DOI:10.20040/j.cnki.1000-7709.2023.20221402

基于决策树及其集成模型的水库调度规则提取方法

戴 领[1],骆光磊[2],周建中[3]

(1. 长江设计集团有限公司,湖北 武汉 430010;2. 中交第二航务工程勘察设计院有限公司,湖北 武汉 430061;3. 华中科技大学土木与水利工程学院,湖北 武汉 430074)

摘要:水库调度规则作为指导水库调度运行的重要工具,不仅是水库规划设计时期的决策参考要素,且是运行管理期影响水库综合效益发挥的关键技术之一。为此,以长江上游水库群历史调度运行数据为基础,结合水库调度原理及运行特征,挑选时段数、前期水位、入库、出库及当前时段入库作为影响因子组建输入因子集,综合考虑运行期数据特征及决策树原理确定时段末水位作为模型输出,同时提出相应基于水库调节库容的水位评价指标,并采用相关系数和互信息作为模型输入因子相关性评定指标,引入树形 Parzen 评估器对输入因子个数和算法超参数进行优化,在此基础上,建立了基于决策树及其集成模型的水库调度规则提取模型,形成了融合历史调度过程和专家经验的水库调度规则。试验结果表明,决策树及其集成模型在水库调度规则提取应用时具有较强的能力和适用性。

关键词:水库调度;规则提取;决策树及其集成模型;贝叶斯优化

中图分类号:TV391 **文献标志码**:A **文章编号**:1000-7709(2023)06-0045-04

1 引言

随着“节能减排”政策的贯彻落实,我国水电能源持续、快速、大规模开发,已建成一大批流域巨型梯级水库群,科学合理地管理这些大型水利设施,对实现水资源综合高效利用起着关键性作用[1]。因此,开展水库调度规则提取研究对于提升实际调度水平,提高水库防洪兴利效益具有重大的现实意义。水库调度规则是决定水库调度方式的具体要求和规定[2],其表征方式、表征形式、提取方法一直是水库调度领域的研究热点。近年来,随着新型人工智能算法理论不断完善,基于机器学习算法的现代智能模型快速发展,大量非线性拟合更强、鲁棒性更高的算法,如决策树、神经网络等算法被引入调度规则提取领域,其中决策树及其集成模型以较好的可解释性和较低的实现难度大受青睐[3,4]。由于决策树模型天然与水库调度决策过程类似,模型结构与一般水库调度规程相近,其在调度规则提取方面效果较优,可用于工程实际。本文以决策树及其集成模型与调度决策过程的相似性为切入点,以长江上游水库群为例,首先根据水库调度原理及运行特征构建输入因子集,采用相关系数和互信息作为模型输入因子筛选方法,然后分别基于决策树、随机森林和极限梯度提升树方法构建输入输出间函数映射关系,并引入贝叶斯优化算法优化输入因子个数和算法超参数,获得模型最优超参数组合,最后根据最优超参数训练得到目标函数最优的调度规则,验证了决策树及其集成模型在水库调度规则提取方面的适用性。

2 研究方法

2.1 决策树及其集成模型原理

决策树(DT)[5]为树形结构,从根节点开始采用自上向下的递归方式将给定特征空间划分为一系列子空间,每个树节点表示对某个特征的划分,每个叶子节点代表一个划分类别。随机森林(RF)算法是一种基于决策树和自举汇聚法(Bagging)的集成模型,适用于解决小样本、高维度特征数据分类和回归问题[6]。极限梯度提升树(XGBoost)是一种基于决策树与提升算法(Boosting)的集成模型[5]。

收稿日期:2022-07-11,**修回日期**:2022-08-16

基金项目:湖北省博士后创新实践岗位(2022CXGW003);国家自然科学基金项目(U1865202,52039004)

作者简介:戴领(1994-),男,博士,研究方向为水库调度,E-mail:dailing2021@qq.com

2.2 基于树形结构 Parzen 估计器的超参数优化算法

超参数优化旨在寻找使算法在验证集上表现性能最佳的超参数组合。为提高算法超参数搜索效率，不同类型贝叶斯优化算法的基本原理如下：首先基于目标函数已有的评估结果建立代理模型，进而通过采集函数寻找代理模型期望收益最大的超参数，然后将其与评估结果作为输入更新代理模型，以此往复交替进行，最终获得目标值最优的超参数。相对于随机或网格搜索，贝叶斯优化算法使用不断更新的代理模型，并通过推断已有结果来“集中”更有希望的超参数，从而大大减少了调参时间[7]。

