

数据驱动的热毒宁注射液金银花浓缩工序黄金批次识别与优化研究

童枫^{1,3}, 徐芳芳^{1,3*}, 闫逸伦^{1,3}, 刘恒旭^{1,3}, 钱雅婷^{1,3}, 宋秋月^{1,3}, 章晨峰^{2,3}, 王振中^{1,3}, 张欣^{1,2*}

1. 中药制药过程控制与智能制造技术全国重点实验室, 江苏 连云港 222001

2. 江苏省海洋药物和现代中药创制重点实验室, 江苏 连云港 222001

3. 江苏康缘药业股份有限公司, 江苏 连云港 222001

摘要: 目的 聚焦于热毒宁注射液 (Reduning Injection, RI) 金银花浓缩工序, 旨在开发并验证一套数据驱动的技术框架, 以实现“黄金批次”的客观评价、智能识别与工艺优化, 并从根本上揭示影响“黄金批次”形成的关键工艺机制与量化控制策略, 以期提升生产过程的效率与稳定性。方法 收集了 170 批金银花浓缩工序生产工艺数据, 创新性地构建了“综合浓缩效能比” (E) 作为工艺性能评价指标。采用一种潜空间引导的效能阈值自适应划分策略 (latent-space guided adaptive performance thresholding, LGAPT), 客观地将生产批次划分为黄金批次与非黄金批次。随后, 构建了基于极限梯度提升决策树算法 (extreme gradient boosting, XGBoost) 的分类预测模型, 并结合 SHAP (Shapley 加性解释) 与部分依赖图 (partial dependence plots, PDP) 等模型解释技术, 深入挖掘影响批次质量的关键工艺特征及其最优调控区间。最后通过非参数检验、效应量分析及核密度估计对模型发现进行统计学验证。结果 确定黄金批次的 E 分界阈值为 $0.1363 \text{ kg}/(\text{m}^3 \cdot \text{min})$, 据此将 170 批数据划分为 61 个黄金批次和 109 个非黄金批次; XGBoost 模型在识别黄金批次方面表现优异 (测试集的 F1 分数为 0.84); 模型解释分析识别出 LT_std-2、LT_skew-1 和 T3_abd-2 等是影响黄金批次形成的核心特征, 并明确了其优化区间。统计验证结果与模型解释高度一致, 证实了该结论的可靠性。结论 建立了一套从性能量化、智能划分到模型寻优的完整技术框架, 成功识别了影响金银花浓缩工序质量的关键特征及其控制范围; 该框架为中药生产过程的标准化与智能化控制提供了有力的理论依据, 推动了从“经验判断”到“数据驱动”的工艺优化模式转变。

关键词: 热毒宁注射液; 金银花; 浓缩; 数据驱动; 黄金批次; 模型解释; 智能识别; 综合浓缩效能比

中图分类号: R283.6 文献标志码: A 文章编号: 0253-2670(2026)11-4172-11

DOI: 10.7501/j.issn.0253-2670.2026.11.009

Data-driven golden batch identification and intelligent optimization for *Lonicerae Japonicae Flos* concentration process of Reduning Injection

TONG Feng^{1,3}, XU Fangfang^{1,3}, YAN Yilun^{1,3}, LIU Hengxu^{1,3}, QIAN Yating^{1,3}, SONG Qiuyue^{1,3}, ZHANG Chenfeng^{2,3}, WANG Zhenzhong^{1,3}, ZHANG Xin^{1,2}

1. State Key Laboratory of Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture, Lianyungang 222001, China

2. Jiangsu Key Laboratory on Innovation for Marine Medicine and Modern Chinese Medicine, Lianyungang 222001, China

3. Jangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

Abstract: Objective This study focuses on the Jinyinhua (*Lonicerae Japonicae Flos*, LJF) concentration process in the production of Reduning Injection (热毒宁注射液, RI), with the objective of developing and validating a data-driven technical framework. The framework is designed to enable objective evaluation, intelligent identification, and process optimization of “golden batches”, while fundamentally elucidating the key process mechanisms and quantitative control strategies influencing their formation. Ultimately, this approach aims to enhance the efficiency and stability of the production process. **Methods** This study collected production process data for 170 batches of LJF concentrate, and innovatively established “comprehensive concentration efficiency ratio” (E) as an

收稿日期: 2025-12-30

基金项目: 国家工信部产业基础再造和制造业高质量发展专项 (TC2308068); 国家科技部长三角科技创新共同体联合攻关项目 (2023CSJGG1700)

作者简介: 童枫, 男, 硕士, 研究方向为中药制药过程新技术。E-mail: tongfeng5324@126.com

*通信作者: 张欣, 博士, 研究方向为中药制药过程新技术。E-mail: zxtcm@126.com

徐芳芳, 副主任药师, 硕士生导师, 从事中药智能制造研究。E-mail: 879164331@qq.com

evaluation index for process performance. A latent-space guided adaptive performance thresholding (LGAPT) strategy was employed to objectively classify the production batches into golden and non-golden categories. Subsequently, a classification prediction model based on extreme gradient boosting (XGBoost) was constructed. By integrating model explanation techniques such as SHAP (Shapley additive explanations) and partial dependence plots (PDP), the key process features affecting batch quality and their optimal control ranges were systematically investigated. Finally, the model findings were statistically verified through non-parametric tests, effect size analysis, and kernel density estimation. **Results** The study established a threshold of $E = 0.1363 \text{ kg}/(\text{m}^3 \cdot \text{min})$, classifying the 170 batches into 61 golden and 109 non-golden batches. The XGBoost model demonstrated excellent performance in identifying golden batches, achieving an F1-score of 0.84 on the test set. Model interpretation identified features such as LT_std-2, LT_skew-1, and T3_abd-2 as core determinants for golden batch formation and defined their optimal ranges. The statistical validation results were highly consistent with the model interpretations, confirming the reliability of the findings. **Conclusion** This study established a comprehensive technical framework encompassing performance quantification, intelligent classification, and model-based optimization. By successfully identifying the key process features and their optimal control ranges, this work lays a methodological foundation for the standardization and intelligent control of traditional Chinese medicine (TCM) manufacturing, driving a paradigm shift in process optimization from an empirical-based to a data-driven approach.

