程中，不同批次浸膏分配和加醇操作的时间存在差异，导致时序参数的数据点数不同。如经统计，不同批次完成浸膏分配所需时间为 749 ~ 900 min，因此不同批次分配温度数据点为 749 ~ 900。为保证数据整齐，利于建模研究，对于温度仪表采集的数据，截取前 749 个温度数据点；对于加醇流速数据，第一阶段加醇速度收集 48 个数据点、第二阶段加醇速度收集 23 个数据点及第三阶段加醇速度收集 23 个数据点。在 1 个生产批次中，时序参数共有(749+749) ×2+(48+23+23) ×2 = 3 184 个数据点；非时序参数(如浸膏密度、浸膏温度、浸膏量、金银花质量等) 共 18 个数据点，总计 3 202 个数据点。259 批总计获取 829 318 个数据点。

2　原理与方法

2.1　变量相关性分析

2.1.1　Pearson 相关系数　Pearson 相关系数用于定量衡量 2 个变量 x 和 y 之间紧密联系程度[11]，当 2 个变量都是正态连续变量，且两者之间呈线性关系时，可用此法来判断过程输入变量与输出变量之间的相关程度，辅助参数的初步分析和筛选，其原理如下。

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) \tag{1}$$

其中 $\frac{x_i-\bar{x}}{\sigma_x}$ 表示对 x_i 样本的均值标准化处理，$\bar{x}$ 和 σ_x 分别是所有 x 变量的平均值和标准差。相关系数的取值范围在(−1,1)，绝对值越大，两变量之间的相关性程度就越强。一般认为绝对值大于 0.7 为高度相关，0.4~0.7 为中等相关，0.2~0.4 为低度相关，绝对值小于 0.2 为极弱相关或不相关。

2.1.2　灰色关联度　灰色关联分析是根据变量变化曲线的几何形状，对序列参数进行分析，从而确定哪些参数起主导作用[12]，其原理如下。

$$\xi_j = \frac{\min_i\min_j|y_{0j}-x_{ij}| + \rho\times\max_i\max_j|y_{0j}-x_{ij}|}{|y_{0j}-x_{ij}| + \rho\times\max_i\max_j|y_{0j}-x_{ij}|} \tag{2}$$

其中 ξ_j 表示第 i 个子序列的第 j 个参数与母序列(即 0 序列) 的第 j 个参数的关联系数，$|y_{0j}-x_{ij}|$ 表示序列 y_{0j} 与 x_i 在 j 点差值的绝对值；$\min_i\min_j|y_{0j}-x_{ij}|$ 表示差值绝对值的两级最小值；$\max_i\max_j|y_{0j}-x_{ij}|$ 表示差值绝对值的两级最大值；ρ 为分辨

系数取值范围在 [0,1]，其取值越小求得的关联系数之间的差异性越显著，一般取 0.5。关联系数的算术平均数为关联度。该方法适用于小样本灰色系统，即因素之间的关系是灰色的、难以定量区分密切程度。

2.2 偏最小二乘

偏最小二乘（partial least squares，PLS）集中了主成分分析、典型相关分析法和多元线性回归为一体的线性模型。在 PLS 模型中，参与建模的参数对模型重要性可用变量投影的重要性（variable importance in the projection，VIP）指标表示。

$$VIP_i = \sqrt{k\sum_{h=1}^{m} R_d(Y;t_h)\,W_{h_i}^2 \Big/ \sum_{h=1}^{m} R_d(Y;t_h)} \quad (3)$$

式中，k 为自变量的个数；$W_{h_i}^2$ 为自变量在主成分上的权重；t_h 为矩阵 T 的第 h 列分量；$R_d(Y;t_h)$ 为第 h 个主元对 Y 的解释能力）表示，VIP_i 越大，i 变量对 Y 预测性的贡献越大[13]。

2.3 统计分析软件

Pearson 相关系数、灰色关联度分析通过 MATLAB 2009a（美国 MathWorks 公司）完成，PLS 建模通过 SIMCA-P 12.0（瑞典 MKS Umetrics 公司）完成。

