基于改进随机森林算法的工业过程运行状态评价

常玉清 [1] 孙雪婷 [1] 钟林生 [1] 王福利 [1, 2] 刘英娇 [1]

摘 要 运行状态评价是指在过程正常生产的前提下, 进一步判断生产过程运行状态的优劣. 针对复杂工业过程定量信息与定性信息共存的情况, 本文提出了一种基于随机森林的工业过程运行状态评价方法. 针对随机森林中决策树信息存在冗余的问题, 基于互信息将传统随机森林中的决策树进行分组, 并选出每组中最优的决策树组成新的随机森林. 同时为了强化评价精度高的决策树和弱化评价精度低的决策树对最终评价结果的影响, 使用加权投票机制取代传统众数投票方法, 最终构成一种基于互信息的加权随机森林算法 (Mutual information weighted random forest, MIWRF). 对于在线评价, 本文通过计算在线数据处于各个等级的概率, 并且结合提出的在线评价策略, 判定当前样本运行状态等级. 为了验证所提算法的有效性, 将所提方法应用于湿法冶金浸出过程, 实验结果表明, 相对于传统随机森林算法, MIWRF 降低了模型的复杂度, 同时提高了运行状态评价精度.

关键词 湿法冶金, 运行状态评价, 互信息, 加权随机森林

引用格式 常玉清, 孙雪婷, 钟林生, 王福利, 刘英娇. 基于改进随机森林算法的工业过程运行状态评价. 自动化学报, 2021, **47**(9): 2214−2225

DOI 10.16383/j.aas.c190066

Industrial Operation Performance Evaluation of Industrial Processes Based on Modified Random Forest

CHANG Yu-Qing[1] SUN Xue-Ting[1] ZHONG Lin-Sheng[1] WANG Fu-Li[1, 2] LIU Ying-Jiao[1]

Abstract Operation performance evaluation refers to further judging the operation performance of process on the premise of normal production. In the view of coexistence of qualitative information and quantitation information during the industrial processes, a method of industrial operation performance evaluation of industrial processes based on modified random forest is proposed. In order to solve the problem of redundancy of decision trees information in random forest, decision trees are grouped based on mutual information, and the optimal decision tree in each group is selected to form a new random forest. Meanwhile, in order to strengthen the decision tree with high evaluation accuracy and weaken the decision tree with low evaluation accuracy, weighted voting mechanism are proposed to replace the traditional mode voting, and finally a mutual information weighted random forest (MIWRF) based on mutual information is formed. To verify the proposed method, the method is applied to hydrometallurgical leaching process. The result shows that MIWRF reduces the complexity of the model and improves the accuracy of operation performance evaluation compared with the traditional random forest algorithm.

Key words Hydrometallurgy, operation performance evaluation, mutual information, weighted random forest

Citation Chang Yu-Qing, Sun Xue-Ting, Zhong Lin-Sheng, Wang Fu-Li, Liu Ying-Jiao. Industrial operation performance evaluation of industrial processes based on modified random forest. *Acta Automatica Sinica*, 2021, **47**(9): 2214−2225

收稿日期 2019-01-27 录用日期 2019-09-09
Manuscript received January 27, 2019; accepted September 9, 2019
国家自然科学基金 (61673092, 61533007, 61304121, 61973057, 61873053), 创新研究群体科学基金 (61621004), 中央高校基础科研业务费 (N150404017), 矿冶过程自动控制技术国家重点实验室开放基金 (BGRIMM-KZSKL-2018-08) 资助
Supported by National Natural Science Foundation of China (61673092, 61533007, 61304121, 61973057, 61873053), Science Foundation for Innovative Research Groups (61621004), Fundamental Research Funds for the Central Universities (N150404017), and Open Foundation of State Key Laboratory of Process Automation in Mining and Metallurgy (BGRIMM-KZSKL-2018-08)
本文责任编委 伍洲
Recommended by Associate Editor WU Zhou
1. 东北大学信息科学与工程学院 沈阳 110819 2. 流程工业综合自动化国家重点实验室 (东北大学) 沈阳 110819
1. College of Information Science and Engineering, Northeastern University, Shenyang 110819 2. State Key Laboratory of Synthetical Automation for Process Industries (Northeastern University), Shenyang 110819

工艺操作的安全性和优化性是工业界近几十年来备受学术界和工业界关注的两个关键问题[1]. 然而, 由于扰动、噪声等不确定因素的存在, 安全性和最优性都可能发生恶化, 这会导致运行性能的下降. 因此, 及时、准确地掌握过程运行状态, 对于提高企业经济效益、生产效益和产品质量都有重要意义. 而传统的过程检测仅仅是监测生产过程是否正常, 对于异常的生产过程剖析原因通过调整使生产继续正常运行. 为了获得更高的经济效益, 需要得到整

个工艺过程的优劣状态, 因此相关学者提出了运行状态评价的概念[2−3]. 运行状态评价是指在生产过程正常运行的前提下, 对实际生产过程的运行优劣进行识别与判断, 当运行状态处于 “非优”时, 通过及时调整操作, 使运行状态达到 “优”. 因此, 对运行状态评价的研究具有重要的理论意义和应用价值.

目前, 针对运行状态评价学者们进行了一定的研究. 案例推理[4] 方法评价速度快, 且推理评价过程包含了学习和归纳, 有一定实际意义. 粗糙集理论[5] 能够有效地分析和处理不完备信息, 并从中发现隐含的知识, 揭示潜在规律. 它的根本思想是在保持条件属性相对于决策属性不变的前提下, 通过属性约简和值约简等过程, 挖掘数据内涵的规则, 在对定性和定量变量混合的数据处理上具有较好的效果. 人工神经网络[6] 学习能力强, 能够表达非线性的映射关系, 在非线性过程的运行状态评价中有应用. 针对小样本情况, 研究者提出灰色关联分析法[7]. 模糊评价[8] 是最常用的评价方法之一, 它既符合决策过程中信息的可用性和不确定性, 又符合人的认知的模糊性. 针对多模态过程运行状态评价, 邹筱瑜等[9] 提出了基于高斯混合模型的评价方法, 确保特征提取的准确性, 避免了模态划分问题. 文献 [10] 提出了多种工业过程运行状态评价方法, 但其是基于定量变量, 对于定性变量的处理并未提及.

但在实际生产中, 复杂工业过程运行环境差、检测技术不完善, 导致过程定量信息与定性信息共存, 限制了传统的运行状态评价方法的应用. 其中, 定量信息指用数值大小描述的变量信息, 定性信息指通过语义定性描述的变量信息. 本文通过对复杂工业过程进行深入研究, 提出了基于随机森林的运行状态评价方法, 随机森林的特点是能够同时处理定量信息和定性信息, 且无需对数据进行复杂的预处理.