Key words: Reduning Injection; *Lonicerae Japonicae Flos*; concentration; data-driven; gold batch; model interpretation; intelligent recognition; comprehensive concentration efficiency ratio

提升生产效率是“中国制造 2025”与智能制造战略的核心目标之一。2025 年 2 月，中华人民共和国工业和信息化部在总结“智能工厂梯度培育行动”成效时明确指出，通过数字化、智能化转型，全国卓越级智能工厂的生产效率平均提升了 22.3%，实现了显著的提质增效^[1]。这一宏观数据凸显了数据驱动技术在赋能传统制造业方面的巨大潜力，也为中药制造行业的转型升级指明了方向^[2]。

热毒宁注射液 (Reduning Injection, RI) 是江苏康缘药业股份有限公司的独家品种，由金银花、青蒿和栀子 3 味药材提取精制而得^[3]。提取精制过程包括提取、浓缩、醇沉、萃取、干燥等工艺，金银花浓缩是 RI 提取精制过程中的关键工艺单元之一，通过三效板式浓缩富集绿原酸等有效成分，去除多余水分^[4]。然而，在实际生产中，即便是使用相同的设备与标准操作规程，不同批次间的生产效率仍存在显著波动，这种波动主要来源于原药材批次差异、工艺参数变化以及操作执行偏差，同时这种不稳定性意味着大量批次未能达到最优生产水平，不仅导致资源利用率参差不齐、生产周期难以预测，更从根本上制约了整体生产效能的提升和成本的有效控制。因此，如何系统性地识别并复制最高效的生产模式，消除低效批次，成为当前智能制造背景下制药企业面临的共性挑战。为应对这一挑战，借鉴“黄金批次” (golden batch) 理念，即识别并学习历史生产中工艺表现最优的批次，成为一种极具潜力的优化路径^[5]。通过系统性地分析黄金批次的工艺数据，可以揭示达成高效生产的关键工艺控制

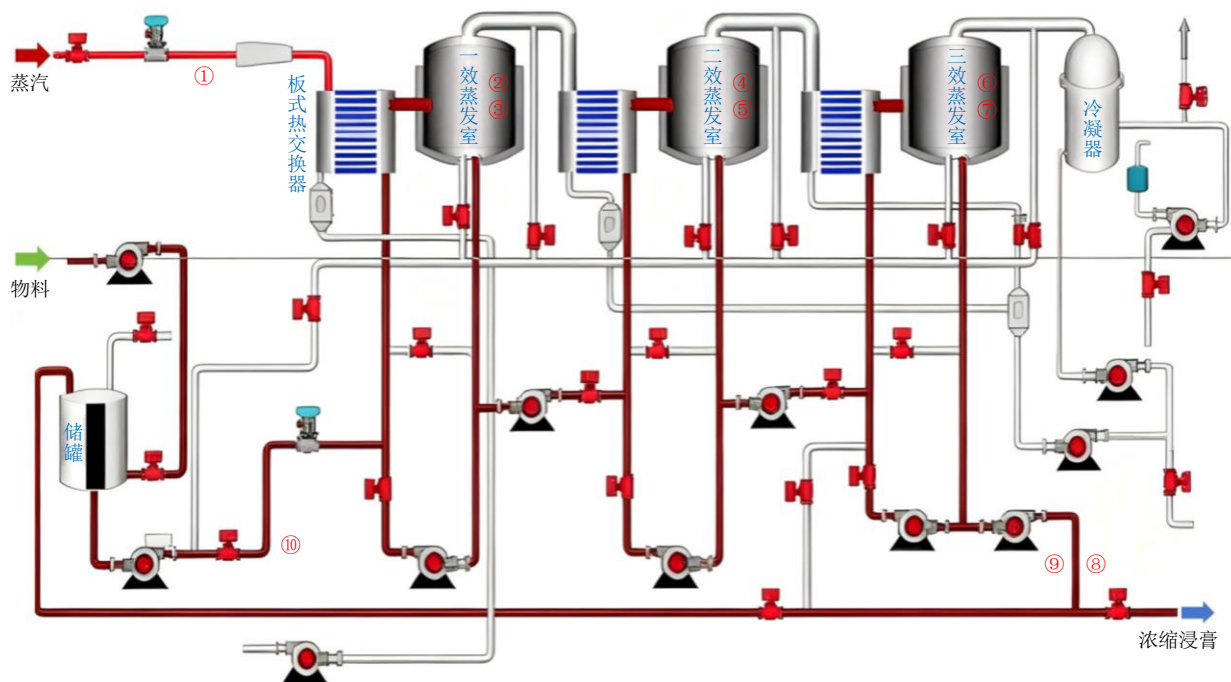
策略，从而指导常规批次的生产向黄金批次看齐，最终在保证产品质量合格的前提下，实现整体生产效率和经济效益的最大化^[6-7]。

如何客观评价批次表现是“黄金批次”判定的核心，其评价指标的选择至关重要。在实际生产中往往存在“高收率 vs 高杂质”“短周期 vs 高能耗”等矛盾，合理的指标选择可以确保选出的黄金批次既满足法规要求，又兼顾经济效益^[8]。金银花浓缩工序包含多个生产步骤，涉及温度、压力、流速等众多相关工艺参数，依托传感器可以实时获取各个生产批次的工艺参数变化情况，已具备数据分析的基础。本研究聚焦于金银花浓缩工序，基于历史生产工艺参数数据，结合工业大数据分析技术，旨在：构建一个能量化生产效率的综合性指标；建立一种客观、数据驱动的黄金批次划分方法；利用机器学习模型识别影响生产效率的关键特征，并确定其优化区间，以实现从工艺表征到特征优化的全流程覆盖，为中药注射剂浓缩工艺的智能优化控制提供理论支撑与方法指导。

1 数据来源

本研究涉及的金银花浓缩工序采用三效外循环蒸发工艺 (图 1)，提取液在真空条件下，依次通过一、二、三效蒸发器，利用蒸汽进行间接加热，药液在蒸发器内循环流动，水分被蒸发后，浓缩液最终在三效蒸发器中进行终点浓缩，直至达到预设密度。整个过程的关键工艺参数均由分布式控制系统 (distributed control system, DCS) 进行监控。

本研究收集了江苏康缘药业有限公司 2024 年生



图中序号分别对应以下参数的在线传感器位置：①蒸汽压力 (SP)；②一效温度 (T1)；③一效压力 (P1)；④二效温度 (T2)；⑤二效压力 (P2)；⑥三效温度 (T3)；⑦三效压力 (P3)；⑧药液温度 (LT)；⑨药液体积流量 (LFR)；⑩进液体积流量 (IFR)。

The serial numbers in the figure respectively correspond to the on-line sensor positions of the following parameters: ① steam pressure (SP); ② primary temperature (T1); ③ primary pressure (P1); ④ secondary temperature (T2); ⑤ secondary pressure (P2); ⑥ three-effect temperature (T3); ⑦ three-effect pressure (P3); ⑧ liquid temperature (LT); ⑨ per-name ruby liquid flow rate (LFR); ⑩ attending inlet flow rate (IFR).