3 结果与讨论

3.1 关键质量属性的确定

药品质量源于设计强调“以终为始”，即首先明确产品 COAs，这是进一步梳理和辨识哪些因素对 COAs 产生影响的前提。对热毒宁注射液金银花和青蒿前处理生产线而言，其“终产品”为萃取后浓缩制得的浸膏，浸膏质量和有效成分含量或转移率可视为该产线的 COAs。在金青醇沉工段，受检测成本和生产周期的制约，仅记录了金青醇沉浓缩后制得浸膏的质量。通过探索性分析发现，259 批金青醇沉浓缩浸膏质量与萃取浓缩浸膏质量、萃取后浸膏中绿原酸的转移率和断氧化马钱子苷的转移率的相关系数 r 分别为 0.69，0.72 和 0.51，相关关系图见图 2。可见金青醇沉浓缩浸膏质量与“终产品”3 个 COAs 之间存在中等至高度的正相关性，因此将金青醇沉浓缩浸膏质量确定为金青醇沉工段的关键质量属性。

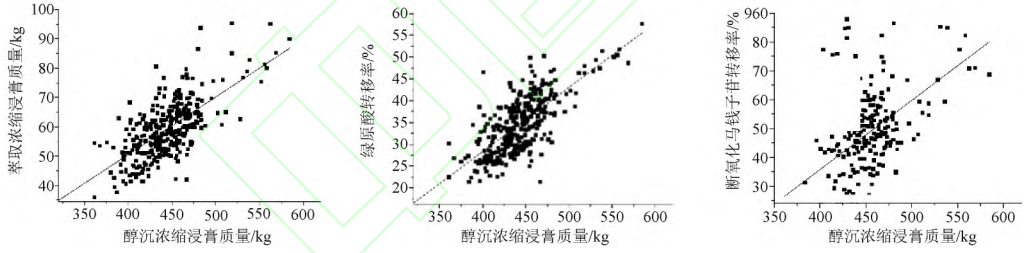


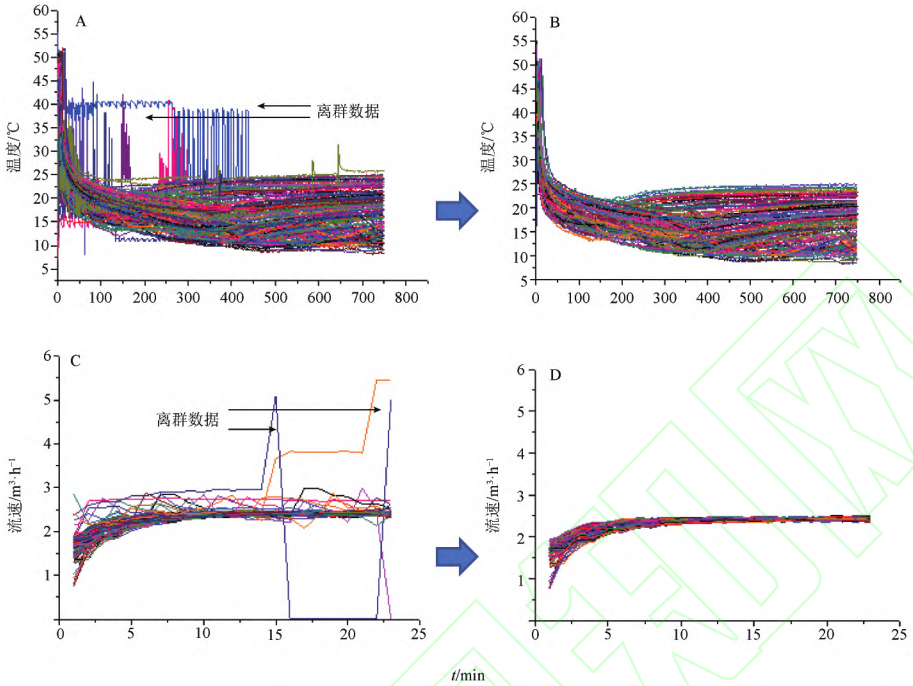
图 2 金青醇沉和金青萃取关键质量属性相关性分析

Fig. 2 Correlation analysis between critical quality attributes from Jinqing alcohol precipitation process and Jinqing extraction process