为了提升随机森林的性能, 学者们提出了多方面的改进方案. 文献 [11] 通过操作训练集和训练特征定义了一组新的随机森林分类器. Tuv 等[12]、Paul 等[13] 提出通过特征选择来去除不重要特征、减小特征冗余以减小特征空间. 文献 [14] 提出了随机森林中一种有效的聚合方法—交替决策森林, 该方法将随机森林的训练归为全局损失最小化问题. 在投票环节, 提出一种加权投票方法为不同决策树赋予不同的投票权值[15]. 但是针对决策树之间的冗余问题, 以上方法未有提及. 为了解决该问题, 本文提出一种基于改进的随机森林运行状态评价算法, 采用互信息计算随机森林中不同决策树之间的相关性以及每棵决策树的评价精度, 剔除相关性大且精度低的决策树. 同时, 为了解决投票权重问题, 将评价精度转换成决策树的权重, 增加精度高的决策树投票权重, 降低精度低的决策树投票权重, 进而提高随机森林的评价精度.

本文主要以湿法冶金的氰化浸出过程为研究对象, 分别利用传统的随机森林算法和改进的随机森林算法建立运行状态评价模型, 通过对评价结果的对比分析, 验证了所提方法的正确性和有效性.

1　随机森林

1.1　随机森林定义及性质

随机森林 (Random forests, RF) 算法是 Breiman 于 2001 年提出的一种分类和预测算法, 其本质是将 Bagging 算法和随机子空间算法结合起来[16−18]. 随机森林是一种以决策树为基分类器的集成学习模型[19]. 它是利用 bootstrap 抽样方法生成多个训练集, 针对每个训练集建立一个决策树模型, 然后通过投票方式集成所有决策树的预测结果. RF 具有很高的预测准确率, 对异常值和噪声具有很好的容忍度[20], 且不容易出现过拟合, 并且训练速度较快, 目前已广泛应用于各种分类及预测问题.

定义 1. 随机森林是一组由多个决策树分类器组成的集成分类器, 随机向量独立同分布. 当输入待分类样本后, 每个决策树分类器通过简单投票 (即少数服从多数) 得出分类结果.

1.2　随机森林的生成

随机森林算法采用 bootstrap 方法从 N 个训练样本中有放回的抽取 n $(n \le N)$ 个样本, 重复 K 次, 生成 K 棵决策树. 在对决策树每个节点分裂时, 从全部 M 个属性中随机抽取 m 个属性 (通常取小于等于 $\log_2(M)+1$ 的最大正整数), 再从 m 个属性中选择最优属性作为分裂属性. 对于分裂属性的选择, 随机森林应用了 CART 算法. 该算法使用基尼指数来选择划分属性. 假设当前样本集合 D 中第 i 类样本所占比例为 p_i $(i=1,2,\cdots,|y|)$, 其中, $\sum_{i=1}^{|y|} p_i = 1$, $|y|$ 表示分类类别总数, 则 D 的基尼值为

$$Gini(D) = 1 - \sum_{i=1}^{|y|} p_i^2 \tag{1}$$

其中, $Gini(D)$ 取值范围为 $[0, 1]$.

用测试属性 a 将随机变量 D 二元划分为 D_1 和 D_1 两类, 则属性 a 的基尼指数为

$$Gini(D,a)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \quad (2)$$

假设候选属性集合为 $A=a_1,a_2,\cdots,a_m$, 则选择使得划分后基尼指数最小的属性作为最优划分属性, 即

$$a^*=\text{argmin}(Gini\_index(D,a)) \quad (3)$$

2 基于改进随机森林的工业过程运行状态评价

工业过程运行状态评价是指在生产过程正常运行的前提下, 通过一定的方法对实际生产过程的运行状态进行识别与判断, 当运行状态处于 "非优" 时, 及时调整生产操作, 以使状态达到最佳. 其实质是通过学习历史数据, 对在线数据的运行状态进行分类评价.

在用传统随机森林算法进行运行状态评价时, 由于其会生成相似度较高的决策树, 造成模型冗余, 同时在投票环节, 所有决策树权重相同, 忽略了不同决策树的性能差异. 基于上述问题, 本文提出了一种基于互信息的加权随机森林算法 (Mutual information weighted random forest, MIWRF). 利用互信息计算任意两棵决策树的相关性, 对于相关性较大的决策树, 只保留评价精度最高的决策树, 从而形成新的随机森林, 并将评价精度转化为投票权重, 最终得到冗余更小, 评价精度更高的随机森林模型.

2.1 改进的随机森林算法

在文献 [21] 中, Krogh 和 Vedelsby 提出了一种误差−分歧分解规则, 该规则表明个体学习器准确性越高, 多样性越大, 则集成效果越好. 基于该结论, 我们提出用互信息来度量个体学习器的多样性和准确度.

2.1.1 算法描述

在信息论中, 互信息用来衡量两个变量之间的相互依赖关系[22], 换而言之, 它表示一个随机变量中包含另一个随机变量的信息量. 对于两组给定随机变量 X , Y, 它们的互信息表示为

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=H(X)+H(Y)-H(X,Y) \quad (4)$$

其中, $p(x,\ y)$ 为 X , Y 的联合概率分布, $p(x)$, $p(y)$ 分别为 X , Y 的边缘概率分布, $H(X)$ 是 X 的信息熵, 其计算式为

$$H(X)=-\sum_{x\in X}p(x)\log p(x) \quad (5)$$

其中, $p(x_i)$ 表示事件 x_i 发生的概率; $H(Y)$ 是 Y 的信息熵, $H(X,Y)$ 是联合熵, 其计算式为

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y) \quad (6)$$

当变量 X 和 Y 完全独立时, 互信息最小, 结果为 0, 说明两个变量之间不包含重复信息. 反之, 互信息越大, 两个变量的相互依赖性越大, 即两个变量之间的重复信息越多.

对于随机森林中的决策树 h_i $(i=1,2,\cdots,K)$, $I(h_i,h_k)$ $(k\neq i)$ 表示决策树 h_i 与 h_k 的互信息. 本文采用 $I(h_i,h_k)$ $(k\neq i)$ 来计算决策树 h_i 与 h_k 之间的相关性, 即重合度. 其计算式为

$$I(h_i,h_k)=I(y_i,y_k) \quad (7)$$

其中, y_i $(i=1,2,\cdots,K)$ 为第 i 棵决策树的输出状态. $I(h_i,h_k)$ 的值越大, 说明两棵决策树的相关性越大, 所描述的信息的重合度越高. 通过计算任意两棵决策树的互信息, 将互信息值大于阈值 ε 的决策树合为一组. $I(h_i,y$ 表示决策树 h_i 与实际标签 y 的互信息, 即决策树 h_i 的输出评价结果与实际评价结果之间的相关性. 其计算式为

$$I(h_i,y)=I(y_i,y) \quad (8)$$

$I(h_i,y)$ 的值越大, 说明决策树 h_i 的评价精度越高. 最后将相关性较小、精度较高的决策树组成新的随机森林.

算法具体过程如下:

步骤 1. 获取训练集 $D=\{(\boldsymbol{x}_1,y_1)$, $(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,\ y_N)\}$, 验证集 $T=\{(\boldsymbol{x}_1,y_1)$, $(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_L,\ y_L)\}$, $\boldsymbol{x}_i(1\times M)$ 为评价属性, $y_i(1\times 1)$ 为评价标签.

步骤 2. 通过 Bootstrap 抽样从 D 中抽取 n ($n\leq N$)个样本, 重复 K 次, 得到 K 个训练集.