图1 金银花浓缩工序设备流程图及关键参数采集点

Fig. 1 Flowchart of *Lonicerae Japonicae Flos* concentration process and locations of key parameter acquisition

产的中间体检合格、符合质量放行标准的金银花浓缩工序数据 170 批，数据源包括 2 部分：其一为生产批记录，由车间工艺员人工录入，包括提取液出液量和浓缩出膏量等；其二为工艺时序数据库，金银花浓缩工序的工艺参数实时数据由生产设备上的检测仪表生成，后传入 DCS 中，再由实时数据采集软件通过 OPC UA 协议从 OPC2web 工具中采集数据并存储至工艺时序数据库，实现从数据源头到传输和存储的全链路自动化，数据涵盖了 10 个关键过程工艺参数，包括蒸汽压力 (SP, MPa)、一效/二效/三效温度 (T1/T2/T3, °C)、一效/二效/三效压力 (P1/P2/P3, MPa)、进液体积流量 (IFR, L/h)、药液体积流量 (LFR, L/h)、药液温度 (LT, °C)。后续的特征工程均是基于上述 10 个原始工艺参数的时序数据构建。

2 方法与结果

2.1 综合浓缩效能比 (comprehensive concentration efficacy ratio, E) 指标构建

进液量 (V, m³)、浓缩时长 (T, min) 与出膏量 (M, kg) 作为金银花浓缩工艺的 3 个核心工艺

性能评价指标，其动态关联性直接决定了浓缩过程的综合效能。传统工艺评价常采用单一维度指标，如出膏量 (M) 衡量产率，或以单位时间处理量 (L/T) 表征效率，但这类指标存在显著的局限性——单纯追求高出膏量可能伴随进液量激增或浓缩时间延长，导致资源过度消耗；而过度压缩浓缩时间虽能提升单位时间处理量，却可能因蒸发不充分造成出膏量不足或有效成分损失。

为解决这一矛盾，本研究提出 E 这一复合参数，以实现多维度工艺指标的有机整合。该参数的构建逻辑源于对浓缩过程本质特征的深度解构：进液量 (V) 代表原料投入规模，浓缩时长 (T) 体现时间资源占用，出膏量 (M) 则反映最终产出效果。通过构建投入-时间-产出的三维评价体系，可有效避免传统指标的片面性：高产率可能隐含高能耗，高效率可能牺牲产品质量。

$$E = M / VT$$

E 的单位为 kg/(m³·min)，该指标的物理意义表征为每立方米进液量在每分钟内所转化的出膏量

各批次 E 值的分布特征如图 2 所示。从分布形

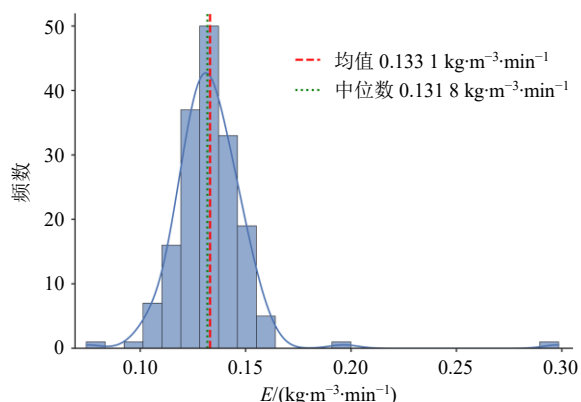


图2 E 的分布特征

Fig. 2 Distribution characteristics of E

态观察可知, E 值呈现右偏单峰分布特征, 核心区间集中在 0.12~0.15 kg/(m³·min), 涵盖约 78% 的生产批次。统计结果显示, 170 批次 E 值的均值为 (0.133 1±0.019 0) kg/(m³·min), 中位数为 0.131 8 kg/(m³·min)。均值略微大于中位数的右偏长尾特征, 表明存在部分效能极为突出的黄金批次, 拉高了整体的均值水平。

进一步的数据离散度分析表明, 批次间呈现出显著的两极分化趋势。虽然主体批次的效能差异控制在±5%以内, 但分布尾端的极差悬殊: 效能最高批次的 E 值达 0.30 kg/(m³·min), 而最低批次仅为 0.07 kg/(m³·min)。值得关注的是, 效能最优的前 5% 批次 [E>0.21 kg/(m³·min)] 与效能最低效的 5% 批次 [E<0.10 kg/(m³·min)] 之间跨度超过 300%, 直观反映出生产工艺的批间差异与不稳定性。

2.2 时序数据特征提取

在金银花浓缩生产过程中, 工艺参数随时间动态变化, 形成的 170 批原始数据具有高维度、强时序性的特点。直接将此类原始数据用于后续的数据

分析工作, 会存在显著局限性: 一方面, 原始时序数据包含大量冗余信息, 易导致模型过拟合, 降低泛化能力; 另一方面, 单一批次的参数曲线无法直接量化其工艺稳定性、均值水平及波动特征, 难以建立工艺参数与批次工艺性能的关联关系^[9]。从工业生产实践需求来看, 黄金批次的核心特征在于工艺参数的稳定性、合理性及一致性, 因此, 需通过特征提取将高维度的原始时序数据转化为低维度、可解释、具有明确物理意义的特征指标, 实现对每批生产过程的关键信息浓缩, 提升模型的准确性与工业应用价值^[10]。

基于 170 批金银花浓缩工艺参数的时序特性, 结合“工艺稳定性-参数合理性-波动特征”的核心需求, 本研究对关键工艺参数时序数据进行特征提取。鉴于金银花浓缩工艺在操作上可分为 2 个彼此独立的亚批, 本研究分别对每个亚批内的时序数据进行提取, 以捕捉不同亚批内工艺的独有特性。

针对每批生产过程中的 10 个关键工艺参数, 即 SP、IFR、T1、P1、T2、P2、T3、P3、LFR、LT, 提取了涵盖 6 大类共 11 种统计特征。这些特征较全面地反映了生产过程信息: 基于最大值/最小值界定安全边界, 均值/标准差刻画中心趋势与离散度, 变异系数评估相对波动, 偏度/峰度揭示分布形态, 曲线下面积量化累积效应, 以及通过上升速率、峰值持续时间、动态偏离时长等时间特征捕捉动态特性。详细的特征类别及其物理意义如表 1 所示。

特征命名遵循“工艺参数缩写_统计量-亚批号”的格式, 如“LT_mean-1”代表第 1 亚批 LT 的均值 (mean)。通过上述特征提取, 最终将 170 批次的高维时序数据, 转化为 1 个 170×220 的结构化特征向量矩阵。该矩阵定量地描述了每批生产在各个关

表 1 浓缩过程时序数据特征提取方案

Table 1 Feature extraction scheme for concentration process time-series data

特征类别	特征名称	物理/统计意义	特征数量
中心趋势	均值 (mean)	反映工艺参数在整个过程中的平均水平	20
波动性	标准差 (std)	衡量参数的绝对波动幅度与相对波动大小, 表征工艺稳定性	40
	变异系数 (cv)		
分布形态	偏度 (skew)	描述参数分布的对称性与尖峭程度, 反映操作是否偏离中心及极端值出现的频率	40
	峰度 (kurt)		
累积效应	曲线下面积 (auc)	参数值对时间的积分, 量化总能量输入或物料累积	20
极值特征	最大/最小值 (max/min)	界定参数运行的安全边界或实际达到的极限工况	40
时间特征	峰值持续时间 (pkd)	捕捉过程的动态响应速度、参数超出警戒限但在规格限内的累计时长等关键动态特性	60
	动态偏离时长 (abd)		
	上升速率 (rr)		