3.2 数据清洗

由于中药生产现场工况环境复杂，物料组分多样，采集到的过程数据因受到多种因素的影响导致数据中包含较多的噪音信息，为了保障数据的准确性和完备性，需预先对空值数据和离群数据进行清洗。空值数据是从数据库导出数据时，数据点为空的数据。离群数据是不合理的或明显偏离正常水平的数据。以金青浸膏分配过程上部温度和第二阶段加醇流速随时间的变化为例，展示了数据清洗前后的效果对比，见图 3。噪音数据示例见图 3，表现出不规则、非连续变化特征，且明显偏离总体趋势，这些数据在清洗过程将被删除。在收集的金青醇沉工段 259 批数据的基础上，首先将存在空值数据的 27 批数据删除，随后发现离群数据 25 批。数据清洗后，保留原始数据 207 批，占比约 80%。

为检验数据清洗效果，运用 PLS 建模法，对清洗前后的数据进行建模比较。清洗前自变量矩阵大小为 259×3 202，清洗后自变量矩阵大小为 207×3 202。对于清洗前数据，按照批次先后顺序和 7∶3 的比例，将样本分成校正集和验证集，建模后校正集 $R^2X=0.83$，$R^2Y=0.46$ 和 $Q^2=0.27$，预测集 $R^2Y=0.008\,0$，预测误差均方根 RMSEP = 32.55；清洗后，同样依批次先后顺序和 7∶3 的比例，划分校正集和验证集后建模，校正集 $R^2X=0.88$，$R^2Y=0.76$ 和 $Q^2=0.32$，预测集 $R^2Y=0.28$，预测误差均方根 RMSEP = 24.83。表明通过数据清洗可有效提高模型校



A. 清洗前上部温度时序; B. 清洗后上部温度时序图; C. 清洗前第二阶段加醇流速时序图速; D. 清洗后第二阶段加醇流速时序图速。

图 3　数据清洗过程

Fig. 3　Examples of the data cleaning process

正和预测性能。

3.3　特征提取

在金青醇沉工段中,时序参数不同时间点记录的数据间具有强相关性。为提高过程数据的代表性、综合能力和可解释性,根据温度和流速时序数据的变化特点、工艺操作阶段和物料所处状态,从时序数据中提取特征变量,基于特征变量进行数据分析。浸膏分配过程中,温度检测仪表记录温度时序曲线,见图 3A,初始温度最高,在 0~200 min 温度呈下降趋势,而 200 min 之后维持稳定,分配过程结束的温度是加醇阶段的开始时浸膏的初始温度。首先提取分配过程温度时序曲线的第一个时间点数据值和结束时间点的数据值。然后分别计算 0~200 min 以及 200 min 之后时间段内的温度数据的几何平均值。对于每个温度检测仪表可提取 4 个特征变量。图 1 中每个醇沉罐的上部和下部,分别安装了温度检测仪表,因此每个醇沉罐可提取 8 个温度特征变量,见表 1。

浸膏分配结束后开始加醇操作,加醇过程分为 3 个阶段。流速检测仪表记录加醇过程流速时序曲线。以图 3B 展示的加醇第二阶段流速时序曲线为例,说明流速特征变量的提取过程。在加醇第二阶段泵入乙醇的过程中,初始流速最低,在 0~10 min 流速呈上升趋势,而 10 min 之后维持稳定。因此分别计算 0~10 min 以及 10 min 之后时间段内的流速数据的几何平均值,作为加醇第二阶段的特征变量。采用同样的方法,提取加醇第一阶段和第三阶段的流速特征变量。在醇沉生产操作过程中观察到第一和第二阶段的分界点时间,产生较多的絮凝物,因此提取第二阶段加醇初始流速值作为特征变量。图 1 中每个醇沉罐的流速仪表,可提取 6 个平均流速和 1 个第二阶段初始流速,共计 7 个特征变量,见表 1。