步骤 3. 对于每一棵决策树的内部节点, 从样本 M 个属性中随机抽取 m 个属性, 选择 m 个属性中的最优分裂属性作为该节点的测试属性.

步骤 4. 生成 K 棵完全随机决策树, 组成随机森林 $R=\{h_1,h_2,\cdots,h_K\}$.

步骤 5. 将验证集 T 输入 R 中得到评价结果.

步骤 6. 根据评价结果, 依次计算每棵决策树与其余决策树的相关性. 将所有 $I(h_i,h_k)$ $(k\neq i)$ 大于阈值 ε 的决策树合为一个决策树组. 若 h_i 与其余所有决策树相关性均小于等于阈值 ε , 则该决策树单独作为一组.

步骤 7. 再将组外决策树按照顺序重复步骤 6,

直至全部决策树分组完成.

步骤 8. 根据精度 $I(h_i, y)$ 获取每组中精度最高的决策树.

步骤 9. 将所获取的决策树组成新的随机森林 $R' = \{h'_1, h'_2, \cdots, h'_P\}$.

2.1.2　加权投票

对输入的样本数据进行评价时, 传统的随机森林在投票时每棵决策树的投票权重相同, 忽略了不同决策树的评价精度对最终结果产生的影响[23], 降低了随机森林整体的评价精度.

为了增加评价精度高的决策树和降低评价精度不高的决策树对最终评价结果的影响, 本文提出了加权投票方法, 将评价精度转化为决策树的投票权值. 在传统随机森林精简后, 新随机森林中的决策树评价精度矩阵 ACC 为[15]

$$ACC = \begin{bmatrix} acc_{11} & acc_{12} & \cdots & acc_{1P} \\ acc_{21} & acc_{22} & \cdots & acc_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ acc_{Q1} & acc_{Q2} & \cdots & acc_{QP} \end{bmatrix} \tag{9}$$

其中, acc_{qp} 表示第 p 棵决策树对于第 q 种运行状态的评价精度, 其中, $p = 1, 2, \cdots, P$, $q = 1, 2, \cdots, Q$, Q 为评价结果的等级数目, P 为改进后决策树数量. 通过将验证集 T 代入新的随机森林, 并计算每棵决策树对于每一类的输出结果的正确率获得.

根据评价精度矩阵, 定义权重矩阵 W 为

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1P} \\ w_{21} & w_{22} & \cdots & w_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ w_{Q1} & w_{Q2} & \cdots & w_{QP} \end{bmatrix} \tag{10}$$

其中, w_{qp} 表示第 p 棵树对于第 q 种运行状态的权重, 其计算式为

$$w_{qp} = acc_{qp} \tag{11}$$

基于互信息的加权随机森林算法 (MIWRF) 的具体流程如图 1.

2.2　运行状态评价模型的离线建模

由于随机森林能够同时处理定量和定性信息, 同时无需对正常数据进行复杂的预处理. 在获取离线建模数据后, 生成包含 K 棵决策树的初始随机森林评价模型 R. 同时用验证集 T 精简初始模型. 根据第 2.1 节所提方法, 计算任意两棵决策树的相关性 $I(h_i, h_k)$ $(k \neq i)$, 将相关性大于阈值 ε 的决策树合为一个决策树组, 再计算组内决策树的评价精度 $I(h_i, y)$, 留下组内精度最高的决策树, 生成包含棵 P 决策树的新的随机森林评价模型 R'. 在投票

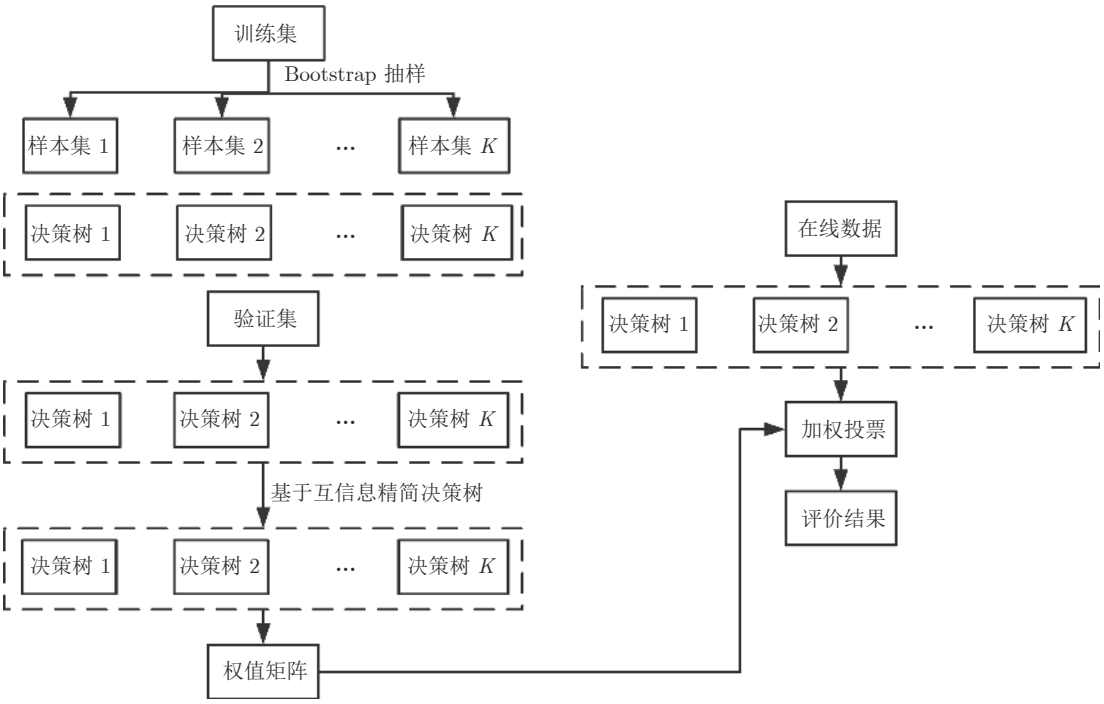

图 1　基于互信息的加权随机森林算法

Fig. 1　Weighted random forest algorithm based on mutual information

环节, 以评价精度为基础生成随机森林的权重矩阵 W.

2.3 在线评价

假设建模数据充分, 评价结果分为 Q 个等级. 将 t 时刻在线数据 x_t 输入评价模型 R' 得到每棵决策树的评价结果 $y_i\ (i=1,2,\cdots,P)\in\{1,2,\cdots,Q\}$, 则 x_t 处于第 q 个状态等级的概率为

$$P(q|x_t)=\frac{\sum\limits_{p}\|(y_i=q)\,w_{qp}}{\sum\limits_{q}\sum\limits_{p}\|(y_i=q)\,w_{qp}} \tag{12}$$

其中, $\|(f)$ 为指示函数, 表示当 f 为真时, 其值为 1, 否则为 0.