键工艺参数上的中心趋势、波动性及分布形态，既实现数据降维，又保留核心信息，可支撑后续黄金批次初步筛选与分类模型构建，为工艺优化与质量控制提供数据支撑。

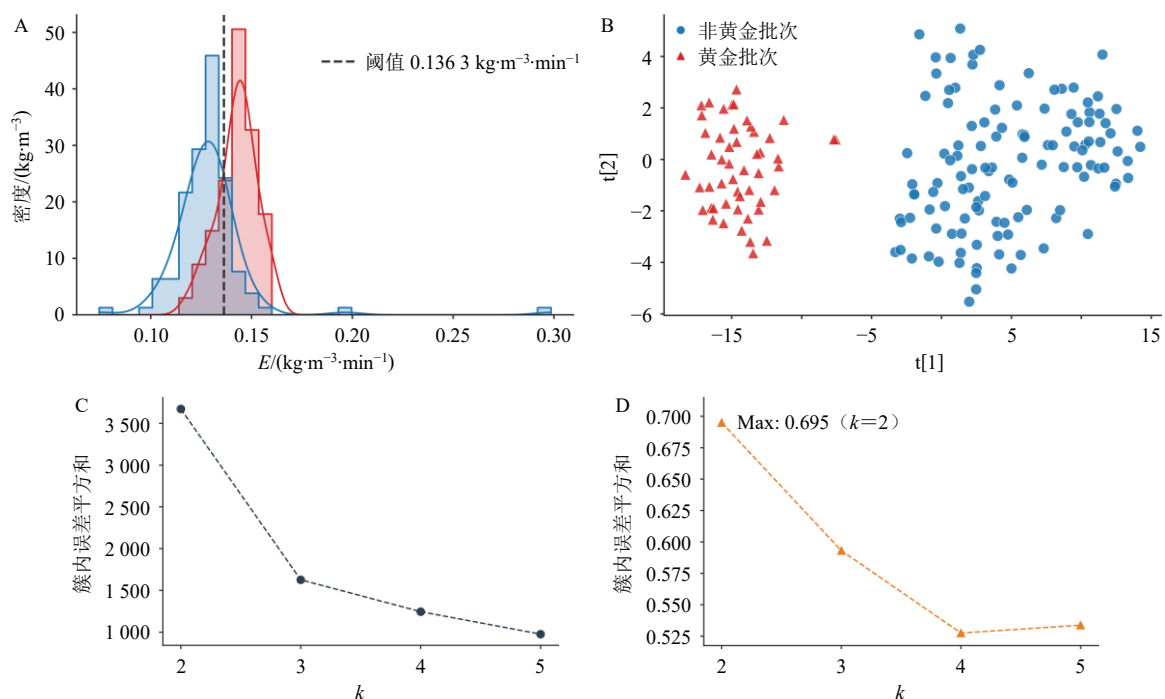
2.3 基于潜空间引导的效能阈值自适应划分策略 (latent-space guided adaptive performance thresholding, LGAPT) 的建立

针对生产过程中批次质量评估依赖主观经验、难以从高维数据中建立客观标准的问题，本研究提出了一种 LGAPT 策略。该策略融合了非线性降维与聚类分析，旨在通过无监督学习的方式，数据驱动地建立黄金批次的客观划分标准，为工艺分级管理提供科学依据。该方法的核心思想是：利用 *t*-SNE (perplexity=30) 算法，将高维、复杂的生产过程数据映射到二维潜空间，以可视化的方式揭示批次之间内在的相似性与差异性结构^[11]；随后，在构建的潜空间上，采用 *K*-means 算法 (采用 *K*-means++ 进行中心初始化，*n*_init=10)，将所有批次自动划分为 2 个性能表现上存在显著差异的自然簇群；最后，将聚类结果与黄金批次质量评价指标 *E* 进行关联，通过计算 2 个簇群效能比的中位数，自适应地生成一个最优划分阈值，将一个依赖于复杂模型的聚类

结果，转化为一个独立、普适且具有明确物理意义的性能标准，不仅提升了“黄金批次”定义在未来应用中的泛化能力，也为后续监督学习模型提供了更具可解释性的业务目标，最终实现对所有生产批次的“黄金批次”与“非黄金批次”的软分类，将无监督的聚类结果转化为有监督的分类标准。

本研究采用 *t*-SNE 与 *K*-means 的组合，旨在为后续的监督学习构建客观且清晰的分类标签。考虑到中药生产数据的高维与非线性特性，首选 *t*-SNE 进行降维，该算法的核心优势在于能将原始数据映射为簇间分离度高的二维可视化空间。在此基础上应用 *K*-means 算法，其作为一种硬聚类方法，能为每一批分配唯一的类别归属，这对于生成明确的“黄金批次/非黄金批次”二元标签以训练分类器，并确定单一性能阈值是必要的。*t*-SNE 的降维结果为 *K*-means 的收敛和稳定性创造了理想条件，二者结合确保了最终划分结果的稳健性与可解释性^[12]。

LGAPT 策略的具体实施与自适应阈值生成结果如图 3 所示。首先，基于算法的无监督学习，所有生产批次被自动划分为 2 个性能表现存在显著差异的自然簇群，算法分别计算 2 个簇群效能比的中位数，并取其均值作为最佳分界线。如图 3-A 所示，



A-效能分布与自适应阈值；B-*t*-SNE 潜空间聚类结果；C-肘部法则确定最佳聚类数；D-轮廓系数验证。

A-efficiency distribution and adaptive threshold; B-*t*-SNE latent space clustering results; C-elbow rule determines the best clustering number; D-profile coefficient verification.

图 3 基于 LGAPT 策略的黄金批次划分结果

Fig. 3 Gold batch identification results via LGAPT strategy

尽管 2 簇效能分布存在部分重叠，但最终确定的自适应划分阈值为 $0.1363 \text{ kg}/(\text{m}^3 \cdot \text{min})$ ，该阈值可有效区分黄金批次与非黄金批次，共有黄金批次 61 批、非黄金批次 109 批。进一步的 *t*-SNE 潜空间聚类可视化（图 3-B）也直观印证了这 2 类批次在特征空间中存在明确的分布界限。

上述聚类过程中将簇群数量 (*k*) 设定为 2 (*k*=2)，是经过严格参数评估的最优选择。图 3-C 显示，肘部法则在 *k*=2 时出现显著“肘点”；同时，图 3-D 中的轮廓系数也在 *k*=2 时达到全局最大值 (0.695)。这 2 项内部评估指标共同证实，将数据划分为 2 个簇群能够最优地刻画生产过程中的批间性能差异模式。

综上所述，本研究确立了黄金批次的客观标准：凡 $E \geq 0.1363 \text{ kg}/(\text{m}^3 \cdot \text{min})$ 的批次，即被界定为黄金批次。本研究建立的量化标准突破了传统经验判据的局限性，为实施差异化的生产管控策略提供了可靠的数据支撑。