将金青醇沉工段的批记录参数和特征提取后获得的特征变量整理见表 1,共获得 48 个参数。其中有 18 个特征参数为点数据,其余 30 个特征参数为基于连续数据推导出的点数据。特征提取后自变量矩阵大小为 48×207,响应变量仍为金青醇沉浓缩浸膏质量,按照批次先后顺序和 7∶3的比例,将 207 个样本分成校正集和验证集,进行 PLS 建模,结果校正集 $R^2X = 0.48$,$R^2Y = 0.57$ 和 $Q^2 = 0.42$,预测集 $R^2Y = 0.30$,预测误差均方根 RMSEP = 24.50,与特征提取前建模效果比较可见,校正集交叉验证 Q^2 明

表 1 金青醇沉工段 48 个特征变量

Table 1 Forty-eight characteristic variables of Jinqing alcohol precipitation section

No.	变量名称	No.	变量名称
1	浸膏密度/kg·m^{-3}	25	罐 2 金银花质量/kg
2	浸膏温度/℃	26	罐 2 青蒿质量/kg
3	金银花出膏量/kg	27	罐 2 加醇量/m^3
4	金青总质量/kg	28	罐 2 浸膏分配前阶段上部平均温度/℃
5	乙醇温度/℃	29	罐 2 浸膏分配后阶段上部平均温度
6	乙醇浓度/%	30	罐 2 浸膏分配上部初始温度/℃
7	罐 1 金银花质量/kg	31	罐 2 浸膏分配上部结束温度/℃
8	罐 1 青蒿质量/kg	32	罐 2 浸膏分配前阶段下部平均温度/℃
9	罐 1 加醇量/m^3	33	罐 2 浸膏分配后阶段下部平均温度
10	罐 1 浸膏分配前阶段上部平均温度/℃	34	罐 2 浸膏分配下部初始温度/℃
11	罐 1 浸膏分配后阶段上部平均温度/℃	35	罐 2 浸膏分配下部结束温度/℃
12	罐 1 浸膏分配上部初始温度/℃	36	罐 2 第一阶段 1~10 min 加醇平均流速/L·h^{-1}
13	罐 1 浸膏分配上部结束温度/℃	37	罐 2 第一阶段 11~48 min 加醇平均流速/L·h^{-1}
14	罐 1 浸膏分配前阶段下部平均温度/℃	38	罐 2 第二阶段 1~10 min 加醇平均流速/L·h^{-1}
15	罐 1 浸膏分配后阶段下部平均温度/℃	39	罐 2 第二阶段 11~23 min 加醇平均流速/L·h^{-1}
16	罐 1 浸膏分配下部初始温度/℃	40	罐 2 第二阶段加醇初始流速/L·h^{-1}
17	罐 1 浸膏分配下部结束温度/℃	41	罐 2 第三阶段 1~10 min 加醇平均流速/L·h^{-1}
18	罐 1 第一阶段 1~10 min 加醇平均流速/L·h^{-1}	42	罐 2 第三阶段 11~23 min 加醇平均流速/L·h^{-1}
19	罐 1 第一阶段 11~48 min 加醇平均流速/L·h^{-1}	43	总加醇量/m^3
20	罐 1 第二阶段 1~10 min 加醇平均流速/L·h^{-1}	44	醇沉上清液体积/m^3
21	罐 1 第二阶段 11~23 min 加醇平均流速/L·h^{-1}	45	醇沉上清液传料体积/m^3
22	罐 1 第二阶段加醇初始流速/L·h^{-1}	46	单效回收时间/min
23	罐 1 第三阶段 1~10 min 加醇平均流速/L·h^{-1}	47	出膏密度/kg·m^{-3}
24	罐 1 第三阶段 11~23 min 加醇平均流速/L·h^{-1}	48	出膏温度/℃

显提高，表明特征筛选可有效减少噪音变量的干扰；而预测集 R^2Y 和 RMSEP 基本保持不变，说明模型自变量由 3 202 个减少为 48 个时，仍可维持预测性能，表明特征提取成功。