数据 x_t 的最大后验概率对应的等级为

$$q_t^*=\arg\left\{\max_{q=1,2,\cdots,Q}P(q|x_t)\right\} \tag{13}$$

为了减少噪声引起的错误评价, 提出一种在线评价策略. 假设运行状态等级只能在相邻等级之间转换, t 时刻的评价结果为 q_t^*. 运行状态等级在线评价遵循以下 2 条规则:

1) 只有连续 H 个样本点最大后验概率对应的等级 $q_{t-H+1}^*,q_{t-H+1}^*,\cdots,q_t^*$ 都与 $t-H$ 时刻的评价结果 q_{t-H}^* 不同, 才认为运行状态等级可能发生转换: 否则, 保持评价结果不变, 记为 $q_t^*=q_{t-1}^*$. 当评价结果个数小于 H 时, 最终评价等级 q_t^* 等于即时评价等级.

2) 根据规则 1) 判断出运行状态等级转换之后, 记 $q_{t-H+1}^*,q_{t-H+1}^*,\cdots,q_t^*$ 中频率最高的等级为 $\underset{\sim}{q}_t^*$. 如果 $\underset{\sim}{q}_t^*$ 是 q_{t-1}^* 的相邻等级, 那么 $\underset{\sim}{q}_t^*$ 则为运行状态等级评价结果, 记为 $q_t^*=\underset{\sim}{q}_t^*$; 否则, 在当前运行状态等级为 q_{t-1}^* 的两个相邻等级中, 更接近 $\underset{\sim}{q}_t^*$ 的等级为运行状态等级评价结果 q_t^*.

通过以上评价策略, 确定最终评价等级为 q_t^*.

3 金湿法冶金浸出过程的运行状态评价

3.1 湿法冶金浸出过程

湿法冶金指使用一定成分的无机水溶剂或有机溶剂与经选矿富集的精矿相接触[24], 通过化学反应使矿石中的有用金属转入溶液中, 再从溶液分离富集所含的金属离子, 最后以单质或其化合物的形式提取的方法. 本文以高铜线为研究对象, 研究其氰化浸出过程. 氰化浸出过程是湿法冶金中比较重要的一步, 浸出的好坏对黄金的产量和综合经济效益的高低有直接影响, 因此对浸出过程进行运行状态评价具有重要意义.

氰化浸出是指将氰化钠溶液与含有待提取金属的矿石进行化学反应, 提取其中的有价金属或其化合物. 浸出工序包含四个浸出槽, 分别向第 1、2 和 4 浸出槽内加入氰化钠溶剂, 并向每个浸出槽中通入空气以使氧气溶于矿浆中并搅拌矿浆, 使金单质与氰化钠充分反应, 最终以金氰离子形式存于液相中. 理论上氰化浸出的化学方程式如式 (14). 为了防止氰化钠被水解产生剧毒气体 HCN 或者被二氧化碳分解消耗, 常用石灰乳作为保护碱.

$$\begin{aligned}4\mathrm{Au}+8\mathrm{NaCN}+2\mathrm{H_2O}+\mathrm{O_2}=\\ 4\mathrm{NaAu(CN)_2}+4\mathrm{NaOH}\end{aligned} \tag{14}$$

3.2 运行状态影响因素分析

在氰化浸出过程中, 影响因素众多, 对可能影响浸出结果的因素分析如下[25−28].

1) NaCN 添加量和浓度

氰化钠添加量和浓度会影响浸出过程的反应速率和浸出率的高低, 氰化钠的添加量直接影响溶液中氰化钠的浓度. NaCN 浓度过低, 金的溶解速率较低, 会使得一定时间内的反应不充分; NaCN 浓度过高, 会增加生产成本. 在实际生产中, NaCN 浓度固定不变, 因此不考虑 NaCN 浓度对浸出率影响; NaCN 添加量是可以在线获得的定量变量.

2) 搅拌强度和氧浓度

湿法冶金的浸出过程主要通过通入空气对浸出槽内的液体进行搅拌, 使其充分反应, 提高反应速率, 其在充入空气的同时也带入了反应所需的氧气. 在这里, 用空气流量大小来体现搅拌强度和氧浓度. 空气流量是可以在线测量的定量变量.

3) 初始金品位和矿浆浓度

初始金品位和矿浆浓度直接影响着整个浸出过程的运行状况及浸出率的高低. 矿浆浓度直接决定其中的反应物的扩散速度, 初始金品位影响着金的浸出量. 在实际生产过程中, 受生产条件的限制, 这 2 个变量无法在线获得, 只能由专家给出定性估计.

3.3 运行状态评价仿真实验

综合第 3.2 节影响因素分析, 本文以氰化浸出过程的浸出效果作为评价指标, 选取 19 个与该指标密切相关的过程变量, 列于表 1 中. 氰化浸出过程是一个定性信息与定量信息共存的过程, 根据评价指标及该过程对运行状态评价的具体要求, 将浸

表 1　浸出过程变量列表

Table 1　Key variables affecting leaching efficiency

分割方法	单位	属性
矿石初始来料量	kg	定量
浸出调浆后矿浆浓度	%	定量
矿石初始金品位	g/t	定性
一浸浸出槽1氰化钠添加量	kg/h	定量
一浸浸出槽2氰化钠添加量	kg/h	定量
一浸浸出槽4氰化钠添加量	kg/h	定量
一浸浸出槽1槽空气流量	m^3/h	定量
一浸浸出槽2槽空气流量	m^3/h	定量
一浸浸出槽3槽空气流量	m^3/h	定量
一浸浸出槽4槽空气流量	m^3/h	定量
二浸前矿浆浓度	%	定量
一浸后金品位	g/t	定性
二浸浸出槽1氰化钠添加量	kg/h	定量
二浸浸出槽2氰化钠添加量	kg/h	定量
二浸浸出槽4氰化钠添加量	kg/h	定量
二浸浸出槽1槽空气流量	m^3/h	定量
二浸浸出槽2槽空气流量	m^3/h	定量
二浸浸出槽3槽空气流量	m^3/h	定量
二浸浸出槽4槽空气流量	m^3/h	定量

出过程的运行分为优、次优和非优三个等级. 本节将所提方法应用于我们课题组开发的金湿法冶金半实物仿真平台中, 此仿真平台模拟了所研究的金湿法冶金生产过程. 经过长时间的实践、修正和完善, 此平台可以较为准确地模拟该湿法冶金生产过程, 为实际生产决策提供参考. 从金湿法冶金仿真平台采集 5 000 组训练数据, 其中, 4 000 组数据作为训练集用来建立传统随机森林模型, 1 000 组数据作为验证集. 实验选取参数决策树数量 $K = 100$, 评价属性数量 $m = 5$, 阈值 $\varepsilon = 0.85$, $H = 5$.

为了验证本文算法的有效性, 从仿真平台采集 850 组数据作为测试样本, 设计了如表 2 的实验, 实验模拟了由于一次浸出浸出槽 4 氰化钠添加量不足导致的运行状态评价等级从 “优”逐渐转化为 “次优”, 再转化为 “非优”. 同时分别建立了传统随机森林算法在不同决策树数量下的评价模型, 对比改进前后模型决策树的数量可知改进算法在提高精度的同时降低了模型的复杂度.