2.3 调度规则提取方法

水库调度决策与水位、入库流量、出库流量、出力、负荷、预报流量等因素密切相关，考虑到水库出力、负荷、预报流量数据难以获取，选择时段数、水库前期水位、入库、出库及当前时段入库作为影响因子构成输入因子集，而时段数和水库前一时刻状态及当前时段入库与当前时段决策关系最密切，因此将其作为模型输入必选因子，采用 F 检验方法[8]（FR）和互信息方法[9]（MIR）定量评估剩余因子与输出因子间的相关性并将其作为备选因子，然后分别采用 DT、RF、XGBoost 算法构建 6 个模型，即 FR-DT、MIR-DT、FR-RF、MIR-RF、FR-XGB、MIR-XGB，最后采用树形 Parzen 评估器同时优化输入因子选择个数与 DT、RF、XGB 算法超参数，寻找模型效果最优的超参数组合。

此外，采用树模型进行规则提取。树模型本质上是一种聚类模型，模型输出是训练样本不同聚类组合下的标签均值，因此其在预测时，输出将被限制在训练样本标签最大、最小值之间，而由于水库出库流量年际变幅较大，未来时段出库流量极大概率在训练集范围外，若采用水库出库流量作为模型输出，将严重降低模型模拟精度，不利于工程实际应用。相反水库水位变动范围在年际间差异较小，基本在正常蓄水位与死水位间波动，故选定时段末水位作为输出。由于上下游水库所处位置不同，上游水库运行水位普遍较高，若采用传统平均绝对百分比误差评估模型，会造成指标虚高，上下游水库模型精度无法公度。为此，研究综合考虑水库自身特性，在原有平均绝对百分比误差的基础上，提出一种基于水库调节库容的模型评价指标 $H_{\text{H-MAPE}}$，表达式为：

$$H_{\text{H-MAPE}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i-y_i|}{y_{\text{base}}} \tag{1}$$

式中，N 为样本总量；y_i、$\hat{y}_i$ 分别为第 i 个样本的实际值、预测值；y_{base} 为水库正常蓄水位与死水位差值。

基于决策树及其集成模型的水库调度规则提取方法技术路线图见图 1，具体步骤如下。

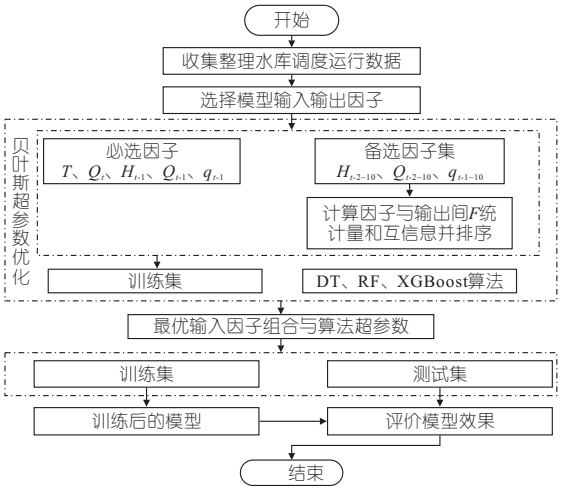

图 1 调度规则提取模型流程

Fig. 1 Flow chart of operation rule extraction model

步骤 1 从水库调度运行数据中挑选时段数 T、前期水位 H、入库 Q、出库 q 等特征组建模型输入因子集，选择当前时段末水位作为模型输出，选择时段数、前 1 时段入库、出库、末水位及当前时段入库作为必选输入因子，选择前 2～10 时段入库、出库、末水位构成备选输入因子集，计算备选输入因子集中各因子与输出变量间的 F 统计量和互信息并排序。

步骤 2 将输入因子选择个数 featureNum 与 DT、RF、XGBoost 本身超参数组合成超参数集，featureNum 含义为在备选因子集中选择 F 统计量和互信息排序靠前的 featureNum 个因子与必选输入因子构成模型输入。

步骤 3 根据模型输入输出构造数据样本并划分训练集和测试集，以训练集上交叉验证的均方根误差均值最小为目标函数，采用树形 Parzen 估计器优化模型超参数。

步骤 4 选用步骤 3 中最优的超参数重新训练模型，计算模型在测试集上的各项指标，评估模型效果。

本文所提调度规则相较于以往大部分研究中的调度规则提取稍有不同，其一般采用大量不同来水优化调度后的调度过程作为数据样本集，此类调度规则侧重于决策支持，表明在当前水库状态下，若按提取的调度规则进行决策，能使优化调度模型的目标值（发电量、弃水量等）更优，其评价调度规则优劣取决于按该调度规则模拟调度后水