3.4 潜在关键工艺参数筛选

在 207 批数据和 48 个特征变量的基础上，分别计算醇沉浓缩浸膏质量与 48 个特征变量的 Pearson 相关系数值见表 2，其中 Pearson>0.20 的包括 21 个参数，即质量参数（$x3$，$x4$，$x7$，$x25$），体积参数（$x9$，$x27$，$x43$，$x44$，$x45$），流速参数（$x18$，$x19$，$x20$，$x23$，$x36$，$x37$，$x38$，$x41$，$x42$），出膏温度（$x48$），浓缩时间（$x46$）和乙醇浓度（$x6$）。

醇沉浓缩浸膏质量与表 1 中 48 个特征变量的灰色关联度分析结果见表 2。保留表中灰色关联度>0.90 的部分[17]，包括密度参数（$x1$，$x47$），温度参数（$x2$，$x16$，$x34$，$x48$），质量参数（$x3$，$x4$，$x7$，$x8$，$x25$，$x27$），乙醇浓度（$x6$），体积参数（$x9$，$x43$，$x44$，$x45$），流速参数（$x19$，$x20$，$x21$，$x37$，$x38$，$x39$）共计 23 个工艺参数。

Pearson 相关分析法对两两参数直接进行相关计算，因此在筛选潜在关键工艺参数时，直接用 Pearson 相关系数筛选的关键工艺参数易造成假阴性的概率的升高。因此本文将 Pearson 相关分析与灰色关联分析结合，共同筛选潜在关键工艺参数筛选，以提高决策可靠性。Pearson 相关分析筛选出的 21 个参数和灰色关联分析筛选出的 23 个参数中，共同部分包括 6 个方面的 15 个参数，见表 3。

计算上述 15 个参数两两之间相关系数，见图 4。其中 4 个药液质量参数两两之间相关系数均大于 0.95，金青总质量（$x4$）是金银花出膏量（$x3$）与青蒿出膏量之和，金银花出膏量（$x3$）又是罐 1 金银花质量（$x7$）与罐 2 金银花质量（$x25$）之和。

乙醇浓度 $x6$ 被筛选出，是因为在实际生产中，受来料波动和储存温度的影响，乙醇在 93%~97% 波动，而非固定使用 95% 乙醇，乙醇浓度的波动会造成加醇量的波动，最终影响醇溶物量[14]。

罐 1（或罐 2）第一阶段 11~48 min 加醇平均流速（$x19$）、以及罐 1（或罐 2）第二阶段 1~10 min 加醇平均流速（$x20$）被筛选出，推测加醇速度的变化影响醇沉过程中絮凝沉淀生成[15]。如徐冰等通过

表 2　Pearson 相关系数与灰色关联度

Table 2　Pearson correlation coefficient and gray correlation calculation results

参数	r	ξ	参数	r	ξ
$x1$	0.07	0.94	$x25$	0.56	0.94
$x2$	−0.12	0.94	$x26$	0.02	0.88
$x3$	0.54	0.93	$x27$	0.52	0.95
$x4$	0.55	0.94	$x28$	0.06	0.89
$x5$	−0.01	0.89	$x29$	0.18	0.83
$x6$	0.55	0.94	$x30$	−0.09	0.87
$x7$	0.53	0.93	$x31$	0.16	0.81
$x8$	−0.07	0.90	$x32$	−0.15	0.89
$x9$	0.48	0.95	$x33$	0.08	0.84
$x10$	0.04	0.89	$x34$	0.07	0.92
$x11$	0.18	0.83	$x35$	0.07	0.82
$x12$	−0.14	0.89	$x36$	0.23	0.72
$x13$	0.16	0.81	$x37$	−0.24	0.92
$x14$	−0.10	0.88	$x38$	0.22	0.94
$x15$	0.08	0.84	$x39$	0.19	1.00
$x16$	0.15	0.92	$x40$	0.01	0.86
$x17$	0.09	0.82	$x41$	0.22	0.80
$x18$	0.21	0.73	$x42$	0.21	0.83
$x19$	−0.22	0.92	$x43$	0.51	0.95
$x20$	0.35	0.95	$x44$	0.63	0.95
$x21$	0.18	0.94	$x45$	0.64	0.95
$x22$	0.19	0.88	$x46$	0.38	0.87
$x23$	0.22	0.84	$x47$	−0.11	0.94
$x24$	0.10	0.84	$x48$	0.29	0.95