表 2　浸出过程实验设计

Table 2　Experiment of leaching process

数据	等级	描述
1~304	优	1~160个样本点, 过程运行状态等级为 “优”; 自第161个样本点, 一浸槽4氰化钠添加量逐渐减少, 运行状态等级优性逐渐减弱, 但始终保持 “优”等级, 直到第304个样本点.
305~621	次优	314~479个样本点, 保持一浸槽4氰化钠添加量不变, 过程稳定运行于等级 “次优”; 480~621个样本点, 持续减少槽4 氰化钠添加量, 运行状态等级优性再次减弱.
622~850	非优	保持一浸槽4氰化钠添加量不再变化, 过程稳定运行于等级 “非优”

3.4　实验结果分析

运行状态等级的概率计算结果如图 2, 最终评价结果如图 3. 由于实际采集的样本会存在一定程度的波动, 因而图中的概率计算会出现一些波动尖

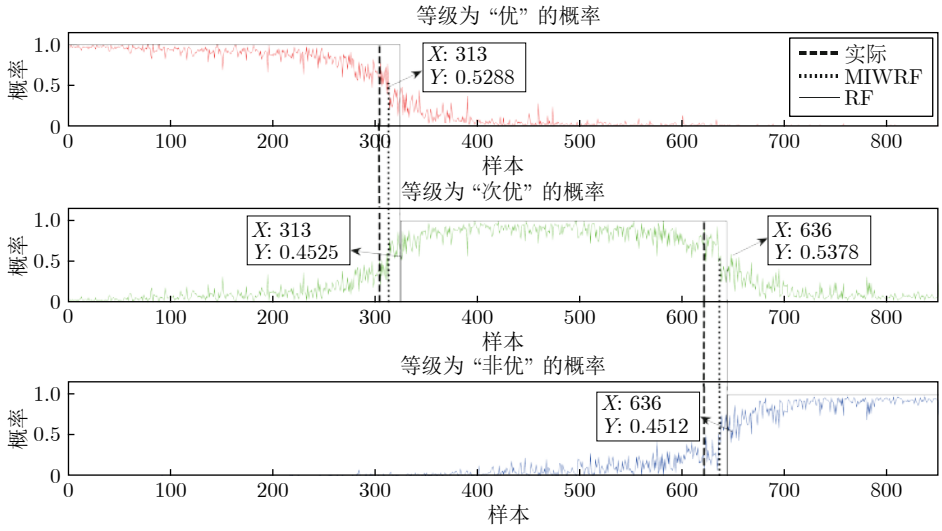

图 2　运行状态等级概率

Fig. 2　Probability of grade of running state

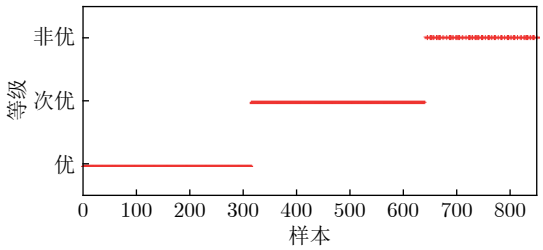

图 3 运行状态评价等级

Fig. 3 Grade of running state

点. 从图 2 的概率计算结果可以看出, 在前 160 个样本点, 等级 “优”的概率最大: 从 161 个样本点开始, 等级 “优”的概率逐渐减小, 等级 “次优”的概率逐渐增大. 314~479 个样本点中, 等级 “次优”的概率最大: 自 480 个样本点起, 等级 “次优”的概率逐渐减小, 等级 “非优”的概率逐渐增大. 从第 637 个样本点起, 等级 “非优”的概率最大. 根据运行状态等级在线评价策略, 运行状态等级评价结果为: 1~317 个样本点, 运行状态等级评价结果为 “优”等级: 318~640 个样本点, 为 “次优”等级: 641 个样本点起, 为 “非优”等级. 与实际运行状态等级对比, 精度达到 96.2%, 而随机森林评价模型精度为 93.6%. 验证了本文所提运行评价方法的有效性.

两种算法的对比实验结果如表 3 所示. 从表 3 可以看出与传统的随机森林算法相比, 本文所提改进算法的决策树数量大大减少, 且改进算法在一定程度上提高了评价精度.

为了进一步验证所提方法的评价性能, 将 MIWRF 方法与 KNN (K near neighbor), ANN (Artificial neural network), RF (Random forest), 基于互信息精减决策树但未加权的随机森林算法 MIRF (Mutual information random forest) 和文献 [2] 中的 FDbD 方法进行了对比实验. KNN 中, 近邻数设置为 10 并采用曼哈顿距离. ANN 中, 神经网络的层数为 2, 隐含层神经元个数为 10. RF、MIRF、MIWRF 的参数设置与第 3.3 节一致. FDbD 是基于 Dempster-Shafer 理论的模糊动态因果图方法, 该方法应用模糊理论减少了定量变量离散化导致的信息损失, 同时针对 DCD (Dynamic causal diagram) 中取值/状态既能通过原因节点进行推理又能被测量/估计的节点, 通过 DST 先将多源信息进行融合, 再进行推理, 弥补了传统 DCD 无法同时利用推理和测量估计所提供的信息的缺陷.

表 4 列出了 6 种算法的运行结果, 包括评价精度、建模时间和测试时间, 可以看出, 通过对决策树的精简和加权投票, 可以得到最佳的运行状态评价结果. 与传统的评价方法 (KNN, ANN) 相比, 基于随机森林的评价方法在精度上有了显著的提高, 这表明基于集成学习的评价方法比单一模型的评价方法具有更好的性能. 对于 3 种基于随机森林的运行

表 3 RF 与 MIWRF 实验结果对比

Table 3 Comparison of experimental results of RF and MIWRF

实验编号	RF决策树数量	MIWRF决策树数量	RF评价精度 (%)	MIWRF评价精度 (%)	决策树减少量 (%)
1	50	19	92.9	94.9	62.0
2	60	24	93.0	95.0	60.0
3	70	31	93.0	95.1	55.8
4	80	32	93.9	95.3	60.0
5	90	36	93.5	95.7	60.0
6	100	39	93.6	96.2	61.0
7	110	43	93.6	95.9	60.9
8	120	46	93.5	95.7	61.7
9	130	51	93.5	96.0	60.8
10	140	54	93.6	95.9	61.4
11	150	58	93.5	95.7	61.3
12	160	62	93.6	95.8	60.6
13	170	63	93.6	96.1	62.9
14	180	65	93.5	96.2	63.9
15	190	69	93.6	96.1	63.7
16	200	71	93.7	96.2	64.5

表 4　6 种评价方法的运行状态评价性能
Table 4　Performances of 6 evaluation methods

方法	KNN	ANN	RF	MIRF	MIWRF	FDbD
精度 (%)	88.8	91.2	93.6	95.2	96.2	93.3
建模时间 (s)	0	51.1519	27.4507	23.0871	23.6831	28.7613
测试时间 (s)	0.69011	0.11933	1.3123	0.9062	0.9639	1.1534

评价方法, 传统的随机森林建模时间相对较长, 主要是因为 RF 的训练集比 MIRF 和 MIWRF 大, 不需要从训练集中划分出验证集. 但 RF 的评价精度比 MIRF 和 MIWRF 低, 这主要是因为 RF 的决策树存在冗余且部分精度较低, 而基于互信息分组选择后的决策树具有多样性且精度高. 高精度和多样性的基学习器能提高集成学习器的泛化能力. 与 MIRF 相比, MIWRF 对于运行状态的精度进一步提高但训练时间与测试时间略长, 这是由于 MIWRF 算法对决策树输出结果进行了加权, 强化了性能好的决策树同时弱化了性能差的决策树. 与 FDbD 方法相比, MIWRF 的精度有一定的提升且在时间上有微弱的优势. FDbD 基于过程知识和数据建立运行评价模型, 在评价模型中引入过程知识一方面充分利用了过程知识, 另一方面增强了模型的解释性, 这些优点在非优原因追溯阶段得到了充分的体现, 基于 FDbD 模型追溯的非优原因更接近引起非优运行状态的根本原因. 但在实际应用中过程知识的提取相对困难且所获得的过程知识难以保证完备, 同时加入过程知识的评价模型难以推广到一般的工业过程. MIWRF 完全基于数据, 能轻易推广到一般的工业过程, 但如何基于 MIWRF 进行非优原因追溯有待进一步研究. 综上所述, MIWRF 不仅降低了评价模型的复杂度, 还提高了评价模型的精度.