2.4 极限梯度提升决策树 (extreme gradient boosting, XGBoost) 模型^[13]的建立

为建立特征与批次质量 (黄金/非黄金批次) 的映射关系，本研究采用 XGBoost 模型构建二分类预测器。XGBoost 作为梯度提升决策树 (gradient boosting decision tree, GBDT) 的优化实现，在解决复杂工业数据的非线性分类问题方面具有显著优势。该算法通过迭代地训练一系列弱分类器，并将这些弱分类器组合成一个强分类器，以提高模型的预测性能。其核心思想是在每一轮迭代中，根据前一轮模型的预测结果与真实值之间的残差，构建一棵新的决策树，然后将这棵新树的预测结果与之前的模型结果进行加权累加，不断减小残差，直至达到预设的迭代次数或满足停止条件^[14]。

为获得最佳模型性能并有效防止过拟合，首先采用分层抽样方法，将包含 170 批工艺参数的时序特征矩阵按照 7 : 3 的比例划分为训练集与测试集，确保训练集与测试集中黄金批次与非黄金批次样本的比例与原始数据集保持一致。随后，利用带交叉验证的递归特征消除 (recursive feature elimination with cross-validation, RFECV) 对原始的 220 个工艺特征进行筛选，该过程采用分层的 5 折交叉验证机制，并在此过程中引入贝叶斯优化算法，以寻找模型关键超参数的最优组合^[15]。在获得优化参数后，构建最终分类模型。特征筛选过程如图 4 所示，

贝叶斯优化的超参数搜索范围如表 2 所示，模型性能如表 3 所示。

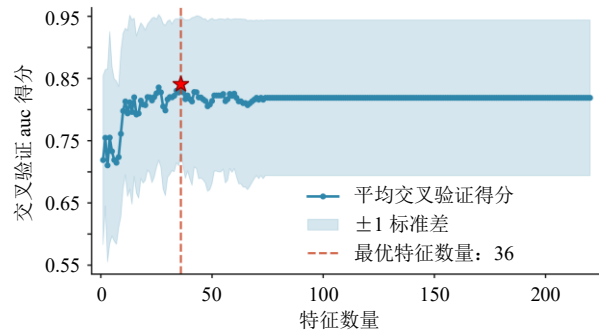


图 4 RFE CV 特征筛选过程

Fig. 4 Feature screening process of RFE CV

表 2 XGBoost 模型贝叶斯优化的超参数搜索范围

Table 2 Search range of hyperparameters for Bayesian optimization of XGBoost model

超参数名称	含义	搜索范围
learning_rate	学习率	(0.005, 0.05)
max_depth	树的最大深度	(2, 5)
subsample	训练每棵树时样本的采样比例	(0.5, 0.8)
colsample_bytree	构建每棵树时特征的采样比例	(0.5, 0.7)
n_estimators	迭代次数	(300, 2 000)
gamma	节点分裂所需的最小损失降低值	(1, 30)
reg_alpha	L1正则化项的权重	(0.1, 100)
reg_lambda	L2正则化项的权重	(10, 200)
min_child_weight	叶子节点所需的最小样本权重和	(2, 20)

表 3 特征筛选前后模型性能对比

Table 3 Comparison of model performance before and after feature screening

特征筛选	特征数	训练集				测试集			
		准确 率/%	精确 率/%	召回 率/%	F1 分数	准确 率/%	精确 率/%	召回 率/%	F1 分数
无处理	220	79.4	73.3	67.3	0.70	76.5	64.3	75.0	0.69
RFE CV	36	88.2	85.4	81.4	0.83	80.4	89.6	78.8	0.84

研究表明，当特征数量为 36 个时，模型的交叉验证性能 (AUC) 达到峰值。这表明由这 36 个特征构成的子集，在保留核心预测信息的同时，最大限度地排除了噪声干扰。基于该特征子集构建的 XGBoost 模型，与使用全部特征的模型相比表现更优，模型性能大幅提升：F1 分数由特征筛选前的 0.69 提升至 0.84，提升幅度达 21.7%。同时，模型表现出优异的泛化能力，训练集与测试集的 F1 分数差异小于 0.02，验证了模型的有效性与鲁棒性。最终建立的轻量化模型不仅参数规模缩减 83.6%，更为后续过程优化提供了可靠的特征维度依据，有助于

深入解析黄金批次的形成机制。

为进一步评估模型的泛化能力，本研究引入了一个独立的外部验证集。该验证集包含了 20 批次全新的生产数据，这些数据未参与模型的任何训练、超参数调优或特征筛选过程。将已训练的基于 36 个特征的最终 XGBoost 模型直接应用于该外部验证集，其混淆矩阵如图 5 所示。结果表明，该模型成功地正确识别了 10 个真实黄金批次中的 9 个，同时正确识别了 10 个非黄金批次中的 7 个。表明模型在面对全新的外部数据时，依然保持了强大的预测性能 ($F1=0.8182$)，展现出良好的泛化能力和在实际生产环境中应用的潜力。

真实类别	非黄金批次 (0)	7	3
	黄金批次 (1)	1	9
		非黄金批次 (0)	黄金批次 (1)
		预测类别	

图 5 外部验证集模型混淆矩阵

Fig. 5 Confusion matrix of model on external validation set

2.5 模型解释

为深入剖析黑箱模型中特征对“黄金/非黄金批次”预测的作用机制，本研究以 SHAP (Shapley 加性解释) 获得样本的可加贡献分解与全局重要性排序^[16]；以部分依赖图 (partial dependence plot, PDP) 刻画单变量或成对变量对模型输出的边际影响曲线，从而识别关键参数、阈值与调控区间，两者互证互补，提升解释的稳健性与可操作性^[17]。SHAP 分析是一种基于博弈论的模型解释技术。该方法能够针对每个特征的每个样本计算相应的 SHAP 值，此值可量化该特征对模型预测结果的贡献程度。SHAP 值的计算以 Shapley 值概念为基础，Shapley 值在博弈论中用于公平分配合作收益，而 SHAP 值则衡量了每个特征在不同特征组合情形下对模型预测结果的平均贡献。PDP 用于呈现单个特征或多个特征对模型预测结果的边际效应。其基本原理是固定其他特征的值，计算目标特征在不同取值时模型的平均预测结果，从而直观地展现该特征与模型预测结果之间的关联。