表 3　Pearson 相关系数与灰色关联分析共同筛选参数

Table 3　Common parameters screened by Pearson correlation coefficient and grey correlation analysis

参数种类	影响参数
药液质量	$x3$-出膏量,$x4$-金青总质量,$x7$-罐 1 金银花质量,$x25$-罐 2 金银花质量
浓度	$x6$-乙醇浓度
加醇速度	$x19$-罐 1 第一阶段 11～48 min 加醇平均流速,$x20$-罐 1 第二阶段 1～10 min 加醇平均流速,$x37$-罐 2 第一阶段 11～48 min 加醇平均流速,$x38$-罐 2 第二阶段 1～10 min 加醇平均流速
加醇量	$x9$-罐 1 加醇量,$x27$-罐 2 加醇量,$x43$-总加醇量
上清液体积	$x44$-醇沉上清液体积,$x45$-醇沉上清液传料体积
温度	$x48$-出膏温度

金银花醇沉过程的在线监控将醇沉过程划分为 4 个阶段[16]，第一阶段(前 10%时间) 析出少量絮状物和颗粒,第二阶段(10%～27%时间) 颗粒增长和絮凝物结块,第三阶段(27%～约 50%时间) 大块絮状物打散,第四阶段(约 50%至结束) 颗粒继续打散至均匀。因此类比金银花醇沉过程,本文金青醇沉第一阶段 11～48 min 加醇过程和第二阶段 1～10 min 加醇过程,位于金青醇沉过程时间轴的约 11%～62%位置,与上述金银花醇沉第二、三阶段和第四阶段初期对应,是醇不溶物析出的关键环节。

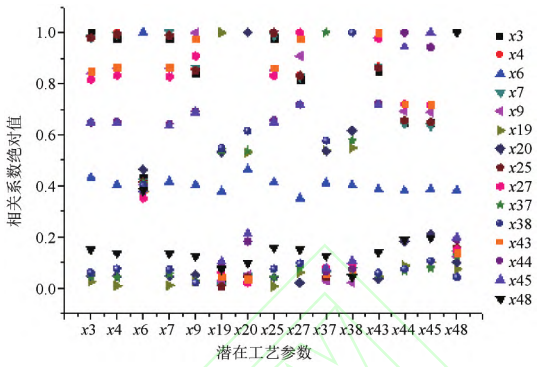


图 4　潜在关键工艺参之间相关系数绝对值

Fig. 4　Absolute value of correlation coefficient between potential key process parameters

总加醇量($x43$) 是罐 1 加醇量($x9$) 和罐 2 加醇量($x27$) 之和,3 个加醇量之间的相关系数均大于 0.95。醇沉前药液密度严格控制在非常窄的范围之内(±0.01),受饮片原料波动的影响,每批提取浓缩后所得浸膏的固形物量是不同的,因此相同密度的药液具有不同体积。此外,金青醇沉终点药液含醇量相同,因此加醇量随初始浸膏体积的变化而在一定范围内波动,这一波动将进一步传递至醇沉上清液体积参数的变化。醇沉上清液体积 $x44$ 和醇沉上清液传料体积 $x45$ 的相关系数为 0.94。

3.5　预测性建模

将数据清洗获得的 207 批数据,按照批次先后顺序和 7∶3的比例,将样本分成两部分,前 145 批用于预测性建模,后 62 批用于验证模型预测性能。以 15 个潜在关键工艺参数作为输入,建立预测醇沉浓缩浸膏质量的 PLS 模型。以校正集的决定系数(R^2)、校正误差均方根 RMSEC 及平均相对预测误差 δ,其中,相对预测误差=(真实值−预测值)/真实值)×100%,预测集的决定系数(R^2)、预测误差均方根 RMSEP 及平均相对预测误差 δ 对不同模型的校正和预测性能进行评价。潜变量因子数对 PLS 模型性能有显著的影响,潜变量因子选择过少,会导致提取信息不全,模型欠拟合,而潜变量因子选择过多,会导致模型过于复杂,出现过拟合现象。以 RMSECV,RMSEC,RMSEP 为指标,优选最佳潜变量因子数,见图 5。可见当潜变量因子数为 3 时,交叉