3.5　实际数据测试

为了进一步验证本文方法的实用性, 从某冶金厂获取 5 组现场数据用于仿真验证, 5 组数据的运行状态评价评价结果如图 4 ~ 8 所示, 评价精度达 96.1%, 能够满足定量信息与定性信息共存的复杂工业过程的运行状态评价要求. 说明本文所提方法在现场实际生产过程中具有应用价值.

4　结束语

本文针对定量信息与定性信息共存的工业生产

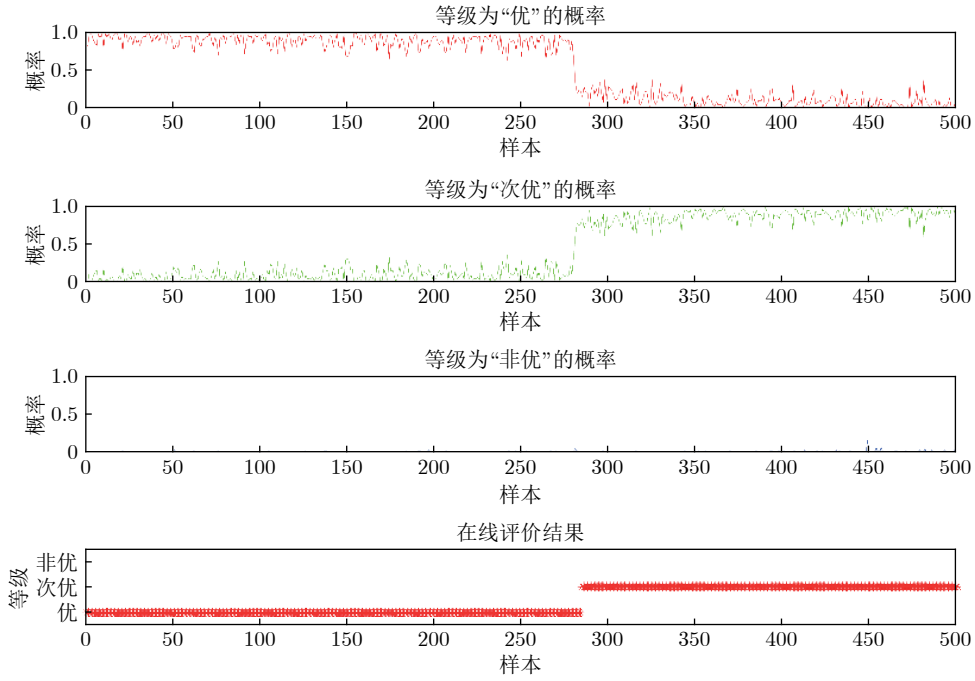

图 4　优到次优的转换
Fig. 4　Transformation from optimal to suboptimal

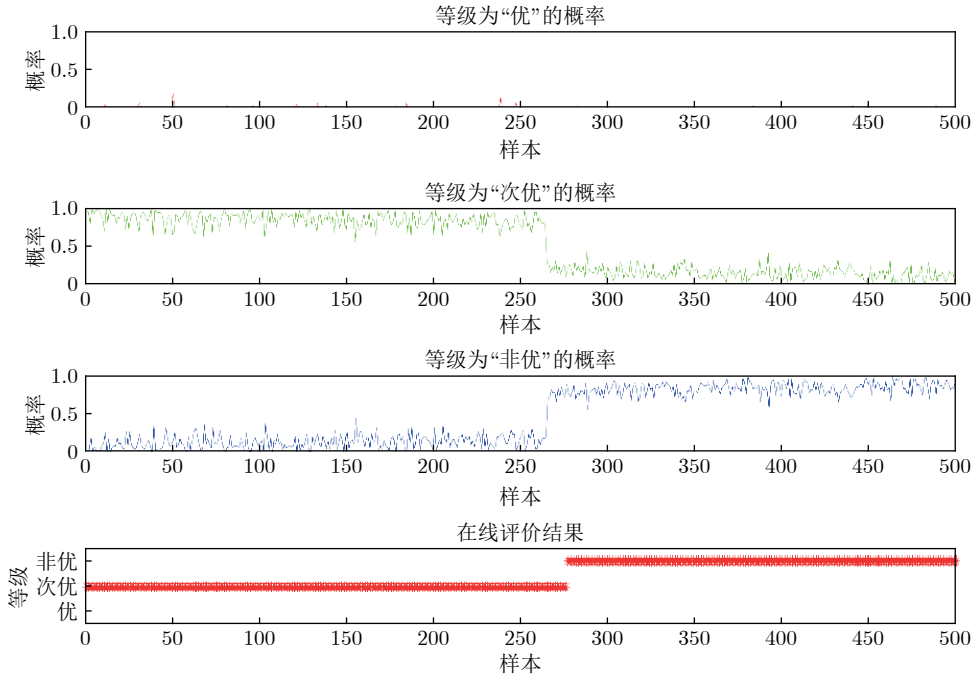

图 5 次优到非优的转换

Fig. 5 Transformation from suboptimal to non-optimal

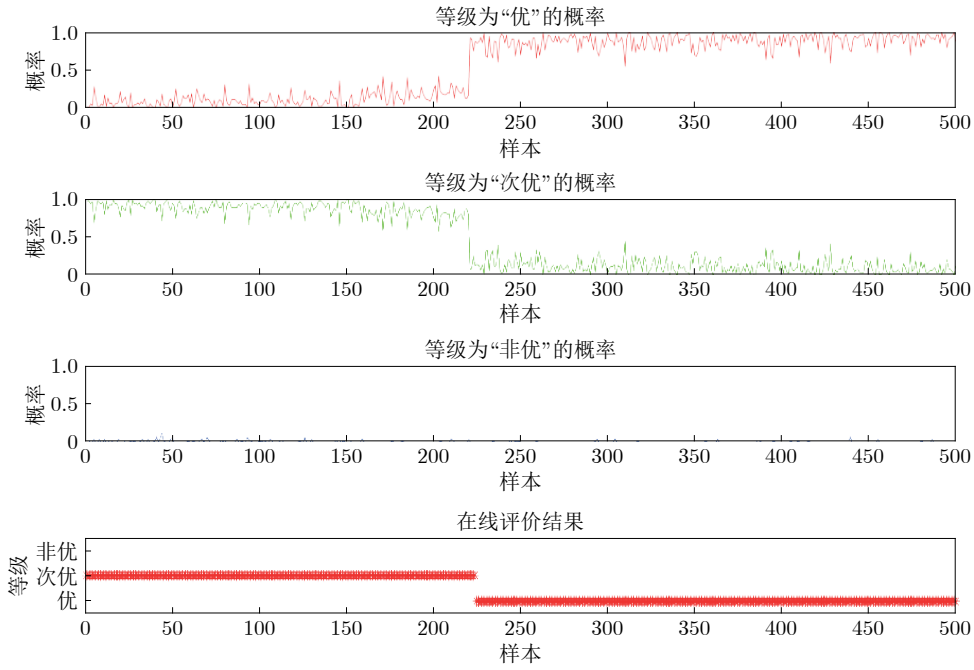

图 6 次优到优的转换

Fig. 6 Transformation from suboptimal to optimal