库发电量、弃水量、发电保证率等[10,11],而本文采用水库历史运行数据作为基础样本,提取的调度规则更侧重于预测和模拟,旨在从历史数据中挖掘水库调度运行规律和调度员人工经验,即模拟任意来水条件下水库按当前调度方式运行的响应过程,其评价标准是按调度规则获得的水位、出库或出力预测值与实际值的接近程度。

3　实例应用

以长江上游梨园、阿海、金安桥、龙开口、鲁地拉、观音岩、锦屏一级、二滩、溪洛渡、向家坝、瀑布沟、紫坪铺、碧口、宝珠寺、亭子口、洪家渡、乌江渡、构皮滩、思林、沙沱、彭水、银盘、三峡、葛洲坝 24 座水库为例(流域拓扑图见图 2),分别以阿海、观音岩、瀑布沟、二滩水库作为日调节、周调节、季调节、年调节类型水库代表,重点分析决策树及其集成模型在不同类型水库、不同调度时期、不同水位等级下的模拟效果。各水库均以 2020 年数据作为测试样本,2015～2019 年数据为训练样本。

图 3、表 1 分别为各水库测试集上实测值与模拟值的对比图及各评价指标结果。由图 3、表 1 可知,不同类型水库模拟精度均较优,确定性系数 R^2 均在 0.95 以上,$H_{\text{H-MAPE}}$ 指标均在 0.05 以下,均方根误差 R_{RMSE} 指标大部分在 0.3～0.6 m 之间,平均相对误差 M_{MAE} 指标大部分在 0.2～0.5 m 之间,不同水库、不同模型间误差分布相近,无较明显的区分特征。各水库模拟水位与实际水位基本保持一致,只有部分时段存在模拟值滞后实际值现象,其中阿海水库最为明显,各模型在 6 月底峰值水位和 8～10 月水位频繁波动时期存在明显偏差,其原因在于该时期水库水位日间变化较小,日水位自相关性较强,导致模型输出基本在前期时段末水位附近取值,难以捕捉水位突变点。对比不同类型水库结果,阿海水库散点图明显较分散,各模型 R^2 及 $H_{\text{H-MAPE}}$ 指标较其他低,二滩水库散点图基本贴近 $Y=X$ 折线,R^2 指标高达 0.999,$H_{\text{H-MAPE}}$ 低于 0.01,随着水库调节能力增