如图 6-A 所示的 SHAP 特征重要性排序表明，

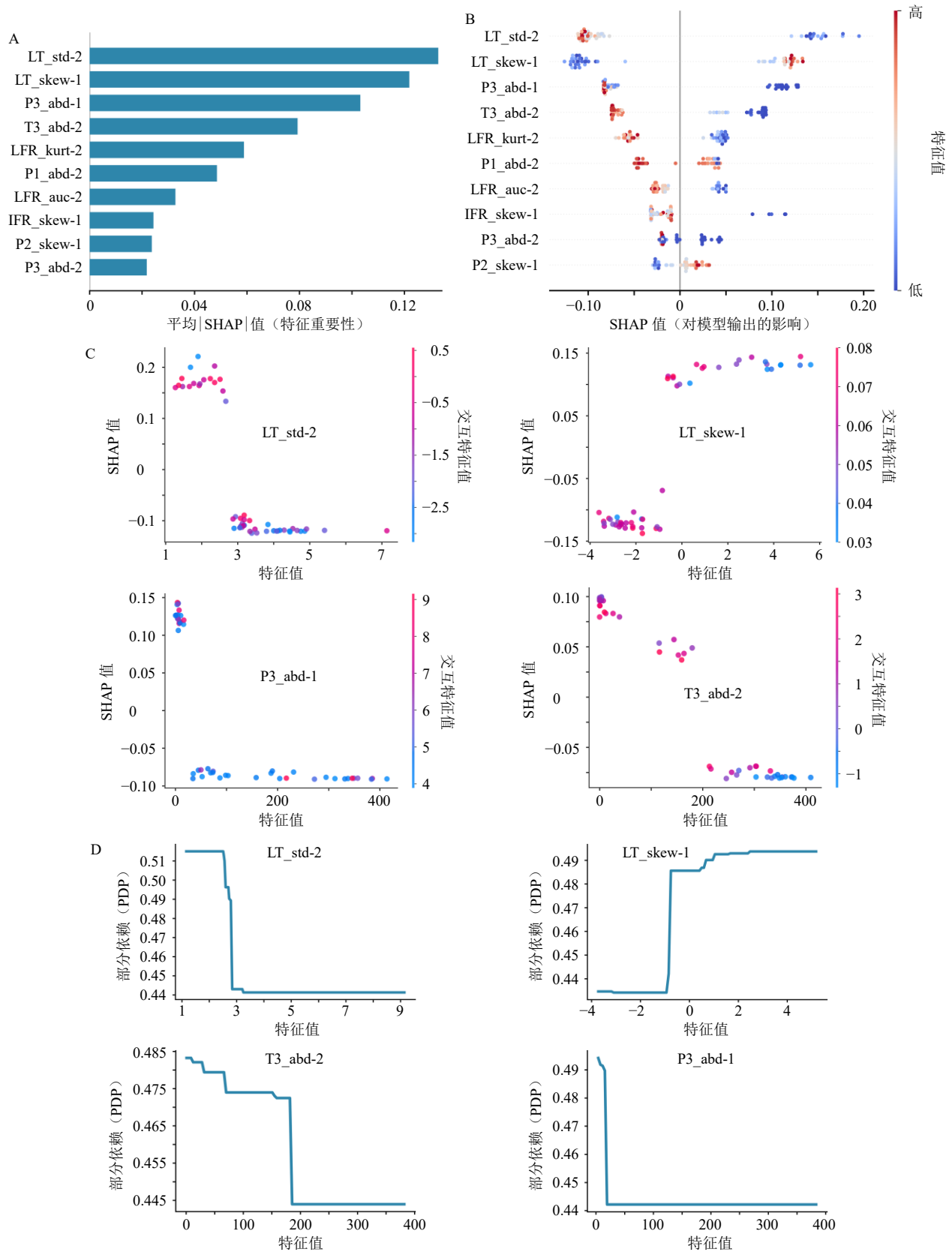
模型预测中影响力排名靠前的特征依次为 LT_std-2、LT_skew-1、P3_abd-1、T3_abd-2、LFR_kurt-2 和 P1_abd-2，这些特征在平均绝对 SHAP 值上的贡献显著高于其他特征，表明它们在区分黄金批次与非黄金批次中占据核心地位。结合图 6-B 的 SHAP 散点分布图，可以观察到不同特征值对模型输出方向的影响存在明显的分界趋势。如 LT_std-2 与 LT_skew-1 在取值较高时显著推动样本向非黄金批次偏移，而取值较低时，则更易预测为黄金批次。

结合图 6-C SHAP 依赖图和图 6-D PDP 分析结果，不难发现，核心特征与预测概率之间普遍呈现非线性变化。如 LT_std-2 的影响曲线在 2.5 °C 左右出现明显转折，高于该阈值后黄金批次形成概率明显下降；而 LT_skew-1 在 -1 附近呈现由负向影响转为正向影响的趋势。进一步地，通过 PDP 分析和 SHAP 依赖图明确各关键特征对黄金批次预测的平均影响趋势，并揭示特征间的显著交互效应，例如 LT_std-2 的正面作用在 LFR_skew-2 较低时会被放大。这些变化趋势反映出工艺过程并非非线性驱动，而是存在关键控制区间，超过或低于特定范围均可能导致批次质量下降，在实际生产中可以通过关注这些阈值点对生产过程进行调控，使其向有利于黄金批次形成的方向发展。

2.6 关键特征的差异性验证与机制分析

为确保 LGAPT 策略划分结果的有效性，并独立验证机器学习模型识别出的关键特征，本研究从统计显著性与实践重要性 2 个层面进行量化评估。首先，对 2 组样本数据 (黄金/非黄金批次) 进行 Shapiro-Wilk 正态性检验，鉴于工业过程数据通常呈现非正态分布且样本量有限，采用传统的参数检验易引入判定偏差。因此，本研究选用更为稳健的非参数方法 Mann-Whitney U 检验评估组间特征分布的差异显著性 (P 值)^[18]。同时，为具体衡量差异的实际大小并消除样本量的干扰，引入效应量 (effect size) 分析，采用等级双列相关系数 (rank-biserial correlation) 进行标准化量化^[19]。结果如图 7 所示。结果表明，由 SHAP 模型识别出的 Top-10 关键特征，如 T3_abd-2、IFR_mean-2 等，均展现出中等至较大 ($|d| > 0.5$) 的效应量，这有力地证明了模型所识别出的重要特征，在统计学上确实是区分黄金与非黄金批次的核心差异点。

此外，效应量的方向性 (正或负) 与“2.5”项下的 SHAP 特征散点分布图中揭示的影响方向 (促



A-SHAP 特征重要性图；B-SHAP 特征散点分布图；C-SHAP 特征依赖图；D-PDP 部分依赖图。

A-SHAP feature importance diagram; B-SHAP characteristic scatter map; C-SHAP feature dependence graph; D-partial dependence graph of PDP.