验证指标 RMSECV 最小。因此选择 3 个潜变量因子建立 PLS 模型，校正集 $R^2X=0.78$，$R^2Y=0.50$ 和 $Q^2=0.42$，预测集 $R^2Y=0.30$，预测误差均方根 RMSEP=22.84；校正集与预测集的相对误差均小于 5%，且 RMSEP/RMSECV 接近 1，表明该模型具有稳健的模型结构和预测性能。以筛选出的 15 个潜在关键工艺参数建立的 PLS 模型预测精度上优于醇沉工艺全部参数(3 202 个参数)，且与 48 个特征参数的预测能力相当。

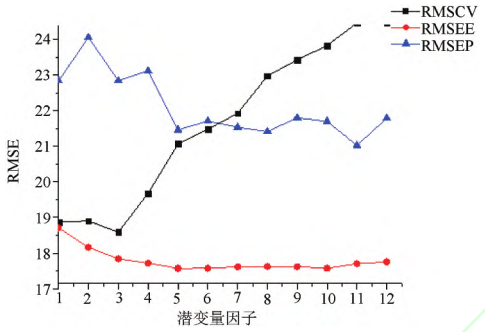


图 5 醇沉浓缩浸膏质量预测模型主成分数

Fig. 5 The number of principal components of the alcohol precipitation concentrated extract weight prediction model

运用 VIP 对 15 个潜在工艺参数进行分析，见图 6。可见对金青醇沉影响最大的工艺参数是 $x45$-醇沉上清液传料体积，传料至浓缩罐的醇沉上清液进入浓缩工序，因此与浓缩液质量直接相关，与之相关的参数为 $x44$；其次为 $x4$-金青总质量，该参数代表金青醇沉的过程输入，与之相关的参数为 $x25$，$x7$ 和 $x3$；再次为 $x43$-加醇量，与之相关的参数为 $x9$ 和 $x27$。上述 9 个参数的 VIP 均大于 1，结合其实际物理意义和作用，可将 9 个参数定义为关键工艺参数。乙醇浓度 $x6$ 和加醇速度类参数 $x20$，$x38$，$x19$ 和 $x37$ 的 VIP 虽然小于 1，但在实际生产中，仍应当作为潜在关键工艺参数予以持续的趋势监控。

4 总结和展望

本文采集热毒宁注射液金青醇沉工段历史生产数据共计 829 318 数据点，呈现出数据量大、价值密度低、来源多样等大数据的部分特征。通过数据清洗和特征提取，数据点减少为 9 936 个。采用 Pearson 相关分析和灰色关联度分析进行综合决策，筛选出影响金青醇沉浓缩浸膏质量的 15 个潜在关键工艺参数，数据点进一步缩减至 3 105 个。进一步

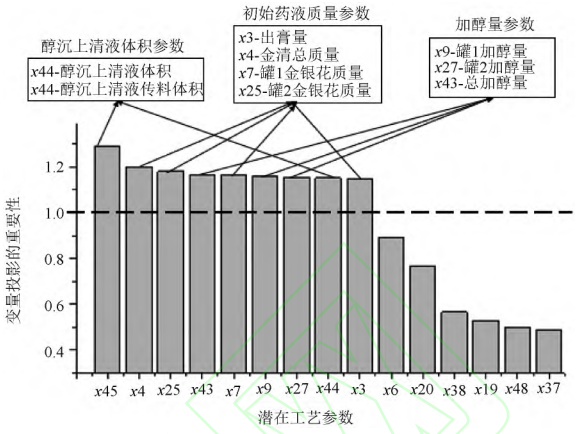


图 6 关键工艺参数辨识

Fig. 6 Identification of critical process parameters

通过 PLS 定量预测建模，辨识出 9 个关键工艺参数，至此数据点为 1 863 个，占原始数据的 0.28%。结果表明醇沉上清液传料体积、金青总质量、罐 2 金银花质量、总加醇量、罐 1 金银花质量、罐 1 加醇量、罐 2 加醇量、醇沉上清液体积、金银花出膏量是对金青醇沉浓缩浸膏质量影响最为显著的参数。