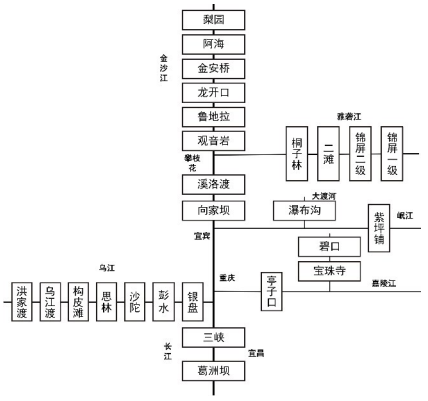

图 2　长江上游流域拓扑图

Fig. 2　Topological map of upper Yangtze River Basin

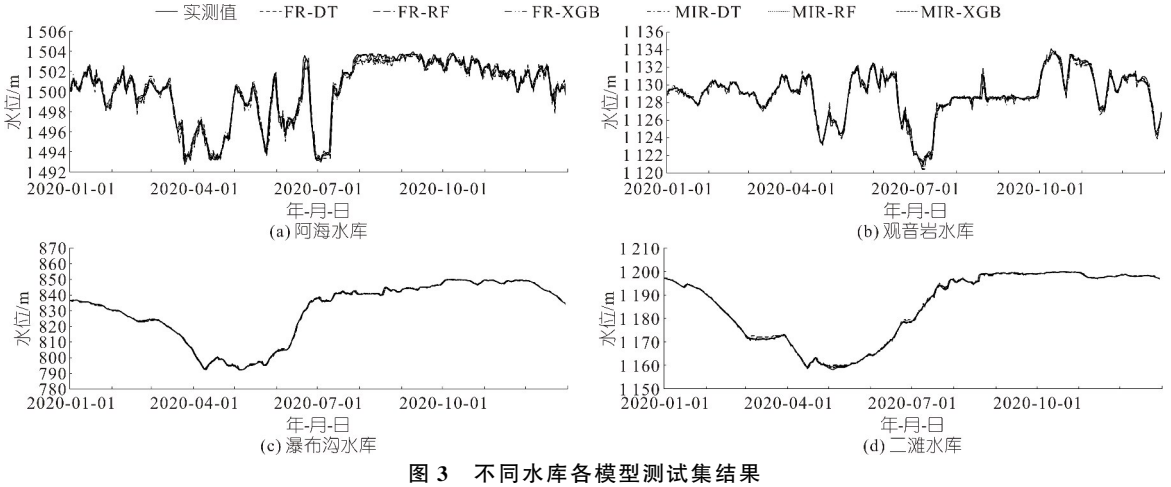

图 3　不同水库各模型测试集结果

Fig. 3　Model results in test dataset of Ahai, Guanyinyan, Pubugou and Ertan

表 1　不同水库各模型测试集指标

Tab. 1　Model index of test dataset of Ahai, Guanyinyan, Pubugou and Ertan

模型	阿海水库				观音岩水库				瀑布沟水库				二滩水库			
	R^2	$H_{\text{H-MAPE}}$	R_{RMSE}	M_{MAE}	R^2	$H_{\text{H-MAPE}}$	R_{RMSE}	M_{MAE}	R^2	$H_{\text{H-MAPE}}$	R_{RMSE}	M_{MAE}	R^2	$H_{\text{H-MAPE}}$	R_{RMSE}	M_{MAE}
FR-DT	0.960 5	0.036 1	0.598 6	0.433 1	0.960 4	0.030 6	0.499 7	0.357 4	0.998 9	0.007 7	0.625 5	0.462 6	0.998 5	0.008 7	0.556 4	0.392 6
MIR-DT	0.960 3	0.037 2	0.600 3	0.445 8	0.961 3	0.030 3	0.494 0	0.354 0	0.998 9	0.007 6	0.608 4	0.456 0	0.998 7	0.008 3	0.533 3	0.372 6
FR-RF	0.972 7	0.029 9	0.498 1	0.358 5	0.975 6	0.022 8	0.392 2	0.267 1	0.999 2	0.006 4	0.526 9	0.381 1	0.999 4	0.005 3	0.355 9	0.238 9
MIR-RF	0.972 3	0.030 4	0.501 6	0.365 2	0.976 4	0.022 4	0.385 8	0.261 7	0.999 2	0.006 3	0.518 2	0.378 8	0.999 4	0.005 4	0.367 1	0.244 5
FR-XGB	0.970 8	0.031 0	0.515 0	0.371 5	0.984 8	0.017 4	0.310 1	0.204 2	0.999 6	0.004 8	0.381 4	0.289 5	0.999 6	0.004 4	0.297 4	0.199 2
MIR-XGB	0.970 9	0.031 5	0.514 3	0.378 3	0.982 6	0.018 4	0.331 5	0.215 6	0.999 5	0.004 9	0.400 0	0.296 4	0.999 6	0.004 5	0.296 7	0.200 8

强，模型评估指标逐渐提高，其原因在于调节能力强的水库运行周期长，水位日间波动较小，变化规律明显，模型训练难度低。对比不同特征提取方法结果，各水库各模型差异不大。对比不同训练模型结果，DT 模型精度普遍低于 RF 和 XGBoost，瀑布沟和二滩水库 XGBoost 模型各项指标均优于 DT 和 RF，阿海和观音岩水库 XGBoost 模型各项指标与 RF 相当，其原因在于 RF 和 XGBoost 作为决策树的不同集成模型，其学习能力优于作为弱评估器的决策树，而 RF 和 XGBoost 的学习能力则无法对比，不同数据集、超参数优化方法均会影响模型的最终结果。

采用相同方法对长江上游其余 20 座水库进行模型训练，图 4 为银盘水库 MIR-RF 模型训练集及测试集实测与模拟水位。由图 4 可知，银盘水库水位波动十分剧烈，峰谷差异大，决策树及其集成模型输出值一般为叶子节点上样本标签的平均值，从而使树模型难以拟合这类数据。图 5 为 MIR-RF 模型各指标与水库调节库容关系。考虑到银盘水库模拟效果较差，图 5 中散点未包含银盘水库。由图 5 可知，随着水库调节库容的提高，