图 6 基于 SHAP 和 PDP 的模型解释结果

Fig. 6 Model interpretation results based on SHAP and PDP

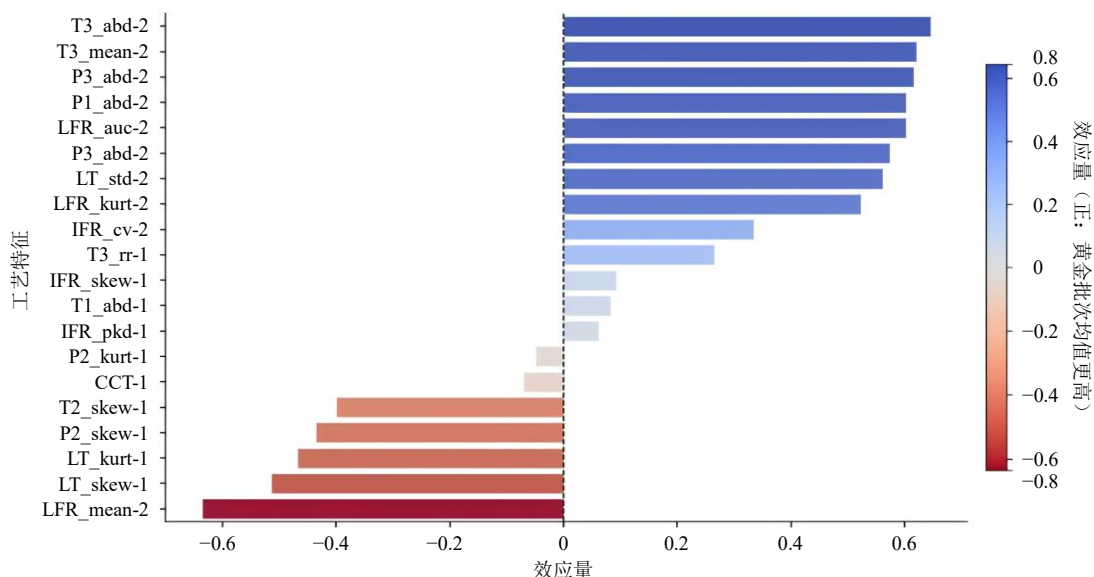


图 7 效应量分析结果

Fig. 7 Results of effect size analysis

进或抑制)表现出高度一致性。例如, IFR_mean-2 呈现显著的正效应, 表明黄金批次在该特征上的均值显著高于非黄金批次, 印证了 SHAP 分析中“高值促进”的结论。反之, T3_abd-2 呈现显著的负效应, 其黄金批次均值远低于非黄金批次, 再次确认了 SHAP 揭示的“低值促进”作用。分类预测模型结果与统计事实的高度一致性, 充分证明了本研究发现的可靠性与稳健性。

在量化差异的基础上, 为从数据分布的视角直观洞察这些特征影响批次性能的内在模式, 本研究进一步采用核密度估计(kernel density estimation, KDE)对效应量最显著的4个正、负向关键特征进行可视化分析^[20]。KDE能够通过拟合数据的概率密度函数, 精细展示特征在不同批次间的分布形态。

如图8所示, 黄金批次(黄色)与非黄金批次(棕色)的分布曲线形成了清晰的分离, 生动揭示了关键特征的作用机制。以代表第2亚批的T3动态偏离时长的特征“T3_abd-2”为例, 黄金批次的分布曲线整体显著位于非黄金批次左侧, 其均值(75.81 min)远低于非黄金批次(247.17 min)。这一显著的差异直观表明, 将该参数严格限制在一个较低的操作窗口内, 是形成黄金批次的决定性条件, 这也与生产实际相吻合, 即减小T3的动态偏离时长有利于黄金批次的形成。

3 讨论

本研究聚焦于RI金银花浓缩工序, 旨在深入探究影响黄金批次形成的关键因素, 并构建有效的

分类预测模型。本研究收集RI金银花浓缩工序的历史生产工艺参数170批, 每批包含2个亚批数据, 通过时序数据特征提取技术, 将原始时序数据转化为固定维度特征向量矩阵(170×220), 创新性地使用E作为工艺性能评价标准, 将LGAPT策略应用于黄金批次的划分, 实现了黄金批次的客观、数据驱动划分, 避免了人为设定阈值带来的偏倚, 最终确定黄金批次分界阈值为0.1363 kg/(m³·min), 该阈值将170批生产批次划分为黄金批次61批、非黄金批次109批。后续采用XGBoost算法构建黄金批次与非黄金批次的分类预测模型, 模型效果良好, 泛化能力强, 进一步通过SHAP分析结合PDP分析对模型进行深度解释, 确定LT_std-2、LT_skew-1、P3_abd-1、T3_abd-2、LFR_kurt-2和P1_abd-2等特征为影响黄金批次的关键特征, LT为影响黄金批次的关键工艺参数, 提示应当保持在LT_std-2 < 2.5 °C, LT_skew-1 > -1的水平, 有利于黄金批次的形成, 这一结论为实际生产中的工艺优化提供了明确的指导方向。最后采用效应量分析和核密度估计, 验证了模型解释结果的可靠性与稳健性。

本研究识别出的关键工艺特征, 与保障RI的药效物质基础具有内在的逻辑一致性。RI“清热解毒”的功效, 在很大程度上来源于金银花中的绿原酸等热敏性成分, 本研究的核心发现是, 过程的稳定性(以更低的LT_std-2为代表)是黄金批次形成的关键。从传热传质学角度分析, 一个稳定、高效的浓缩过程, 能显著减少物料的总受热时间并避免