在热毒宁注射液数字化制造的基础上，如何挖掘生产大数据的潜在价值，实现数据和知识的模型化，以辅助生产精准质量控制和智能决策，是热毒宁注射液由数字化制造向智能化制造迈进的关键。从全局数据出发，采用大数据分析的方法筛选得到的关键工艺参数有助于准确地描述金青醇沉过程质量传递规律。本文研究发现，提取浓缩工段制得的金青浸膏的变化是影响金青醇沉所得浸膏变化的关键因素，未来的工作将进一步加强对原料质量波动规律的理解；进一步研究过程质量监控点布局的合理性，并探索先进质量传感器的在线应用；提高对过程数据噪音波动规律的理解，探索包含缺失数据和异质异构数据的处理算法，进一步提高大数据的价值密度[17]。

[参考文献]

[1] 辛国斌. 智能制造探索与实践：46 项试点示范项目汇编[M]. 北京：电子工业出版社，2016.

[2] International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human use(ICH). Pharmaceutical development. Q8[EB/OL]. [2019-09-01]. https://www.ich.org/page/quality-guidelines.

[3] 徐冰，史新元，吴志生，等. 论中药质量源于设计[J]. 中国中药杂志，2017，42(6)：1015.

[4] International Conference on Harmonization of Technical Require-

ments for Registration of Pharmaceuticals for Human Use (ICH). Lifecycle Management. Q12 [EB/OL]. [2019-09-01]. https://www.ich.org/page/quality-guidelines.

[5] 崔向龙，徐冰，孙飞，等. 质量源于设计在银杏叶片制粒工艺中的应用(Ⅲ)：基于设计空间的过程控制策略[J]. 中国中药杂志，2017，42(6)：1048.

[6] 崔雅华，王茜，徐冰，等. 质量源于设计：基于知识组织的中药生产潜在关键工艺参数的辨识[J]. 中国实验方剂学杂志，2016，22(15)：1.

[7] 刘爽悦，沈金晶，李文龙，等. 3种关键工艺参数辨识方法的比较研究[J]. 中草药，2016，47(18)：3193.

[8] 陈勇，陈明，王钧，等. 基于灰色关联分析法辨识中药生产过程关键工艺参数[J]. 中草药，2019，50(3)：45.

[9] 严斌俊，郭正泰，瞿海斌，等. 丹红注射液醇沉关键工艺参数筛选方法[J]. 中国中药杂志，2013，38(11)：1672.

[10] 徐冰，崔向龙，杨婵，等. 质量源于设计在银杏叶片制粒工艺中的应用(Ⅱ)：颗粒关键质量属性辨识[J]. 中国中药杂志，2017，42(6)：1043.

[11] PEARSON W R，LIPMAN D J. Improved tools for biological sequence comparison [J]. Proc Natl Acad Sci USA，1988，85(8)：2444.

[12] 邓聚龙. 灰理论基础 [M]. 武汉：华中科技大学出版社，2002.

[13] 刘晓红. 环境规制情景下我国农村居民间接碳排放研究——基于STIRPAT模型和PLS-VIP方法[J]. 资源开发与市场，2016，32(12)：1471.

[14] 肖琼，沈平孃. 中药醇沉工艺的关键影响因素[J]. 中成药，2005，27(2)：143.

[15] 孙秀玉，王英姿，乔延江，等. 正交试验法优化清开灵注射液中金银花提取液的醇沉工艺[J]. 世界科学技术——中医药现代化，2014，16(1)：187.

[16] 徐冰，史新元，乔延江，等. 金银花醇沉多阶段多变量统计过程控制研究[J]. 中华中医药杂志，2012，27(4)：784.

[17] 徐冰，史新元，罗赣，等. 中药工业大数据关键技术与应用[J]. 中国中药杂志，2020，45(2)：221.

[责任编辑　孔晶晶]