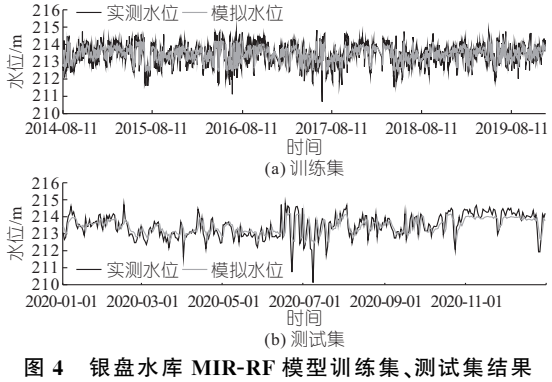

图 4 银盘水库 MIR-RF 模型训练集、测试集结果

Fig. 4 MIR-RF model results in train dataset and test dataset of Yinpan

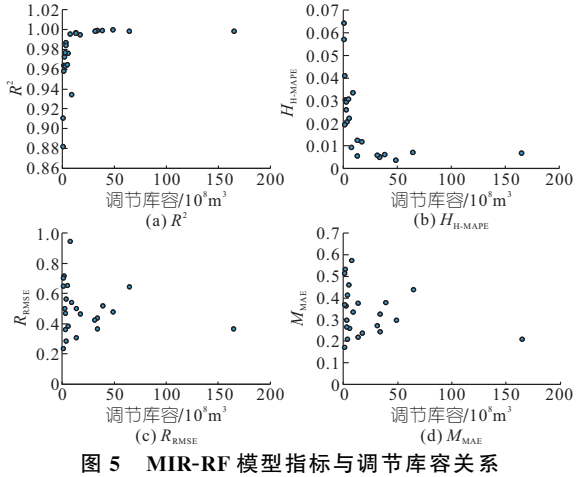

图 5 MIR-RF 模型指标与调节库容关系

Fig. 5 Relation of MIR-RF model index and reservoir regulation capacity

模型 R^2 指标呈急剧增长后稳定，$H_{\text{H-MAPE}}$ 急剧下降后稳定，R_{RMSE}、M_{MAE} 指标无明显规律，调节库容较大的水库，年运行过程稳定，水位消落、上涨有较明显的规律，树及其集成模型对于该类数据具有较强的学习能力，反之，水位过程波动频繁，无明显的变化规律，且树模型输出平均值难以拟合剧烈振荡过程。

4 结论

a. 从决策树及其集成模型结构与调度决策过程的相似性分析着手，研究了其在水库调度规则提取方面的适用性，引入贝叶斯优化理论，建立了模型输入因子与算法超参数协同优化的水库调度规则提取模型，提出了基于水库调节库容的模拟水位评价指标 $H_{\text{H-MAPE}}$，分析了不同模型在长江上游 24 座水库中的拟合精度。

b. 基于决策树及其集成模型提取的水库调度规则效果较好，精度较高。且水库调节能力越强，水位日间波动较小，变化规律明显，模型训练难度低，水库调节能力越弱，水位波动越频繁，模型误差越大。

c. 未来应在模型输入因子选取对模型精度的影响和如何考虑梯级水库间的相互影响等方面进行深入研究。

参考文献：

[1] 何中政. 水库群中长期发电调度优化方法研究[D]. 武汉：华中科技大学，2020.

[2] 李子婷. 安康水电站能量指标复核及调度规则研究[D]. 西安：西安理工大学，2007.

[3] 丁胜祥，董增川，张莉. 基于决策树算法的洪水预报模型[J]. 水力发电，2011，37(7)：8-11.

[4] 郭旭宁. 水量调度规则的建模理论与求解方法[D]. 武汉：武汉大学，2013.

[5] TIANQI CHEN，CARLOS GUESTRIN. XGBoost：A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016：785-794.

[6] BREIMAN L. Random forests[J]. Machine learning，2001，1(45)：5-32.

[7] 浮盼盼，司琪，王鑫赛. 机器学习算法的超参数优化：理论与实践[J]. 电脑编程技巧与维护，2020(12)：116-117.

[8] 龚雪娇，朱瑞金，唐波. 基于贝叶斯优化 XGBoost 的短期峰值负荷预测[J]. 电力工程技术，2020，39(6)：76-81.

（下转第 44 页）

virtual-water management-a case study of Hubei Province, China [J]. Journal of cleaner production, 2021,293:126244.

[2] ALLAN J A. Fortunately there are substitutes for water otherwise our hydro-political futures would be impossible[C]//Priorities for water resources allocation and management,ODA,1993:13-26.