过程的运行状态评价问题, 提出了基于互信息的改进随机森林方法, 并且建立了运行状态评价模型以及相应的在线评价策略. 解决了定量与定性信息共存下的运行状态评价问题, 与传统随机森林方法相比, 不但提高了运行状态评价的精度, 而且降低了模型的复杂度. 将本文所提出的方法与传统运行评价方法进行了对比仿真实验并应用于湿法冶金氰化浸出过程, 仿真和实际应用结果证明了所提方法的

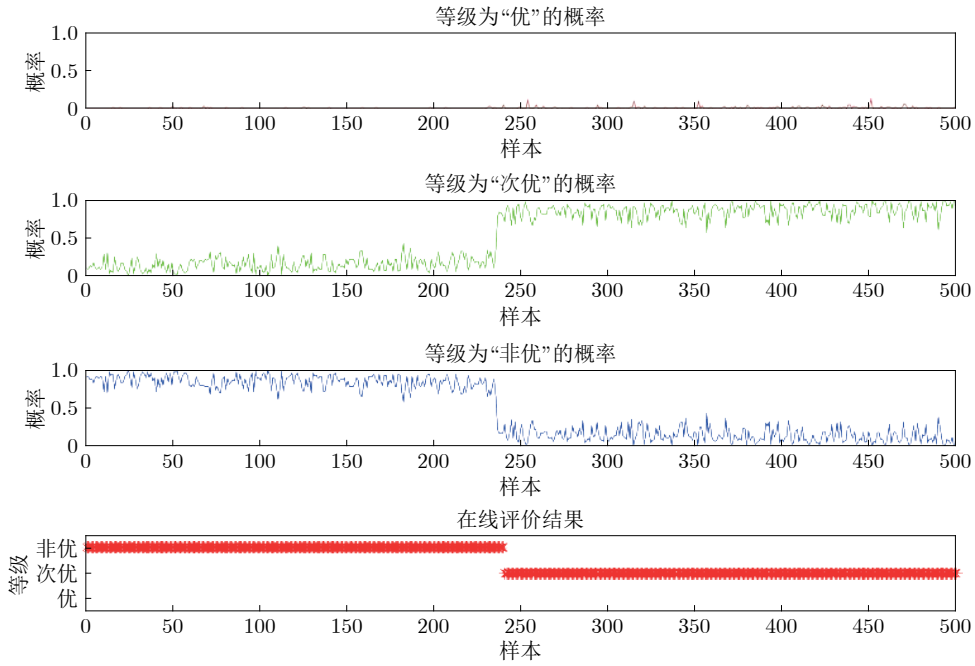

图 7　非优到次优的转换

Fig. 7　Transformation from non-optimal to suboptimal

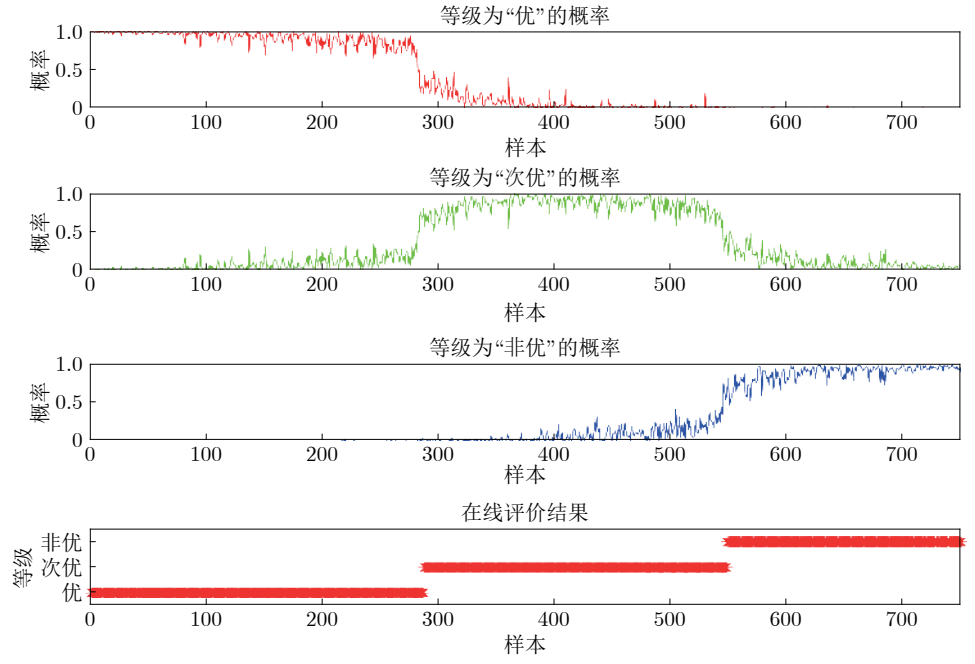

图 8　优、次优到非优的转换

Fig. 8　Transformation from optimal, suboptimal to non-optimal

正确性与有效性.

此外, 在对上述工作进行研究时, 本文并未考虑如何采取有效措施进行非优运行状态原因追溯. 即当运行状态评价结果为非优时, 通过非优原因追溯, 追溯出导致非优的变量, 并结合过程特性给出操作指导建议. 未来的研究将会围绕如何基于随机森林进行非优原因追溯展开.

综上所述, 本文针对复杂工业过程提出的运行状态评价方法, 在现有成熟方法上进行改进和应用, 具有科研价值, 但仍存在研究空间.