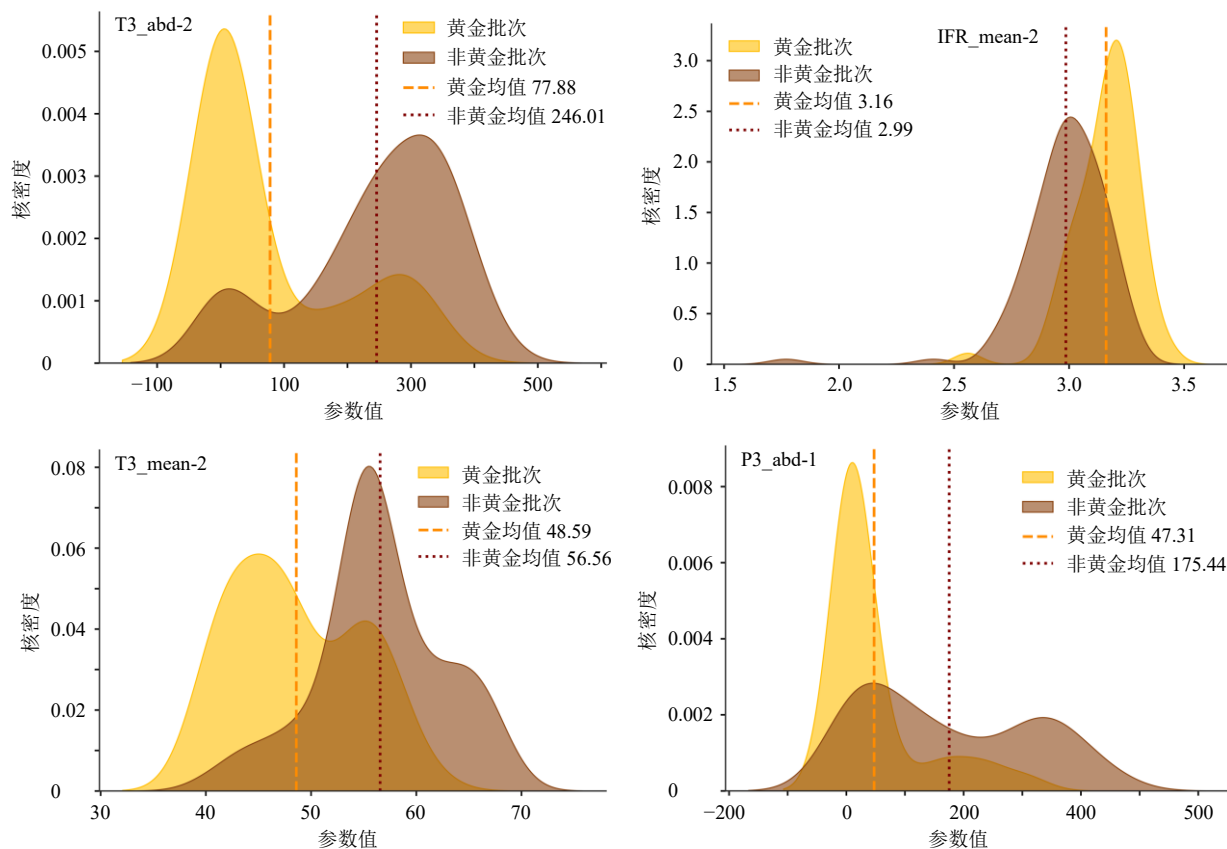


图8 核密度估计结果

Fig. 8 Results of kernel density estimation

局部过热，这在机制上是最大化保留热敏性成分、抑制其降解或异构化的必要条件。因此，可以合理推断，本研究筛选出的以“过程平稳、传热高效”为特征的黄金批次，在客观上有利于保障金银花中热敏性有效成分的稳定性与转移率。

尽管本研究取得了显著的成果，但仍存在一些不足之处。虽然研究寻找到了关键工艺参数与关键特征，并确定其控制区间，在一定程度上回答了“控什么、怎么控”的问题，但关于“控多少”这一更精细的问题，仍有待深入探讨，需要在后续深入研究控制参数与状态参数的相互作用和因果关系，逐步细化为更加定量、可操作的控制策略。此外，应推动本研究成果向在线应用的转化，建立基于关键特征的黄金轨迹实时监控与预警系统，实现对生产过程偏离的早期识别，并在此基础上，进一步构建基于模型的推荐与决策支持系统，最终实现对生产过程的实时监控与动态优化。

总体而言，本研究从复杂多变的浓缩过程数据中，提炼出少量具有清晰物理含义、又便于现场理解和操作的关键特征，确定的关键特征和工艺参数

为企业优化生产工艺提供了具体的方向和目标，实现了从“经验驱动”到“数据驱动”的转变。此外，这一过程也探索了将原始生产数据转化为可用于高级分析的高质量数据集的方法。通过构建 E 这一指标对生产效率进行量化，并利用 LGAPT 策略对批次进行客观的优劣等级标注，成功将无标签的过程数据转化为一个包含明确业务含义（即“黄金批次”）的结构化数据集，为后续开发更复杂的预测与优化模型积累了高质量的数据基础，有助于推动生产过程的智能化进程。

实践表明本研究形成的“效率量化-智能划分-模型寻优”整体框架，已经在金银花浓缩工序中得到了成功应用，具有较强的可移植性，未来可推广至 RI 的其他生产环节，如提取、醇沉、萃取等工序，最终构建覆盖全生产链的智能质量控制体系，为中药制剂的稳定生产和质量提升提供更加有力的技术支撑。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 中华人民共和国工业和信息化部. 智能工厂梯度培育

- 行动取得初步成效 [EB/OL]. (2025-02-07) [2025-02-08]. https://www.miit.gov.cn/jgsj/zbys/gzdt/art/2025/art_4341c71b91d64262b7d205be011c7992.html.
- [2] 谭琳, 万鑫浩, 王学成, 等. 新质生产力视角下的中药绿色数字化制造实现路径研究 [J]. 中草药, 2025, 56(5): 1782-1792.
- [3] 余俭. 抗菌抗病毒新药: 热毒宁注射液 [J]. 中南药学, 2010, 8(7): 548-550.
- [4] 童枫, 徐芳芳, 闫逸伦, 等. 热毒宁注射液金银花和青蒿(金青)萃取过程中固形物含量近红外光谱在线监测模型的建立及萃取终点判断研究 [J]. 中草药, 2024, 55(19): 6555-6565.
- [5] Tu Y J, Li L N, Wang Z T, *et al.* Advances in analytical techniques and quality control of traditional Chinese medicine injections [J]. *J Pharm Biomed Anal*, 2021, 206: 114353.
- [6] Redchuk A, Walas Mateo F. Industry 5.0 and digital twins in the chemical industry: An approach to the golden batch concept [J]. *Chem Engineering*, 2025, 9(4): 78.
- [7] 赵雅兰. 非高斯批次过程控制及性能评价研究 [D]. 太原: 太原理工大学, 2019.
- [8] Bush X, Fratz-Berilla E J, Kohnhorst C L, *et al.* Defining golden batches in biomanufacturing processes from internal metabolic activity to detect process changes that may affect product quality [J]. *Biotechnol Bioeng*, 2025, 122(2): 298-305.
- [9] 黄一躬. 数据驱动的高强钢高速干铣表面粗糙度预测与工艺参数优化 [D]. 重庆: 重庆大学, 2023.
- [10] 蒋锦. 三维荧光光谱结合机器学习实现黑茶品牌及工艺参数区分 [D]. 长沙: 中南大学, 2023.
- [11] 陈俊佟. t-SNE 结合支持向量机的降维分类研究 [D]. 大连: 大连理工大学, 2021.
- [12] 朱卫坪, 陈晓峰, 张萍, 等. 基于 t-SNE 降维与 k-means 聚类算法的化工厂节能减排分析 [J]. 自动化应用, 2022, 63(1): 10-13.
- [13] 南淑荷, 李进军, 魏佳芳, 等. 基于 XGBoost-SHAP 模型的流域水质指标对 DO 的驱动与协同影响分析 [J]. 水电能源科学, 2025, 43(7): 52-56.
- [14] 刘文新, 徐文辉, 陈朝晖, 等. 基于 RFECV-XGBoost 和 SHAP 的火电厂电力输灰预测模型 [J]. 计算机与现代化, 2025(4): 63-69.
- [15] 常家康, 吕宁, 詹跃东. 基于 XGBoost-RFECV 算法和 LSTM 神经网络的 PEMFC 剩余寿命预测 [J]. 电子测量与仪器学报, 2022, 36(1): 126-133.
- [16] 袁溢, 潘从元, 章新宇, 等. 基于 SHAP 可解释性分析与多特征融合的焦粉水分识别模型 [J]. 中国冶金, 2025, 35(7): 189-201.
- [17] 王鹏新, 王颖, 田惠仁, 等. 基于 LightGBM 的冬小麦产量估测与可解释性研究 [J]. 农业机械学报, 2023, 54(12): 197-206.
- [18] 于昆. 基于 Mann-Whitney 算法的非参数控制图 [D]. 天津: 天津大学, 2009.
- [19] 方杰, 温忠麟. 基于两水平回归模型的调节效应分析及其效应量 [J]. 心理科学进展, 2022, 30(5): 1183-1190.
- [20] 杨瑞君, 张楚, 杨评, 等. 非参数核密度估计模型预测双酚 A 的物种敏感度分布规律 [J]. 生态毒理学报, 2024, 19(4): 120-130.

[责任编辑 郑礼胜]