[3] 王丽川，侯保灯，周毓彦，等. 基于水足迹理论的北京市水资源利用评价[J]. 南水北调与水利科技(中英文)，2021,19(4):680-688.

[4] 朱永楠，姜珊，赵勇，等. 我国煤电生产水足迹评价[J]. 水电能源科学,2019,37(9):28-31.

[5] 许爽爽，马树才，付云鹏. 基于投入产出法的辽宁省水足迹和虚拟水核算[J]. 沈阳师范大学学报(自然科学版)，2018，36(1)：58-62.

[6] ZHAO D, LIU J, YANG H, et al. Socioeconomic drivers of provincial-level changes in the blue and green water footprints in China [J]. Resources, conservation and recycling, 2021, 175: 105834.

[7] 陈倩云，安婷莉，王玉宝，等. 我国北方重点煤电基地发展伴生的水资源压力分析[J]. 水电能源科学，2019，37(7)：30-34.

Characteristics and Change Driving Force of Water Footprint in Hubei Province

YUAN Yan-bin[1], LIAN Yi-wen[1], YUAN Xiao-hui[2], ZHOU Han[1], DONG Heng[1], ZHANG Xiao-pan[1]

(1. School of Resources and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China; 2. School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Water footprint evaluation is of guiding significance for the rational development of regional water resources. The input-output method was used to account for the water footprint of Hubei Province from 2007 to 2017, and the driving factors of water footprint changes were analyzed using the structural decomposition model. The results show that the virtual water content in Hubei Province decreased significantly, and water use efficiency improved from 2007 to 2017. The total water footprint showed a trend of rising and then falling, with a net increase of 19.9%, and the main growth sectors were construction and services. Technology level and economic scale were the main factors inhibiting and promoting the increase of water footprint, respectively. However, their effects on water footprint changes gradually weakened, and both showed a reduction effect. From 2007 to 2017, the sectoral linkages changed from positive to negative driving force, and the inhibitory effect on water footprint was revealed, which shows that optimizing the industrial production process helps carry out water conservation. The impact of driving factor on different sectors showed heterogeneity. The adjustment of industrial structure inhibited the increase of water footprint in high water-consuming sectors such as agriculture while promoting the increase in construction and services. In the future, the industry scale and residential demand in sectors producing high-value-added products will grow, such as services, and water-saving technology development should be shifted to these sectors promptly.

Key words: input-output; water footprint; virtual water; structural decomposition; driving factor

(上接第 48 页)

[9] SATTARI M T, APAYDIN H, OZTURK F, et al. Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir[J]. Lake and reservoir management,2012, 28(2):142-152.

[10] 李力，周建中，戴领，等. 金沙江下游梯级水库蓄水期多目标生态调度研究[J]. 水电能源科学，2020，38(11):62-66.

[11] 刘志刚，胡斌奇，伍永刚，等. 基于云模型的水库调度函数拟合方法研究[J]. 水电能源科学，2017，35(3):53-56，23.

Reservoir Operation Rule Extraction Method Based on Decision Tree and Its Integrated Model

DAI Ling[1], LUO Guang-lei[2], ZHOU Jian-zhong[3]

(1. CISPDR Corporation, Wuhan 430010, China; 2. CCCC Second Harbor Consultants Co., Ltd., Wuhan 430061, China; 3. School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Reservoir operation rules, as an important tool to guide reservoir operation, are not only the decision-making reference in the reservoir planning and design period, but also one of the key technologies affecting the comprehensive benefits of the reservoir in the operation and management period. Therefore, based on the historical operation data of reservoirs in the upper reaches of the Yangtze River, the number of periods, early water level, inflow, outflow and current inflow were selected as the influence factors to form the input factor set combined with the reservoir operation principle and operation characteristics. Comprehensively considering the characteristics of operation data and the principle of decision tree, the end of period water level was determined as the model output and then the corresponding simulated water level evaluation index was proposed based on the reservoir regulation capacity. The correlation coefficient and mutual information were used as the correlation evaluation index of model input factors, and the tree Parzen evaluator was introduced to optimize the number of input factors and algorithm super parameters. Finally, the reservoir operation rule extraction model based on decision tree and its integrated model was established and the reservoir operation rule integrating historical operation process and expert experience was formed. The experimental results show that the decision tree and its integrated model have strong ability and applicability in the extraction and application of reservoir operation rules.

Key words: reservoir operation; rule extraction; decision tree and its integrated model; Bayesian optimization