References

1 Ye L B, Liu Y M, Fei Z S, Liang L. Online probabilistic assessment of operating performance based on safety and optimality indices for multimode industrial processes. *Industrial and Engineering Chemistry Research*, 2009, **48**(24): 10912−10923

2 Zou X Y, Wang F L, Chang Y Q, Zhang B. Process operating performance optimality assessment and non-optimal cause identification under uncertainties. *Chemical Engineering Research and Design*, 2017, **120**: 348−359

3 Chang Y Q, Zou X Y, Wang F L, Zhao L P, Zheng W. Multimode plant-wide process operating performance assessment based on two-level multi-block hybrid model. *Chemical Engineering Research and Design*, 2018, **136**: 721−733

4 Wang Dong, Liu Huai-Liang, Xu Guo-Hua. Application of case-based reasoning in faulty dignoses system. *Computer Engineering*, 2003, **29**(12): 10−12
(王东, 刘怀亮, 徐国华. 案例推理在故障诊断系统中的应用. 计算机工程, 2003, **29**(12): 10−12)

5 Jiang Tao, Han Fu-Chun, Fan Wei-Xing. Evaluating the condition of overhead transmission lines based on rough sets theory. *Computer Engineering*, 2007, (12): 58−60
(姜涛, 韩富春, 范卫星. 基于粗糙集理论的架空输电线路运行状态评估. 电气技术, 2007, (12): 58−60)

6 Guo L J, Gao J J, Yang J F, Kang J X. Criticality evaluation of petrochemical equipment based on fuzzy comprehensive evaluation and a BP neural network. *Journal of Loss Prevention in the Process Industries*, 2009, **22**(4): 469−476

7 Xu G, Yang Y P, Lu S Y, Li L, Song X N. Comprehensive evaluation of coal-fired power plants based on grey relational analysis and analytic hierarchy process. *Energy Policy*, 2011, **39**(5): 2343−2351

8 Liu Y, Wang F L, Chang Y Q. Online fuzzy assessment of operating performance and cause identification of nonoptimal grades for industrial processes. *Industrial and Engineering Chemistry Research*, 2013, **52**(50): 18022−18030

9 Zou Xiao-Yu, Chang Yu-Qing, Wang Fu-Li, Zhou Yang. Operation performance assessment for multimode processes based on GMM and Bayesian inference. *Control Theory and Applications*, 2016, **33**(2): 164−171
(邹筱瑜, 常玉清, 王福利, 周阳. 基于 GMM 和贝叶斯推理的多模态过程运行状态评价. 控制理论与应用, 2016, **33**(2): 164−171)

10 Liu Y, Wang F L, Chang Y Q. Operating optimality assessment and nonoptimal cause identification for multimode industrial process with transitions. *The Canadian Journal of Chemical Engineering*, 2016, **94**(7): 1342−1353

11 Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123−140

12 Tuv E, Borisov A, Runger G, Torkkola K. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 2009, **10**(3): 1341−1366

13 Paul A, Mukherjee D P, Das P, Gangopadhyay A, Chintha A R, Kundu S. Improved random forest for classification. *IEEE Transactions on Image Processin*, 2018, **27**(8): 4012−4024

14 Schulter S, Wohlhart P, Leistner C, Saffari A, Roth P M, Bischof H. Alternating decision forests. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, Oregon, USA: IEEE, 2013. 508−515

15 Liu Y, Ge Z Q. Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection. *Journal of Process Control*, 2018, **64**: 62−70

16 Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5−32

17 Tao S. Steve H. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 2006, **15**(1): 118−138

18 Ho T K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(8): 832−844

19 Wang Li-Ting, Ding Xiao-Qing, Fang Chi. A novel method for robust and automatic facial features localization. *Acta Automatica Sinica*, 2006, **35**(1): 9−16
(王丽婷, 丁晓青, 方驰. 一种鲁棒的全自动人脸特征点定位方法. 自动化学报, 2006, **35**(1): 9−16)

20 Fang Kuang-Nan, Wu Jian-Bin, Zhu Jian-Ping. A review of technologies random forests. *Statistics and Information Forum*, 2011, **26**(3): 32−38
(方匡南, 吴见彬, 朱建平. 随机森林方法研究综述. 统计与信息论坛, 2011, **26**(3): 32−38)

21 Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning. In: Proceedings of the 1995 International Conference on Neural Information Processing Systems. MIT Press: 1995. 231−238

22 Cover T, Thomas J, Wiley J. *Elements of Information Theory*. Tsinghua University Press, 2003.

23 Robnik-Šikonja M. Improving random forests. In: Proceedings of the 2004 European Conference on Machine Learning (ECML 2004). Berlin, Germemy: Springer Press, 2004. 359−370

24 Chen Jia-Yong. *Research and Development of Hydrometallurgy*. Beijing: Metallurgical Industry Press. 1998.
(陈家镛. 湿法冶金的研究与发展. 冶金工业出版社, 1998.)

25 Liu Wei-Ping, Qiu Ding-Fan, Lu Hui-Min. New advances in hydrometallurgical technology. *Mining and Metallurgical Engineering*, 2003, **23**(5): 39−42
(刘维平, 邱定蕃, 卢惠民. 湿法冶金新技术进展. 矿冶工程, 2003, **23**(5): 39−42)

26 Ma Rong-Jun. New development of hydrometallurgy. *Hydrometallurgy of China*, 2007, **26**(1): 1−12
(马荣骏. 湿法冶金新发展. 湿法冶金, 2007, **26**(1): 1−12)

27 Chang Yu-Qing, Xu Di-Wu. New development of hydrometallurgy. *Chinese Journal of Scientific Intrument*, 2018, **39**(10): 18−26
(常玉清, 许弟伍. 含不确定信息的湿法冶金金泥品位软测量. 仪器仪表学报, 2018, **39**(10): 18−26)

28 Huang Li-Huang. Gold and Silver Extraction Technology. Beijing: Metallurgical Industry Press. 2001.
(黄礼煌. 金银提取技术. 冶金工业出版社, 2001.)

常玉清 东北大学教授. 2002 年于东北大学获得博士学位. 主要研究方向为复杂工业过程建模、监测、运行状态评价及过程优化.
E-mail: changyuqing@ise.neu.edu.cn
(**CHANG Yu-Qing** Professor at Northeastern University. She received her Ph.D. degree from Northeastern University in 2002. Her research interest covers industrial process modeling, monitoring, operation performance evaluation, and optimization.)

孙雪婷 东北大学硕士研究生. 2017 年于大连交通大学获得学士学位. 主要研究方向为复杂工业过程的建模控制与优化. 本文通信作者.
E-mail: xuetingsun111@163.com
(**SUN Xue-Ting** Master student at Northeastern University. She received her bachelor degree from Dalian Jiaotong University in 2017. Her research interest covers modeling control and optimization of complex industrial processes. Corresponding author of this paper.)

钟林生 东北大学信息科学与工程学院博士研究生. 主要研究方向为复杂工业过程运行状态评价和故障诊断.
E-mail: zhonglinsheng_neu@163.com
(**ZHONG Lin-Sheng** Ph.D. candidate at the College of Information Science and Engineering, Northeastern University. His research interest covers operating performance assessment and fault diagnosis of complex industrial processes.)

王福利 东北大学信息科学与工程学院教授. 主要研究方向为复杂工业过程建模、优化与故障诊断.
E-mail: wangfuli@ise.neu.edu.cn
(**WANG Fu-Li** Professor at Northeastern University. His research interest covers modeling control, optimization, and fault diagnosis of complex industrial processes.)

刘英娇 东北大学硕士研究生. 2015 年于东北大学秦皇岛分校获得学士学位. 主要研究方向为复杂工业过程的建模控制与优化.
E-mail: liuyingjiao1992@163.com
(**LIU Ying-Jiao** Master student at Northeastern University. She received her bachelor degree from Northeastern University at Qinhuangdao in 2015. Her research interest covers modeling control and optimization of complex industrial processes.)